

Thèse de Doctorat

Jean-Baptiste GUILLON

*Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le label de L'Université Nantes Angers Le Mans*

École doctorale : Cognitions, Educations, Interactions (ED 504)

Discipline : Philosophie

Spécialité : Métaphysique, Epistémologie

Unité de recherche : Centre Atlantique de Philosophie (EA 2163)

Soutenue le 21 novembre 2014

Thèse N° :

L'Epistémologie du Libre-Arbitre dans la Tradition analytique

JURY

Rapporteurs :	M. Jérôme DOKIC , Directeur d'études à l'EHESS M. Roger POUIVET , Professeur, Université de Lorraine
Examineurs :	M. Pascal ENGEL , Directeur d'études à l'EHESS M. Derk PEREBOOM , Professeur, Université Cornell M. Carlos MOYA , Professeur, Université de Valencia
Directeur de Thèse :	M. Cyrille MICHON , Professeur, Université de Nantes

Remerciements

Je tiens à remercier tout d'abord mon directeur de thèse, Cyrille Michon, sans qui je n'aurais sans doute pas fait de métaphysique analytique. Sa grande disponibilité pour discuter de mes recherches, ainsi que ses encouragements réguliers, ont été une aide constante dans ce voyage au long cours.

Ma reconnaissance va également à l'Université de Notre Dame, et à ses professeurs, qui m'ont accueilli comme visiteur lors de deux semestres qui constituent ma première et véritable formation en philosophie analytique. Je pense en particulier à Robert Audi et à Peter van Inwagen, à qui je dois à peu près tout ce que je sais en épistémologie et en métaphysique (respectivement).

Je tiens également à remercier l'Université CEU de Budapest, ainsi que les organisateurs de l'école d'été « Aspects of Responsibility », en particulier Derk Pereboom, Timothy O'Connor et Mark Balaguer. Cette école d'été m'a appris en quelques jours plus de choses sur le libre arbitre que je n'avais pu en apprendre par moi-même en plusieurs mois.

La rédaction de la thèse aurait été beaucoup plus longue et difficile si je n'avais pas eu les conditions de travail que j'ai eues au département de philosophie de l'Université de Nantes. Je n'oublierai pas l'accueil si chaleureux que m'ont réservé mes collègues, lors de mes premières années comme enseignant, notamment Denis Moreau, Patrick Lang et Bruno Gnassounou.

Emmanuel Baierlé et Martine Nida-Rümelin m'ont permis de réaliser, à l'Université de Fribourg, le rêve de beaucoup de thésards : organiser une conférence internationale sur son sujet de thèse. Je ne saurais trop leur exprimer ma gratitude.

Les arguments de ce travail ont bénéficié de nombreuses discussions avec des amis de Notre Dame, de Budapest, ou de Nantes : Ghislain Le Gousse, Marcin Iwanicki, Ide Lévi-Fouche, Matthew Lee, Emmanuel Baierlé, Oisín Deery, Stefano Cossara, Joshua Rasmussen, Andrew Bailey, Paul Clavier et Bob Lockie.

Plusieurs amis m'ont rendu le grand service de relire des chapitres : Ghislain Le Gousse, Jérôme Moreau, Raphaëlle Abitbol, Stefano Cossara, Guy de Tonquedec, Ide Lévi-Fouche, Astrid Aschehoug et Alexis Frémont.

Enfin, je tiens à remercier mes proches, mes parents, mes frères et sœur, beau-frère et belles-sœurs, pour leur soutien moral au cours de cette longue tâche, en particulier aux moments les plus éprouvants.

Introduction : la question épistémologique	2
PREMIERE PARTIE : PRESOMPTION DE CONNAISSANCE DU LIBRE ARBITRE	39
Chapitre 1 : L'expérience commune et la définition de « libre arbitre »	40
Chapitre 2 : Justification <i>prima facie</i> à croire que nous sommes libres	79
Chapitre 3 : Justification <i>prima facie</i> à croire que nous savons que nous sommes libres	178
DEUXIÈME PARTIE : ABSENCE DE RÉFUTATIONS	247
Chapitre 4 : Le contenu incompatibiliste de notre expérience du libre arbitre	248
Chapitre 5 : Absence de réfutation directe du libertarianisme	346
Chapitre 6 : Absence de réfutation fondationnelle : la croyance libertarienne comme connaissance	400
Conclusion : le libertarianisme, à l'égal des croyances perceptives	466
ANNEXES	489
Bibliographie	520
Table des matières générale	534

Introduction : la question épistémologique

“A libertarian, it seems, must claim that he knows from his armchair that causal determinism is false; but how could we know in advance such an empirical thesis?”
(Fischer et al. 2007, 81)

I.	Le libre arbitre : concept et problèmes	4
1.	Difficultés pour définir le concept de « libre arbitre ».....	4
2.	Aux origines du débat contemporain	6
3.	L'étrange discrédit de la position libertarienne	7
4.	Le problème épistémologique : un Défi d'Intégration.....	10
II.	Connaissance et Justification : définitions	16
1.	une notion minimale externaliste de « connaissance »	18
2.	une notion internaliste non déontologique de « justification »	22
III.	Le Problème d'Intégration du Libre arbitre	27
1.	Justification et Rationalité dans le débat philosophique	28
2.	Rationalité et Connaissance : la contrainte de méta-cohérence	30
IV.	Plan général de cette étude	32

- Etes-vous libre ?
- Oui bien sûr !
- Comment le savez-vous ?
- ...

Je sais par expérience que ce petit dialogue aboutit, bien souvent, à un regard d'incompréhension voire de mécontentement, le genre de regard que depuis plusieurs siècles les disciples de Socrate et de Diogène se sont habitués à rencontrer lorsqu'ils pressent leurs semblables de questions indiscretes. Ce petit dialogue n'utilise pourtant que des termes bien simples et bien communs, et suit un schéma tout ce qu'il y a de plus banal : « Avez-vous encore de l'essence ? Oui, bien sûr ! Comment le savez-vous ? J'ai vérifié le niveau ce matin. » « Etes-vous prêt pour l'examen terminal ? Oui, bien sûr ! Comment le savez-vous ? J'ai été classé premier à l'examen blanc. » Ce schéma de dialogue peut avoir lieu plusieurs fois par jour sans que personne ne songe à s'offusquer de la question « comment le savez-vous ? ». Comment donc se fait-il que cela ne « marche » pas aussi bien lorsqu'il s'agit de la liberté ? Sans doute la question est-elle trop « philosophique ». C'est à cette question philosophique que cette étude sera consacrée. On peut la formuler ainsi :

Comment savons-nous que nous sommes libres ? Quel genre de connaissance pourrait être la source de notre croyance au libre arbitre ?

Si cette question laisse perplexe le commun des mortels, on pourrait s'attendre à trouver plus de réponses chez les philosophes eux-mêmes. Pourtant, lorsqu'on consulte la tradition philosophique, les questions les plus traitées sur le libre arbitre concernent ses conditions *métaphysiques* : comment le libre arbitre s'inscrit-il dans la trame des événements ? Est-il

compatible avec diverses formes de « déterminismes » (théologique, causal, ou logique) ? La question de *connaissance* proprement dite n'apparaît que fort rarement. Il serait exagéré de dire qu'elle est totalement absente : elle a notamment été l'objet d'un débat assez vif à l'époque moderne entre Descartes (et les cartésiens), tenants d'une connaissance du libre arbitre « par la seule expérience que nous en avons » (Descartes 1953, 588), et d'autre part toutes sortes d'objecteurs (au premier rang desquels Spinoza et Leibniz) qui ne pouvaient accorder à notre sentiment intérieur une telle autorité épistémologique. Un autre débat fondamental dans la période moderne est celui qui opposa Thomas Reid à David Hume, le premier défendant une conception de la liberté où l'agent peut se *connaître* lui-même comme cause de l'action, le second étant, comme on sait, sceptique à l'égard de toute possibilité de connaissance de la causalité. Dans chacun de ces deux débats, débat « Cartésien » et débat « Humien », la place de la théorie de la connaissance, de « l'épistémologie » au sens général que lui donnent les auteurs analytiques, est au moins aussi importante que la théorie du libre arbitre proprement dit et de ses conditions métaphysiques. Ce sera également le cas de cette étude qui peut être considérée comme une étude « d'épistémologie appliquée », au sens où l'on parle aujourd'hui « d'éthique appliquée ».

Dans le débat actuel en philosophie analytique, la question « épistémologique » du libre arbitre n'est que rarement formulée de manière directe et explicite. L'enjeu de mon travail est de la remettre sur le devant de la scène, et de montrer que les avancées récentes en philosophie du libre arbitre d'une part, et en épistémologie d'autre part, peuvent aujourd'hui être réunies pour fournir un modèle vraiment développé d'épistémologie du libre arbitre. En m'appuyant sur ces avancées récentes et sur des travaux plus anciens dans la tradition anglo-saxonne, je défendrai *une* réponse possible à la question épistémologique. Cette réponse comportera, ultimement, à la fois une certaine théorie du libre arbitre (une théorie dite « libertarienne ») et une certaine théorie de la connaissance du libre arbitre (une théorie que j'appelle « introspectionniste »). Mais mon intention, par ce travail, n'est pas seulement de défendre une position particulière ; c'est aussi de fournir une « carte du territoire » des positions possibles, pour un débat qui jusqu'à présent était largement inexploré.

La question de la connaissance du libre arbitre peut paraître à première vue une question très locale et très spécialisée, peut-être moins importante que le traditionnel « problème du libre arbitre » : n'est-il pas plus important de déterminer si nous sommes libres que de déterminer comment nous le savons ? Pourtant, la conviction qui a guidé mes recherches est que la question épistémologique est aujourd'hui la question cruciale à se poser pour pouvoir répondre à la question métaphysique elle-même. C'est-à-dire que le débat sur le libre arbitre se trouve, à mon sens, à un « tournant épistémologique ». Un tournant épistémologique se produit, pour un problème philosophique donné, lorsque les philosophes, ne parvenant pas à trancher leur différend par des arguments sur la chose elle-même – par exemple, le monde est-il fini ou infini ? – passent à des questions de niveau épistémologique – que sommes-nous *justifiés* à penser à propos de l'infini du monde ? Comment pourrions-nous la *connaître* ? Dans certains cas, le tournant épistémologique entraîne un abandon de la question de départ : on peut penser ici à Kant, dont la « révolution copernicienne » est précisément un tournant épistémologique pour toute la métaphysique, ayant pour conséquence de renvoyer dos à dos toutes les thèses opposées des débats métaphysiques traditionnels (aussi bien la thèse de l'infini du monde que celle de la finitude du monde, par exemple). Mais dans d'autres cas, le tournant épistémologique peut se faire au profit d'une des thèses antérieurement débattues. C'est le cas à mon sens de la philosophie du libre arbitre, où le passage au niveau épistémologique peut permettre de mieux soutenir une thèse libertarienne. Aussi, mon ambition véritable dans ce travail n'est pas de traiter une question d'épistémologie, mais bel et bien de proposer une solution au problème traditionnel du libre arbitre. L'épistémologie n'est à mes yeux qu'un moyen, mais un moyen indispensable, pour parvenir à un tel résultat.

Pour pouvoir aborder la question de la connaissance du libre arbitre, il est indispensable de poser, en introduction, quelques bases pour chacun de ces deux domaines que sont la connaissance d'une part, et le libre arbitre d'autre part. Chacun constitue à soi seul un continent de la pensée philosophique ; il nous faut donc, avant de nous lancer dans l'aventure, quelques indications et définitions initiales.

I. Le libre arbitre : concept et problèmes

1. Difficultés pour définir le concept de « libre arbitre »

Commençons par le libre arbitre. Le « problème du libre arbitre et du déterminisme » est si connu que l'honnête homme a généralement une idée vague du sens de l'expression « libre arbitre ». Mais on ne peut se contenter de cette idée vague. Pour pouvoir fonder notre étude sur des bases solides, il nous faudrait une définition. Comment définir alors le concept de libre arbitre ?

Malheureusement, la tâche de définir le libre arbitre s'avère extrêmement délicate. Il existe en effet, dans la littérature sur le libre arbitre, des conceptions extrêmement variées de ce qu'est le libre arbitre : pour certains auteurs, le libre arbitre est une faculté qui nous permet de déterminer ou de sélectionner nos volontés, celles-ci étant auparavant indéterminées ; pour d'autres, le libre arbitre ne concerne pas la sélection (en amont) de nos volontés mais l'influence de nos volontés (ou de nos désirs) sur l'action qui s'ensuit (en aval) ; pour d'autres encore, le libre arbitre concerne la structure hiérarchique de nos désirs pris dans leur ensemble, en non pas l'effet de ces désirs dans l'action.¹ Parmi ces diverses conceptions, certaines sont telles que la description complète du fonctionnement du libre arbitre contient des indéterminismes causaux, et d'autres au contraire peuvent être décrites sans aucune forme d'indéterminisme. Autrement dit, les désaccords sur la nature du libre arbitre entraînent une division importante entre philosophes incompatibilistes (ceux pour qui le libre arbitre est incompatible avec le déterminisme) et philosophes compatibilistes (ceux pour qui les deux notions sont compatibles). Mais comment trancher un tel débat s'il y a un désaccord sur la définition même du libre arbitre.

Une solution classique en philosophie consiste à distinguer définition nominale et définition réelle : la définition nominale est ce qui permet de fixer la référence d'un mot, de déterminer de quelle chose on parle, et elle est donc un préalable indispensable à la discussion ; la définition réelle est ce qui décrit la nature ou l'essence d'une chose, et peut faire l'objet de désaccords. Par exemple, on peut très bien fixer la référence du mot « eau » sans avoir recours à la description de l'eau comme composée de molécules H₂O : il suffit pour cela de disposer d'une définition nominale, telle que « l'élément translucide qui coule dans les rivières et les robinets, et qui nous désaltère » ; la description H₂O ne vient que dans un deuxième temps, comme une définition réelle donnant la nature de la chose. De même, on peut considérer que la question du fonctionnement réel du libre arbitre, et notamment la question de savoir s'il engage des indéterminismes causaux, est une question qui concerne la nature profonde de la chose « libre arbitre » (sa définition réelle), et que l'on peut peut-être

¹ Nous aurons l'occasion de revenir abondamment sur ces diverses conceptions du libre arbitre, au chapitre 1 (voir p.43) et de manière plus détaillée au chapitre 4 (voir p.253). Les trois descriptions que je viens de donner de manière informelle et préparatoire correspondent respectivement aux conceptions *libertarienne*, *conditionnaliste*, et *hiérarchique* du libre arbitre.

donner une définition nominale qui fixe la chose dont on parle tout en laissant ouverte cette question du compatibilisme et de l'incompatibilisme. Les philosophes ont donné diverses définitions réelles du libre arbitre, mais cela n'est pas un problème si l'on peut trouver une définition nominale commune.

Malheureusement, lorsqu'on laisse de côté les définitions réelles du libre arbitre pour se tourner vers la question de la définition nominale, on trouve une diversité presque aussi grande parmi les philosophes. Un certain nombre de philosophes définissent nominalement le libre arbitre à partir de la responsabilité morale : « Le libre arbitre est souvent considéré comme la capacité, ou l'ensemble de capacités, requises pour remplir la condition de contrôle pour la responsabilité morale » (Timpe 2008, 10).² D'autres considèrent que le concept de libre arbitre est intrinsèquement, avant toute investigation sur la définition réelle, un concept modal, qui s'explicite nécessairement à l'aide du verbe « pouvoir » : « Quand je dis d'un homme qu'il 'a le libre arbitre', je veux dire que, bien souvent, sinon toujours, lorsqu'il doit choisir entre deux actions mutuellement incompatibles, [...] chacune des deux actions est telle qu'il peut [...] l'entreprendre » (van Inwagen 1983, 8). C'est la définition du libre arbitre comme « pouvoir des alternatives ». Il est clair que ces deux premières définitions nominales sont distinctes et indépendantes, dans la mesure où il n'est pas du tout évident que la responsabilité morale requière le pouvoir des alternatives.³ Selon la définition nominale qu'on adopte, cette question ouverte pourra s'exprimer de deux manières différentes : si le « libre arbitre » désigne la faculté (quelle qu'elle soit) requise par la responsabilité morale, alors ce sera une question ouverte de savoir si le libre arbitre implique le pouvoir des alternatives. Inversement, si le « libre arbitre » désigne, par définition, le pouvoir des alternatives, alors ce sera une question ouverte de savoir si la responsabilité morale requiert le libre arbitre. Il existe encore d'autres définitions nominales dans le débat sur le libre arbitre,⁴ mais elles ne font que multiplier le type de problèmes que nous venons de voir, à tel point qu'on peut légitimement se demander si les différents auteurs parlent finalement « de la même chose » lorsqu'ils parlent « du libre arbitre ».

A mon sens, ce problème de définition est un problème profond qu'on ne peut pas régler en introduction en « choisissant » simplement telle ou telle attitude. Je traiterai ce problème de manière détaillée au chapitre 1. En fin de compte, la réponse que j'apporte aura des conséquences importantes sur l'ensemble de mon argument. Je renvoie donc au corps de cette étude pour ce qui est de la définition pleinement satisfaisante du libre arbitre. Plus précisément, ma définition nominale occupera le chapitre 1, et pour ce qui est de la définition réelle, elle fera l'objet des chapitres 4 et 5. Cela veut-il dire que nous ne pouvons donner aucune notion préliminaire de ce dont nous allons parler ? Non heureusement. Je crois qu'il est possible, en introduction, de donner une idée de ce dont il s'agit, non pas par une définition en bonne et due forme, mais plutôt par un « récit » de la genèse du problème. C'est seulement au chapitre 1 que nous verrons pourquoi ce récit préliminaire était tout à fait adéquat pour présenter le concept de libre arbitre. Voici donc le récit qui permet de savoir de quoi il s'agit lorsque l'on utilise l'expression « libre arbitre ».

² Voir également (Pereboom 2001, xiv) ou encore (Smilansky 2000, 3).

³ Cette question, comme on le verra au chapitre 4 (sec. II.2), est devenu un des plus grands débats en philosophie du libre arbitre, suite à un article de Harry Frankfurt (1969) qui tendait à montrer que l'on peut très bien être moralement responsable même si l'on n'avait pas d'alternative possible.

⁴ La troisième définition nominale importante est celle de Robert Kane, non plus en termes de possibilités alternatives mais en termes d'origine ou de source de l'action : « the power of agents to be the ultimate creators (or originators) and sustainers of their own ends or purposes » (Kane 1996, 4). D'autres, comme Randy Clarke (2003, 3), utilisent une définition nominale « mixte », qui réunit les conditions de possibilités alternatives et de source de l'action à la Kane. Daniel Dennett, quant à lui, semble considérer que le débat sur le libre arbitre concerne (ou devrait concerner), par définition, « the sort of free will we all want » (Dennett 1984, 16).

« A diverses époques de l'histoire de la pensée, les hommes ont considéré la possibilité que toutes leurs actions soient déterminées par des facteurs qui les dépassent. Cette possibilité a pu se présenter à eux de différentes façons ; par exemple, dans le cadre de réflexions théologiques sur les décisions des Parques, ou au cours de réflexions philosophiques sur le Destin, ou encore au cours de réflexions tirées de l'étude de la nature qui suggèrent un déterminisme (total) par la société ou par les forces atomiques. Quelle que soit l'origine de ces considérations, elles ont eu à chaque fois un effet comparable : elles ont plongé ces penseurs dans une forme de perplexité : cette possibilité qui s'ouvrait à eux n'était pas en harmonie avec l'image qu'ils avaient auparavant d'eux-mêmes et de leurs actions. Par ailleurs, cette possibilité suscitait également en eux une inquiétude concernant la responsabilité morale : cette possibilité du déterminisme, en effet, n'était pas en harmonie avec l'idée très générale selon laquelle les hommes sont moralement responsables, une idée que les hommes n'auraient autrement jamais pensé à mettre en question. Ce genre d'hypothèses a donc suscité chez les penseurs, aux différentes époques où elles sont apparues, une double perplexité, concernant d'une part l'image qu'ils avaient d'eux-mêmes et de leurs actions, et d'autre part leur conception de la moralité. »

C'est cette double perplexité qu'on peut appeler le « problème du libre arbitre et du déterminisme » ou, plus brièvement, le « problème du libre arbitre ». Ce problème est à l'origine d'un débat quasiment aussi vieux que la réflexion philosophique, dont l'enjeu est de savoir si les hommes sont libres, ou bien s'ils sont déterminés. Et c'est donc à ce débat que j'entends apporter une contribution, après avoir rappelé brièvement quelques étapes importantes dans le développement historique des discussions.

2. Aux origines du débat contemporain

Le problème du libre arbitre a été débattu à diverses époques de l'histoire de la philosophie, notamment à l'époque hellénistique et à l'époque médiévale. Mais les termes du débat contemporain sont hérités essentiellement d'une tradition qui commence à la période moderne, avec Hobbes, Locke et Hume. Avant ces auteurs, on considère souvent (de manière simplifiée) que la discussion sur le libre arbitre consistait en une opposition frontale entre défenseurs du libre arbitre et défenseurs du déterminisme. Dans ce débat, que van Inwagen (1983, 2) appelle le « problème traditionnel », il est *présupposé* que le libre arbitre et le déterminisme sont incompatibles, et qu'il faut donc choisir entre les deux. C'est ce présupposé que les auteurs modernes anglo-saxons sont venus contester en défendant une conception « compatibiliste » du libre arbitre.⁵

Pourquoi cette apparition, à l'époque moderne, d'une conception compatibiliste de la liberté ? Une raison est sans doute à chercher dans le développement des sciences expérimentales. L'autorité que prennent les sciences à l'époque moderne a pour conséquence la prééminence d'une « vision scientifique du monde » dans laquelle toute la réalité doit tomber sous le coup de l'explication causale. Dans les termes du « problème traditionnel », cette prééminence de la vision scientifique du monde aurait pu entraîner un abandon total de l'idée de liberté au profit des conceptions déterministes. Mais il semble que l'être humain résiste fortement à abandonner totalement l'idée de liberté ; aussi, pour ne pas abandonner cette idée, il fallut bien l'accommoder au déterminisme. Dans un environnement intellectuel

⁵ En réalité, les débats hellénistiques et médiévaux étaient bien plus complexes. En particulier, les Stoïciens semblent avoir défendu une position qui est très proche du compatibilisme contemporain (voir en particulier Salles 2005). Mais les articulations essentielles du débat contemporain, celles dont j'aurai besoin pour cette étude, sont marquées principalement par l'influence directe des discussions de la période moderne. C'est pourquoi je ne présente pas ici l'histoire des débats hellénistiques et médiévaux.

où le déterminisme scientifique s'impose de manière presque unanime, les deux options qui s'affrontent ne sont plus « libre arbitre (indéterministe) » contre « déterminisme », mais plutôt « déterminisme sans libre arbitre » et « déterminisme avec libre arbitre » – ce que William James (1956) a appelé « déterminisme dur » et « déterminisme doux ». Certes, il existe dans la période moderne des auteurs qui défendent encore un libre arbitre indéterministe, mais cette position est de plus en plus perçue par les défenseurs de la vision scientifique du monde comme une position désespérée voire irrationnelle. Et entre déterminisme dur et compatibilisme, une grande majorité des philosophes s'est portée vers l'option subtile et ingénieuse qui permet de préserver la responsabilité morale.⁶

Ainsi, après Hobbes, Locke et Hume, le débat philosophique sur le libre arbitre devient essentiellement un débat sur la *compatibilité* entre déterminisme et libre arbitre, avec une majorité de philosophes qui était convaincue que le présupposé d'incompatibilité entre déterminisme et libre arbitre n'avait jamais reposé que sur des confusions conceptuelles grossières. Cette situation a perduré approximativement jusqu'aux années 1960.⁷ A partir des années 60 sont apparus des arguments extrêmement puissants (Ginet 1966; van Inwagen 1975) tendant à montrer que le libre arbitre était bel et bien incompatible avec le déterminisme. Ces arguments n'ont, bien sûr, pas convaincu la totalité des philosophes ; de fait, si l'on considère la totalité des philosophes (toutes spécialités confondues) la majorité d'entre eux encore aujourd'hui est, semble-t-il, compatibiliste (voir Bourget and Chalmers 2013) ; mais si l'on considère uniquement les philosophes spécialistes du libre arbitre – ceux qui ont pu consacrer toutes leurs énergies aux arguments *pro* et *contra* – il semble bien qu'une majorité soit devenue *incompatibiliste* à la suite des arguments de cette époque (voir van Inwagen 2008, 338). C'est donc un revirement considérable qui s'est amorcé dans les années 60. Un revirement qui a eu pour conséquence de remettre sur le devant de la scène le problème traditionnel et son alternative : libre arbitre ou déterminisme ?

Suite à ces diverses évolutions, la communauté philosophique se divise principalement en trois grandes familles : les nouveaux compatibilistes (qui ne peuvent plus se contenter de reprocher la « grossière confusion » incompatibiliste, mais doivent désormais répondre aux arguments incompatibilistes), les nouveaux incompatibilistes durs (qui ont été convaincus par les récents arguments incompatibilistes et ont refusé l'existence du libre arbitre), enfin une certaine résurgence des conceptions indéterministes du libre arbitre – ce qu'on appelle la position incompatibiliste « libertarienne ».

3. L'étrange discrédit de la position libertarienne

Malgré la résurgence de positions libertariennes que je viens de mentionner, il me semble qu'on peut dire que ce genre de thèses garde une très mauvaise réputation aux yeux de la majorité des philosophes : non seulement ils ne croient pas au libertarianisme, mais surtout ils la considèrent comme une option qui n'est pas sérieusement envisageable, ou qui est contraire à l'esprit scientifique. Au regard de l'évolution des débats au siècle dernier, ce discrédit persistant est étonnant et mérite une explication.

⁶ Les proportions sont peut-être inverses parmi les intellectuels *scientifiques* de la même période.

⁷ Un bon représentant de la conviction compatibiliste ferme et confiante que l'on pouvait avoir jusqu'au début des années 70s est Donald Davidson : « There are [...] those who believe they can see, or even prove, that freedom is inconsistent with the assumption that actions are causally determined, at least if the causes can be traced back to events outside the agent. I shall not be directly concerned with such arguments, since I know of none that is more than superficially plausible. Hobbes, Locke, Hume, Moore, Schlick, Ayer, Stevenson, and a host of others have done what can be done, or ought ever to have been needed, to remove the confusions that can make determinism seem to frustrate freedom. » (Davidson 1973, 67)

Ce discrédit est étonnant si l'on pense à la découverte de la mécanique quantique, qui est venue renverser l'idée selon laquelle la science fondamentale, notre meilleure conception scientifique de la nature, serait déterministe. En effet, si le déterminisme scientifique est le facteur qui est venu discréditer la position libertarienne à l'époque moderne, alors l'abandon de cette vision scientifique déterministe aurait dû réhabiliter le libertarianisme. Lorsqu'il existe deux options sérieuses dans un débat et qu'une objection vient rejeter l'une de ces deux options, si l'objection tombe, alors normalement la position qu'elle rejetait redevient une option sérieuse. Or les choses ne se sont pas vraiment passées ainsi pour la position libertarienne. Cette position est très loin d'avoir retrouvé le prestige qu'elle pouvait avoir dans la période médiévale ou dans l'antiquité avec Epicure. Tout se passe comme si le débat avait évolué de la manière suivante :

- Etat 1 : le libre arbitre indéterministe est une des deux visions de la réalité tout à fait envisageables.
- Etat 2 : en fait, la science montre que la nature est déterministe, donc le libre arbitre indéterministe n'est plus envisageable.
- Etat 3 : finalement, la science ne montre pas que la nature soit déterministe.
- Etat 4 : ... mais vous n'avez toujours pas avancé de raison sérieuse de croire au libre arbitre indéterministe.

S'il y avait des raisons sérieuses de croire au libre arbitre indéterministe *avant* l'objection déterministe, pourquoi ces raisons ne jouent-elles pas le même rôle une fois que cette objection est dépassée et rejetée ? Pourquoi la remise en cause du déterminisme scientifique ne nous a-t-elle pas ramené à l'état antérieur à cette objection, à « l'état épicurien » de la discussion, pourrait-on dire ? La présente étude peut être lue comme un effort pour replacer le débat du libre arbitre dans son « état épicurien ». Mais avant de critiquer la situation présente du débat, encore faut-il comprendre les raisons pour lesquelles nous en sommes arrivés là.

Une première raison pour laquelle le libertarianisme n'a pas repris sa place tient certainement au développement du compatibilisme : nous avons vu que l'influence croissante du déterminisme scientifique n'avait pas entraîné un rejet total de l'idée de liberté, mais plutôt une évolution du concept de liberté pour rendre celui-ci compatible avec la vision scientifique déterministe du monde. En particulier, les auteurs compatibilistes se sont attachés à montrer qu'on avait eu tort, dans les siècles précédents, de présupposer que le libre arbitre requerrait l'indéterminisme, et que l'indéterminisme pouvait « aider » à être libre. Ces remises en cause, qui avaient initialement pour but de « sauver » la liberté contre la pression du déterminisme scientifique, ont également constitué par la suite des objections à la position libertarienne. L'évolution du débat serait donc la suivante :

- Etat 1 : le libre arbitre indéterministe est une des deux visions de la réalité tout à fait envisageables.
- Etat 2 : en fait, la science montre que la nature est déterministe, donc le libre arbitre indéterministe n'est plus envisageable.
- Evolution : mais contrairement à ce qu'on pensait à l'état 1, le déterminisme ne menace pas, en réalité, le libre arbitre, et même l'indéterminisme ne saurait « aider » à être libre.
- Etat 3 : finalement, la science ne montre pas que la nature soit déterministe.
- Etat 4 : ... mais il n'en demeure pas moins que l'indéterminisme (qui existe peut-être) ne peut en rien aider le libre arbitre.

Autrement dit, une fois que l'objection déterministe a été rejetée par la science elle-même, la position libertarienne s'est trouvée confrontée à deux nouvelles objections, indépendantes de l'objection déterministe, mais auxquelles on n'avait pas vraiment (ou pas autant) songé avant l'apparition de l'objection déterministe. On peut utiliser la terminologie de Robert Kane (1996, 13) pour désigner ces deux nouvelles objections :

- La question de compatibilité : pourquoi le déterminisme menacerait-il le libre arbitre ?
- La question d'intelligibilité : en quoi l'ajout d'indéterminismes dans la nature pourrait-il aider à rendre intelligible le libre arbitre ?⁸

Ces deux défis ont continué, pendant quelques décennies, à convaincre une majorité de philosophes qu'une conception indéterministe du libre arbitre n'était pas sérieusement envisageable. Pourtant, nous avons vu aussi qu'à partir des années 60 de très forts arguments incompatibilistes ont été avancés, relevant de manière convaincante (sinon de manière absolument contraignante) le défi de compatibilité. Le « présupposé », qui prévalait à l'origine de la discussion, selon lequel une action totalement déterminée ne serait pas vraiment libre, n'était sans doute pas un pur préjugé, mais repose au contraire sur des intuitions très fortes.

Restait le défi de l'intelligibilité : si une action déterminée n'est pas libre, est-ce qu'une action *indéterminée* le serait vraiment davantage ? Peut-on vraiment comprendre que l'ajout d'indéterminismes augmente le contrôle libre de l'action ? Ce deuxième défi a été considéré quelques années plus tard par les auteurs libertariens, et principalement par Robert Kane à partir des années 1980. Il s'agissait alors de donner un modèle intelligible du libre arbitre avec indéterminismes. Les débats se sont poursuivis au cours des années 2000 avec l'apparition d'autres modèles à côté de celui proposé par Kane.⁹ On peut dire qu'à l'issue de ces discussions, ce défi a été relevé de manière satisfaisante : bien sûr, les modèles proposés n'ont pas convaincu tous les philosophes, mais un certain nombre de philosophes non libertariens reconnaissent aujourd'hui l'intelligibilité de l'un ou l'autre de ces modèles (Derk Pereboom, Mark Balaguer), et surtout il me semble que la communauté philosophique a été largement convaincue que cette tâche de modéliser le libre arbitre indéterministe pouvait être entreprise de manière sérieuse et rationnelle, et qu'il ne s'agit pas d'une option désespérée. Du point de vue de la crédibilité ou de la rationalité du libertarianisme, le défi de l'intelligibilité ne devrait donc plus constituer un obstacle majeur.

Malgré tout, le libertarianisme garde pour un certain nombre de philosophes le caractère peu enviable d'une option non envisageable ou non scientifique, et n'a toujours pas retrouvé le crédit qu'il pouvait avoir avant la période moderne. Pourquoi ? Maintenant que l'objection déterministe est tombée, maintenant qu'on a apporté des réponses défendables et sérieuses aux deux défis issus du revirement compatibiliste, qu'est-ce qui empêche encore la réhabilitation du libre arbitre indéterministe ? Pourquoi le libertarianisme est-il toujours confronté à la réaction suivante : « il se peut que l'indéterminisme soit possible, et soit requis de manière intelligible pour un certain genre de liberté, mais vous n'avez toujours pas avancé de raison sérieuse qui indiquerait que nous avons effectivement un tel type de liberté ». La réponse à notre question se trouve sans doute dans l'idée de « raison sérieuse ». Que serait, pour l'objecteur, une « raison sérieuse » de croire que nous avons un libre arbitre indéterministe ? Ce serait probablement une observation scientifique directe de l'indéterminisme du libre arbitre. Pourquoi cette demande de raison sérieuse ne posait-elle pas obstacle au libertarianisme avant la période moderne ? Parce qu'on avait une autre conception de ce qui pourrait compter comme « raison sérieuse » de croire positivement au libertarianisme. En particulier, les philosophes antiques et médiévaux semblaient considérer

⁸ Robert Kane (*ibid.*) parle également du problème de la « montée » et du problème de la « descente » de la « montagne libertarienne » : la montée consiste à élever les exigences ou les conditions de possibilité du libre arbitre de telle sorte que ces conditions impliquent l'indéterminisme causal ; la descente consiste à faire redescendre cette exigence indéterministe dans la réalité telle que nous la connaissons, pour donner une vision intelligible du libre arbitre.

⁹ Je pense en particulier au modèle de la « causalité agentive » que nous aurons l'occasion d'approfondir au chapitre 5 (sec. I.1.b.), avant d'envisager le modèle de Robert Kane (sec. I.1.c.).

comme tout à fait acceptables certaines données de sens commun en faveur du libre arbitre, et du libre arbitre indéterministe plus précisément.

Nous tenons donc probablement la réponse à notre question de départ : pourquoi le libertarianisme n'est-il pas réhabilité une fois que la grande objection moderne du déterminisme a été rejetée ? C'est parce que la science moderne n'a pas simplement changé notre vision du monde, elle a aussi changé notre conception de ce qui doit compter comme « raison sérieuse » d'adopter telle ou telle croyance. Autrement dit, elle a soulevé contre le libertarianisme pré-moderne une nouvelle objection qui n'est plus d'ordre métaphysique (comment le libre arbitre indéterministe pourrait-il s'intégrer dans la réalité ?) mais qui est d'ordre *épistémologique* : les données de sens commun sur lesquelles reposait la croyance libertarienne ne peuvent pas constituer, en fait, un bon fondement rationnel. Si ce diagnostic est exact, alors l'épistémologie se trouve au cœur même du débat sur le libre arbitre ; le problème qui permet de comprendre un élément fondamental des positions actuelles est un problème épistémologique.

La présente étude sera tout particulièrement consacrée à ce problème particulier. Ultimement, je défendrai que ce problème peut être résolu – ce qui veut dire que le libertarianisme *devrait* retrouver sa crédibilité pré-moderne. Mais avant de le résoudre, je consacrerai beaucoup d'efforts à le formuler de manière adéquate, car jusqu'à ce jour, il n'a été que très rarement exposé de manière explicite. Dans la section suivante, je vais donner à ce problème le cadre formel qui lui convient le mieux, à savoir celui d'un « Défi d'Intégration ».

4. Le problème épistémologique : un Défi d'Intégration

Un des rares auteurs anti-libertariens qui ait explicitement mentionné le problème épistémologique comme sa motivation principale pour refuser le libertarianisme est John M. Fischer. Voici la manière dont il le formule.

Nous ne savons pas si le déterminisme causal est vrai ou non. Bien que certains physiciens doutent explicitement qu'il soit vrai, d'autres estiment que, en fin de compte, les indéterminations apparentes posées par la mécanique quantique s'avèreront n'avoir été que de simples indéterminations épistémiques (des trous dans notre connaissance). [...] Etant donné que nous ne savons pas avec certitude que le déterminisme causal est faux, il devrait s'ensuivre, semble-t-il, [si la conception indéterministe du libre arbitre était la bonne] que nous ne pouvons pas être (légitimement) sûrs de notre statut d'agents libres et moralement responsables.
(Fischer 2006, 5)

Un libertarien, semble-t-il, doit affirmer qu'il sait dans son fauteuil que le déterminisme causal est faux ; mais comment pourrions-nous savoir à l'avance une thèse aussi empirique ? C'est là une vertu importante du [compatibilisme]¹⁰ : le défenseur de cette

¹⁰ Pour être exact, Fischer ne défend pas le *compatibilisme*, mais le *semi-compatibilisme* ; c'est-à-dire qu'il distingue deux notions de « liberté » ou de « contrôle », celle requise par la responsabilité morale, et la conception des alternatives, et considère que seule la première est compatible avec le déterminisme. Nous aurons l'occasion, au chapitre 4 (sec. II.2.), de discuter avec précision la thèse semi-compatibiliste. Mais ce qui nous intéresse pour l'instant est seulement la structure de l'argument que donne Fischer pour *une* notion de liberté. Le fait qu'il ne l'emploie pas pour *une autre* notion de liberté est une subtilité de sa doctrine qui n'affecte pas la nature de l'argument. Pour l'autre notion de liberté, celle des alternatives, Fischer adopte plutôt le raisonnement « sceptique » que nous allons voir dans quelques paragraphes, parce que son intuition incompatibiliste est trop forte pour cette notion.

doctrine n'a pas besoin de prétendre qu'il sait *a priori* une thèse (vraisemblablement) empirique de physique. [...]

Une des vertus principales du compatibilisme est que nos conceptions les plus profondes et basiques de notre action – notre liberté et notre responsabilité morale – ne sont pas tenues en otage par les données de la physique.

(Fischer in Fischer et al. 2007, 81)

Dans ces citations, Fischer exprime clairement que ce qui l'empêche d'être libertarien est, d'abord et avant tout, une intuition épistémologique que l'on peut formuler de la manière suivante :

(IE) Les données préscientifiques qui fondent notre croyance au libre arbitre ne pourraient pas nous apporter légitimement une connaissance du libre arbitre indéterministe.

ou pour le dire sous la forme d'un conditionnel :

(IE') Si le libre arbitre impliquait des indéterminismes causaux, alors les données préscientifiques qui fondent notre croyance ne seraient pas des sources légitimes de connaissance (mais de pures illusions).

Or Fischer exprime nettement qu'il ne peut pas se résoudre à ce que notre conception de nous-mêmes comme libres (et moralement responsables) soit une pure illusion. C'est-à-dire qu'il semble faire, à partir du conditionnel (IE'), le *modus tollens* suivant :

(IM) Or notre conception de nous-mêmes comme libres n'est pas une pure illusion mais repose sur des données (préscientifiques) qui donnent une vraie connaissance.¹¹

(C) Donc le libre arbitre n'implique pas d'indéterminismes causaux.¹²

¹¹ Le nom (IM) de cette prémisse signifie « Intuition Mooréenne » : comme nous le verrons dans la suite de cette section, puis au chapitre 3, cette appellation repose sur le fait que cette intuition est de la forme « nous savons que *p* », c'est-à-dire qu'il s'agit d'un type d'intuition de niveau directement épistémique, auxquelles G.E. Moore a donné un statut très particulier (voir ch.3, note 50).

¹² Il y a d'autres manières d'interpréter le propos de Fischer : à première vue, Fischer ne présente pas l'inconnaissabilité du libre arbitre comme une conséquence (du libertarisme) qui est *fausse*, mais plutôt comme une conséquence qui est *malencontreuse* ou *indésirable*. Ainsi, Fischer n'aurait pas tant l'intuition (IM) « nous savons que nous sommes libres », mais plutôt le *desideratum* (DM) « il serait bon de trouver une théorie de la liberté qui permette de dire que nous savons que sommes libres ». Si Fischer s'en tenait là, alors son argument n'apporterait pas vraiment une raison de croire au compatibilisme, mais plutôt une raison (pratique) de rechercher des raisons (théoriques) de croire au compatibilisme. On pourrait appeler cela un argument de « motivation de recherche ». Dans un des endroits où Fischer exprime ses motivations compatibilistes, il semble aller dans le sens de la simple « motivation de recherche » (« Given the motivation of seeking *resiliency* of our fundamental conception of ourselves as possessing control and being morally responsible agents, I have addressed the challenges posed by causal determinism. » (Fischer 2006, 6)). Dans d'autres endroits, les formulations de Fischer donnent le sentiment que son objection fondamentale est en fait une « objection d'intelligibilité », tendant à montrer qu'on ne peut comprendre que l'ajout d'un indéterminisme soit en tant que tel un facteur donnant le libre arbitre (« A compatibilist can capture the intuitive idea that a tiny difference (between 100 percent and 99.9 percent) should not make such a huge difference (between our having more than one genuinely available pathway into the future and this pervasive phenomenological fact being just a big delusion). How could such a small change – of the sort envisaged in the probabilities associated with the arcane conditionals of theoretical physics – make such a big difference? » (Fischer et al. 2007, 46)). Je doute qu'il y ait « une » bonne interprétation des considérations que Fischer a réunies pour exprimer « sa » motivation en faveur du compatibilisme. Je crois plutôt qu'il y a plusieurs considérations, l'une d'entre elles étant l'argument épistémologique que je viens de reconstruire.

Ce qui est intéressant, dans cette formulation du problème, c'est qu'elle suggère immédiatement une réaction possible : là où Fischer croit fermement à la connaissabilité (préscientifique) du libre arbitre et fait le *modus tollens* (contre la conception incompatibiliste), d'autres auteurs estimeront que l'intuition incompatibiliste est une intuition analytique beaucoup trop forte et seront plutôt amenés à faire le *modus ponens* que Fischer essaye à tout prix d'éviter. C'est-à-dire que certains auteurs raisonneront ainsi :

(IE') Si le libre arbitre implique des indéterminismes causaux, alors notre croyance préscientifique au libre arbitre est une pure illusion.

(IA) Or le libre arbitre implique des indéterminismes causaux.

(C) = ($\sim$ IM) Notre croyance préscientifique au libre arbitre est une pure illusion.

La conclusion d'un tel raisonnement sera donc une conclusion incompatibiliste sceptique (typiquement incompatibiliste dure) concernant le libre arbitre.¹³

On pourrait interpréter ces deux arguments contraires comme un simple conflit d'intuitions : certains auteurs auraient l'intuition (IM) d'une connaissance du libre arbitre *sans avoir* l'intuition (IA) en faveur de l'incompatibilisme ; d'autres auteurs auraient l'intuition incompatibiliste (IA) *sans avoir* aucune intuition (IM) en faveur de notre connaissance préscientifique du libre arbitre. En réalité, il me semble plus juste de dire que nous trouvons assez naturellement en nous-mêmes l'une et l'autre intuitions, et que le différend entre ces auteurs est un différend sur l'intuition qu'il convient d'*abandonner* au profit de l'autre lorsqu'elles entrent en conflit. Fischer est prêt à abandonner toute intuition incompatibiliste parce qu'il préfère sauver la connaissance du libre arbitre. Un auteur incompatibiliste dur considèrera au contraire que l'intuition de connaissance doit céder à la force de l'intuition analytique. Autrement dit, le problème épistémologique serait un dilemme, dans lequel nous devons abandonner soit l'intuition (IM) soit l'intuition (IA).

Mais il existe manifestement une troisième possibilité pour résoudre le conflit d'intuition. Les deux arguments reposent sur une troisième prémisse, à savoir l'intuition épistémologique (IE) dont nous avons vu dans la section précédente qu'il s'agissait d'un défi apparu à la modernité mais qui n'est peut-être pas absolument incontestable. Il semble donc qu'on pourrait raisonner ainsi :

(IM) Notre croyance préscientifique au libre arbitre est une source de connaissance.

(IA) Le libre arbitre implique des indéterminismes causaux.

(C) = ($\sim$ IE) Donc notre croyance préscientifique au libre arbitre nous permet de connaître légitimement (d'une manière ou d'une autre) les indéterminismes causaux.

Cette conclusion peut paraître très étonnante, peut-être même inacceptable à une oreille de moderne. Tout l'objet de cette étude sera de voir si elle est véritablement inacceptable. Mais ce qui m'intéresse en introduction, c'est simplement de faire remarquer que le problème épistémologique du libre arbitre a, en réalité, la structure d'un *trilemme*, c'est-à-dire d'un conflit entre *trois* intuitions que nous avons, ou pouvons avoir, assez facilement : une intuition analytique (IA), une intuition mooréenne (IM) et une intuition épistémologique (IE).

¹³ Pour être précis, le défenseur de cet argument n'est pas obligé de soutenir que *nous ne sommes pas libres*, mais seulement que *nous ne le savons pas*. Par conséquent, il n'est pas obligé d'avoir une position *incompatibiliste dure*, mais pourrait soutenir une position que j'appellerai, au chapitre 3, « l'empirisme libertarien », à savoir : « il se pourrait que nous soyons libres de manière libertarienne, mais notre conviction préscientifique ne nous justifie pas à le croire ; pour trancher cette question, il faut attendre les résultats futurs des sciences ».

Ce type de trilemme n'est pas propre au libre arbitre : il appartient à une catégorie de problèmes que Christopher Peacocke (1999) a très clairement caractérisée et étudiée, à savoir les « Défis d'Intégration » :

Dans un grand nombre de domaines de la philosophie, nous rencontrons un problème commun. Il s'agit d'un problème de réconciliation. Nous avons à réconcilier une théorie plausible de ce qu'engage la vérité des énoncés d'un certain type avec une théorie crédible de la manière dont nous pouvons connaître ces énoncés, quand nous les connaissons. [...]

La tâche générale consistant à fournir simultanément, pour un domaine donné, une métaphysique et une épistémologie acceptables, et de montrer qu'elles le sont, je l'appelle le *Défi d'Intégration* [Integration Challenge] de ce domaine.

(Peacocke 1999, 1)

L'exemple le plus célèbre d'un Défi d'Intégration en philosophie contemporaine est l'article de Paul Benacerraf (1965) à propos des nombres et de la connaissance mathématique. Un des arguments centraux de l'article de Benacerraf consiste à montrer qu'une certaine théorie des énoncés mathématiques (une certaine *métaphysique* des nombres comme objets abstraits), si on l'accompagne d'une épistémologie causaliste (une certaine théorie sur la manière dont nous acquérons des connaissances), aurait pour conséquence que nous ne pourrions pas connaître la vérité des énoncés arithmétiques.¹⁴ Il faut donc abandonner soit la métaphysique des nombres qui est en cause, soit l'épistémologie causaliste... soit la prétention de connaître les énoncés arithmétiques – cette solution sceptique étant peu attirante dans le cas de l'arithmétique. Peacocke reconnaît clairement le problème de Benacerraf comme « archétype » des Défis d'Intégration : « Le Défi d'Intégration est la généralisation, à un domaine arbitraire, du défi que Paul Benacerraf a si finement identifié et discuté à propos de la vérité mathématique » (*ibid.*). Et de fait, ce problème se généralise aisément à un grand nombre de domaines philosophiques.¹⁵

Ce qui nous intéresse ici, c'est de garder la structure générale de ces problèmes, et de considérer les solutions possibles. De manière générale, un problème d'Intégration est un conflit entre trois types d'intuitions :

- (IM) une intuition mooréenne (ou épistémologique particulière) sur un domaine d'objets. Par exemple : que « la croyance ' $2+2=4$ ' constitue une connaissance », ou que « la croyance 'je suis libre' constitue une connaissance ».

¹⁴ Le problème peut être résumé ainsi : si connaître un objet suppose d'être affecté causalement par lui, et si la causalité ne peut mettre en relation que des objets *concrets*, alors les nombres comme objets abstraits ne pourront faire l'objet de connaissance.

¹⁵ Dans *Being Known*, Peacocke applique ce problème général à quatre domaines principaux : le passé, la modalité, la connaissance de soi, et ... la liberté. Le même problème a été identifié un peu plus tôt par (Pust 2000, sec. 5.5) sous le nom de « recherche du lien empiriquement étudiable ». Joël Pust l'applique à l'épistémologie des *intuitions* philosophiques : « The source of this concern in its many variants is the fact that empirical investigation fails to reveal to us *how* our intuitions could in any way depend upon their apparent truth conditions ». Pust montre que ce problème à propos des intuitions philosophiques est du même genre que le problème de Benacerraf, et il le généralise ensuite à toutes les croyances qui portent sur des objets abstraits. Les objets abstraits posent un problème d'Intégration particulier parce que nos croyances ne peuvent pas y être reliées par un lien causal. Cependant, Pust ne voit pas que le problème peut être généralisé au-delà des objets abstraits : même pour un objet très concret, nous pouvons être confrontés à un problème de dépendance entre l'objet et notre croyance. Seul Peacocke a caractérisé ce problème dans toute sa généralité.

- (IA) une intuition analytique sur ce domaine d'objets. Par exemple : que 'les nombres (2 ou 4) sont des objets abstraits', ou que 'être libre implique l'indéterminisme'.
- (IE) une intuition épistémologique générale sur les conditions de possibilité de la connaissance. Par exemple : que « connaître un objet *O* suppose d'avoir un rapport causal avec *O* », ou que « connaître l'indéterminisme supposerait de le connaître comme résultat de la méthode d'investigation scientifique ».

Peacocke (1999, 7–11) distingue sept types de réponses possibles face aux défis d'intégration. Il me semble qu'on peut plus simplement les réduire à trois¹⁶ :

- abandonner l'intuition analytique (IA), c'est-à-dire « réviser notre métaphysique » (solution (1) de Peacocke) : en donnant à l'objet considéré des conditions métaphysiques moins exigeantes, il devient plus facilement « intégrable » avec notre théorie de la connaissance. Par exemple, abandonner la théorie des nombres comme objets abstraits.
- abandonner l'intuition d'inconnaissabilité (IE), soit en « révisant notre épistémologie » (solution (2) de Peacocke), soit en montrant que notre épistémologie de départ n'est pas en fait incompatible avec la connaissance de l'objet considéré (solution (3) de Peacocke). Par exemple pour le problème de Benacerraf, abandonner l'épistémologie causaliste, ou soutenir qu'un objet abstrait peut entrer dans une relation causale, et donc être connu même dans une épistémologie causaliste.
- abandonner l'intuition (IM), c'est-à-dire nier « la possibilité d'une connaissance significative dans le domaine en question » (solution (6) de Peacocke). C'est la solution sceptique. Par exemple, nier que nous puissions *connaître* une vérité telle que « $2 + 2 = 4$ ». ¹⁷

¹⁶ Au chapitre 3 (p.239 sqq), je donnerai une classification un peu plus fine des solutions possibles au défi d'intégration. Pour l'instant ces détails ne sont pas utiles.

¹⁷ Les trois autres « solutions », ou plutôt « réactions » envisagées par Peacocke sont les suivantes. Une solution « révisionniste » (4) qui consiste à « donner des conditions de vérité au rabais, que l'on présente comme rendant compte d'une partie, mais pas de l'intégralité, du contenu intuitif des phrases problématiques ». A mon sens, le révisionnisme n'est pas une « solution » à part entière : sur le fond, le révisionnisme part du principe que l'intégration du contenu métaphysique (complet) est impossible ; la réponse au problème de fond reste donc la réponse sceptique (6). Le révisionnisme répond seulement à une autre question que la question de fond, à savoir : « comment convient-il de parler une fois que, sur le fond, on adopte le scepticisme ». Je parlerai davantage du révisionnisme au chapitre 4 (sec. II.5.). Peacocke présente ensuite une solution « pragmatiste » (5) qui « plutôt que de rechercher une intégration des conditions de vérité avec l'épistémologie, [...] assigne à ces énoncés un certain rôle dans la pensée et dans l'action, un rôle qui explique leur importance pour nous ». Là aussi, il me paraît clair que le pragmatisme n'est pas, en soi, une réponse au problème d'intégration mais répond à une question toute différente. Typiquement, un auteur pourra se reporter sur les questions pragmatiques s'il est convaincu que, au niveau théorique, aucune intégration n'est possible – c'est-à-dire que sa réaction pragmatiste reposera en fait, sur le fond, sur une réponse sceptique : « les conditions de vérité de cet énoncé ne sont pas connaissables, mais elles ont tout de même un certain rôle important ». L'auteur peut aussi devenir pragmatiste s'il estime que les énoncés problématiques, au lieu d'avoir des conditions de vérité inconnaissables, n'ont *pas du tout* de conditions de vérité identifiables. C'est la solution (7) envisagée par Peacocke : « déclarer que tout le domaine pour lequel l'intégration est en question est fallacieux et fondé sur des illusions d'une si profonde incompréhension que même les solutions (4), (5) et (6) ne sont pas disponibles ». De nouveau, il ne me semble pas qu'il y ait là à proprement parler une solution au problème d'intégration : cette réaction consiste plutôt à dire qu'il n'y a jamais eu de problème d'intégration, puisqu'il n'y a un tel problème que lorsque des conditions métaphysiques *déterminées* entrent en conflit avec une épistémologie *déterminée*. En tout cas, mon propos dans cette étude repose sur l'idée qu'il existe un problème d'intégration déterminé pour le libre arbitre, et par conséquent, je laisserai de côté cette réaction « confusionniste ». Peacocke estime que Galen Strawson a adopté cette réaction pour l'intégration du libre arbitre ; mais, comme on le verra au chapitre 5 (sec. I.2.), l'argument de

Ces trois types de solutions sont également disponibles pour le problème d'intégration du libre arbitre : solution métaphysique (rejeter l'analyse incompatibiliste du libre arbitre), solution épistémologique (accepter un modèle épistémologique de la connaissance du libre arbitre indéterministe) et solution sceptique (le libre arbitre est une chose trop complexe pour être connue préscientifiquement ; la réponse est « prise en otage » par les résultats futurs des sciences empiriques).

Comprendre le problème épistémologique du libre arbitre comme un Défi d'Intégration présente deux intérêts considérables. Premièrement, cela permet de prendre conscience que le problème épistémologique n'est pas seulement un « problème de libertarien », contrairement à ce que pourrait faire croire la section précédente (où l'on a vu que l'intuition (IE) se dressait comme une ultime objection au libertarianisme). Le Défi d'Intégration concerne tout théoricien du libre arbitre : quelle que soit la conception que l'on défende du libre arbitre, il faudra donner pour cette conception une épistémologie adaptée, ou à défaut accepter la conséquence sceptique. Même si l'on s'attache au problème particulier de l'inconnaissabilité de l'indéterminisme – problème de l'intuition (IE) – ce problème ne surgit pas seulement pour les auteurs libertariens ou incompatibilistes, mais pour toute personne qui est capable de ressentir, à un degré ou à un autre, l'intuition incompatibiliste (IA), quel que soit son jugement final sur le sort à réserver à cette intuition (la conserver ou la rejeter à cause d'autres intuitions plus fortes). Un auteur compatibiliste comme Fischer peut très bien partager initialement l'intuition (IA) et ne la rejeter que sur la base du même problème d'intégration que le libertarien entend résoudre en rejetant (IE). Nous verrons même au chapitre 4 (sec. IV.3.d.) que l'on peut sans doute repérer de manière universelle une intuition incompatibiliste. Donc le conflit d'intuition entre (IA), (IE) et (IM) est très probablement un problème épistémologique qui concerne tous les auteurs, et pas seulement les libertariens. Deuxième avantage de la présentation sous forme de Défi d'Intégration : cette présentation permet de bien mettre en évidence la diversité des solutions que l'on peut apporter. Tant que le problème épistémologique est présenté comme un argument anti-libertarien, on peut avoir le sentiment que les intuitions incompatibiliste (IA) et d'inconnaissabilité (IE) sont nécessairement « les prémisses » et que ce problème doit mener, d'une manière ou d'une autre, au scepticisme ($\sim$ IM). Mais la présentation que nous venons de donner rappelle que nous avons également une intuition de connaissance (IM) et que le problème est donc fondamentalement un trilemme, que l'on peut prendre dans trois sens différents pour arriver à trois solutions différentes. Bien sûr, il ne s'agira de « choisir » de manière arbitraire quelle proposition on préfère rejeter, et nous aborderons au chapitre 3 (p.239 sqq) la question difficile de savoir quelle solution est la mieux justifiée. Mais à ce stade, nous ne pouvons exclure aucune des trois solutions, et ce trilemme pourrait donc aboutir aussi bien à une défense du scepticisme ($\sim$ IM), que du compatibilisme ($\sim$ IA), ou que du libertarianisme ($\sim$ IE).

En somme, la présentation du problème épistémologique comme un Défi d'Intégration nous rappelle qu'il ne s'agit pas d'un problème local, qui ne concernerait que les auteurs intéressés spécifiquement par la théorie de la connaissance, mais que l'angle épistémologique n'est qu'une manière d'aborder le débat fondamental, le débat métaphysique, du libre arbitre. Si le problème du libre arbitre exige un traitement épistémologique, il nous faut donc nous mettre au clair également sur les concepts épistémologiques que nous utiliserons.

G. Strawson peut, beaucoup plus simplement, s'interpréter comme établissant la conclusion « sceptique » : les conditions de vérité du libre arbitre sont contradictoires et donc impossibles, et par conséquent notre croyance préscientifique au libre arbitre est une pure illusion.

II. Connaissance et Justification : définitions

Comme on vient de le voir, la question qui nous occupera principalement dans cette étude sera la question de savoir si les données préscientifiques dont nous disposons nous donnent une connaissance du libre arbitre, ou si elles nous justifient à croire au libre arbitre. On peut donc donner, essentiellement, deux formulations de cet ensemble de questions :

(K) Pouvons-nous *savoir* que nous avons le libre arbitre ? (et si oui, comment ?)

(J) La croyance au libre arbitre peut-elle être (épistémiquement) *justifiée* ? (et si oui, comment ?)

A première vue, on pourrait penser que les deux formulations sont à peu près équivalentes, et que les termes de Connaissance et de Justification sont à peu près substituables.¹⁸ Il est en tout cas incontestable que les deux termes sont étroitement liés. Tout d'abord, il faut noter que l'un et l'autre sont des termes qui servent à *évaluer des croyances* ; à ce titre, ils font partie d'une longue liste de « termes d'évaluation épistémique »¹⁹, tels que « être raisonnable (ou déraisonnable) », « être rationnel (ou irrationnel) », « être défendable (ou indéfendable) », etc. Tous ces termes (ou plutôt ces couples de termes) sont *normatifs* au sens ils nous permettent de porter un jugement *positif* (ou négatif) sur une croyance, selon que cette croyance *satisfait* (ou ne satisfait pas) à certaines normes ou à certains objectifs :

Dire qu'une croyance est [une *connaissance*] ou qu'elle est *justifiée* pour une personne, c'est évaluer cette croyance ou cette personne (ou les deux) de manière *positive* ; le fait d'avoir cette croyance est *correct*, ou *approprié*, ou *acceptable*, ou *louable* ou *conforme aux standards*. (Plantinga 1993, 3)²⁰

Pourtant, s'il est incontestable que les deux concepts sont étroitement associés, l'épistémologie contemporaine a fait apparaître qu'ils peuvent néanmoins être disjoints et qu'ils semblent correspondre à deux pôles différents de l'évaluation épistémique. Au moins en première approche, il semble qu'on puisse établir le contraste suivant. Lorsqu'on dit que la croyance d'un sujet S est une *connaissance* (lorsqu'on dit que S *sait* que *p*, ou *est au courant* que *p*), on pense principalement au canal par lequel S a reçu l'information : qui lui a dit ? comment la nouvelle est-elle parvenue à ses oreilles ? Lorsqu'on dit que la croyance de S est *justifiée*, on pense plutôt à la manière dont S a *traité* l'information reçue, comment sa croyance est étayée (ou au contraire menacée) à l'intérieur même de son propre système cognitif. En ce sens, on peut dire que la notion de justification tend davantage vers un pôle « *internaliste* » de l'évaluation épistémique, et la notion de connaissance vers un pôle « *externaliste* ». Comme toujours en philosophie, il est difficile de trouver un consensus absolument universel, et il existe donc des théories importantes qui contestent l'internalisme de la justification²¹ ou l'externalisme de la connaissance²². Mais il semble au moins y avoir un

¹⁸ A condition bien sûr de préciser qu'il s'agit ici du concept « épistémologique » de Justification, et non pas de la justification morale, ou juridique, ou théologique.

¹⁹ L'expression « term of epistemic appraisal » est de Roderick Chisholm (1989, 8). Voir également (Goldman 1979, 90) et (Plantinga 1993, 3).

²⁰ Ce qui est ici traduit par « connaissance » est, dans le texte de Plantinga, le mot « warrant » (« garantie »). Comme il s'agit d'un terme technique désignant « ce qui fait d'une croyance vraie une *connaissance* », il est clair que le couple « warrant » / « justification » correspond exactement à notre couple « connaissance » / « croyance justifiée ». J'ai préféré éviter d'introduire ici la terminologie plantingienne qui n'est pas universelle.

²¹ Une défense récente de l'externalisme de la justification est celle de (Bergmann 2006).

²² Les *Méditations Métaphysiques* de Descartes offrent certainement une des théories les plus internalistes de la connaissance, puisque tout l'édifice de la connaissance y est fondé ultimement sur l'introspection et la raison.

consensus assez large sur le fait que « les alternatives plausibles ne sont pas [...] un externalisme pur (ou absolu) de la justification et un internalisme pur (ou absolu) de la connaissance » (Audi 1998, 238).²³

La terminologie dont j'aurai besoin pour l'économie de cette étude reposera sur ce contraste entre une notion évaluative externaliste (la connaissance) et une notion évaluative internaliste (la justification). Mais cette décision pose un problème : en partant d'une conception externaliste de la connaissance et d'une conception internaliste de la justification, ne suis-je pas en train de faire dépendre l'ensemble de mes conclusions éventuelles de présupposés extrêmement lourds ? De fait, l'opposition entre une tradition épistémologique internaliste (Chisholm, Bonjour, etc.) et une tradition épistémologique externaliste (Goldman, Dretske, etc.) est une des pommes de discorde de l'épistémologie contemporaine. Il semble donc délicat de *présupposer* l'une ou l'autre doctrine, car on s'engagerait alors dans un projet strictement « régional », au sens où il n'aurait de pertinence que pour la moitié des épistémologues. Ce problème apparent repose sur une description très contestable de la querelle entre Internalistes et Externalistes, la description selon laquelle Internalistes et Externalistes s'intéresseraient à une même notion (la connaissance, ou la justification), et seraient en désaccord sur le caractère internaliste ou externaliste de *cette* notion. La plupart des auteurs reconnaissent aujourd'hui que cette description du débat est très problématique : quand on prête attention aux analyses que différents auteurs donnent pour le terme « justification » (lesquelles analyses conduisent à une conclusion internaliste ou externaliste), on constate des disparités si grandes qu'il est quasi impossible de croire que les différents auteurs avaient tous à l'esprit le *même* concept et l'ont analysé différemment.²⁴ Ce diagnostic a été porté de manière assez radicale par William Alston (2005)²⁵ : d'après Alston, il n'y a aucun désaccord d'analyse conceptuelle entre épistémologues ; il existe seulement une liste de normes ou « *desiderata* » épistémiques différents – les uns *clairement* internalistes, les autres *clairement* externalistes – et le différend entre épistémologues porte seulement sur les *desiderata* auxquels ils accordent ou non de l'intérêt. Sans aller aussi loin qu'Alston, la plupart des auteurs acceptent aujourd'hui qu'il y a *plusieurs* normes épistémiques en jeu dans le débat et qu'un point de départ indispensable est de distinguer clairement ces différentes notions.²⁶

Ma démarche reposera sur ce diagnostic : parmi les normes épistémiques qui ont été soigneusement distinguées dans la littérature, deux d'entre elles m'intéresseront particulièrement – l'une clairement externaliste, l'autre clairement internaliste – et ce sont elles que je choisirai d'appeler « connaissance » et « justification » *par stipulation* (mais une stipulation non arbitraire, bien sûr). Pour introduire ces deux notions, le plus simple est de repartir brièvement de l'évolution de la discussion épistémologique dans les 50 dernières années.

Aujourd'hui, suite à l'influence d'un article d'Edmund Gettier dont nous allons parler dans un instant, tous les épistémologues s'accordent à penser que la connaissance doit avoir au minimum *une* condition externaliste importante.

²³ On peut noter, en outre, que les internalistes de la connaissance les plus radicaux sont également (*a fortiori*) internalistes pour la justification, et que les externalistes de la justification les plus radicaux sont également (*a fortiori*) externalistes pour la connaissance. Si bien qu'il y a, à ma connaissance, un consensus réellement universel sur le point suivant : la notion de justification est *au moins aussi internaliste* que la notion de connaissance (qui, réciproquement, est *au moins aussi externaliste* que la notion de justification).

²⁴ Il est particulièrement troublant de noter que les analyses proposées ne ressortissent même pas au même *domaine* conceptuel : les unes sont du domaine éthique, les autres du domaine de la norme fonctionnelle biologique, d'autres encore purement descriptives.

²⁵ Voir également (Bonjour and Sosa 2003, 35) et les références données par Bonjour en note 10.

²⁶ Voir notamment (Pryor 2001, 112). Même Bergmann (pour qui il existe un authentique désaccord à propos d'une *même* notion de Justification) admet qu'il y a « plusieurs notions de justification » en circulation et qu'il faut donc commencer par isoler *la* notion sur laquelle il y a, d'après lui, désaccord (Bergmann 2006, 3–4).

1. une notion minimale externaliste de « connaissance »

L'épistémologie contemporaine a été révolutionnée par un article de trois pages d'Edmund Gettier, « Is Justified True Belief Knowledge ? » (1963). Gettier part d'une analyse tripartite du concept de connaissance – la connaissance comme Croyance *plus* Vérité *plus* Justification – et montre que cette analyse est directement contredite par plusieurs contre-exemples, les fameux « cas de Gettier ».²⁷ Les cas de Gettier sont des cas de croyances vraies justifiées qu'aucun locuteur compétent du français (ou de l'anglais) ne catégoriserait comme des cas de « connaissance » (respectivement : « knowledge »).

Un bel exemple se trouve chez Oscar Wilde²⁸ : John Worthing, un dandy originaire de province, fait croire à la haute société de Londres qu'il se prénomme Ernest. Gwendolen, sa fiancée, est donc parfaitement justifiée à croire qu'il s'appelle Ernest (comme n'importe qui est justifié à croire son interlocuteur lorsque ce dernier dit son nom), bien que cette croyance repose en fait sur un mensonge. A la fin de la pièce, on apprend que M. et Mme Worthing ne sont pas réellement les parents de John Worthing : ils l'ont trouvé dans un berceau lorsqu'il était bébé, et lui ont donné le nom de John, ignorant quel était son véritable prénom. On apprend également que « John Worthing » est en fait le fils aîné disparu du Général Moncrieff ; en tant que fils aîné, il porte donc le prénom de son père... c'est-à-dire Ernest. La *croyance* de Gwendolen (que John s'appelle Ernest) était donc non seulement *justifiée*, mais (par pur hasard) il se trouve qu'elle était également *vraie*... pourtant on ne peut pas dire que Gwendolen *savait* que son fiancé s'appelle Ernest.

La conclusion immédiate à tirer des cas de Gettier est simple : les trois conditions données dans l'analyse JTB (Justified True Belief) ne sont pas suffisantes pour pouvoir parler de connaissance. Si elles étaient suffisantes, la croyance de Gwendolen serait une connaissance, mais ce n'est clairement pas le cas. Cette conclusion est inattaquable, mais elle est négative : elle nous dit ce que la connaissance *n'est pas*. Pour passer à une conclusion *positive* sur ce que la connaissance *est*, il faut s'appuyer sur un diagnostic de ce qui *ne va pas* dans les cas de Gettier. Or il y a évidemment de nombreux diagnostics différents.

Heureusement, malgré leur diversité, les diagnostics semblent converger autour d'une notion générale, la notion de « hasard épistémique ». Dans l'exemple d'Oscar Wilde, nous avons dit que la croyance de Gwendolen se trouvait être vraie « par pur hasard » : la raison pour laquelle elle croit que son fiancé s'appelle Ernest (elle le croit sur sa parole) n'a absolument rien à voir avec le fait qu'il s'appelle Ernest (puisque lui-même ignore ce fait et croit avoir menti). Les cas de Gettier sont ainsi à l'origine d'un des très rares consensus de l'épistémologie contemporaine, que Duncan Pritchard résume ainsi :

« Le consensus philosophique qui est au centre de ce livre [...] concerne l'intuition semble-t-il universelle que le hasard empêche la connaissance, ou, pour le dire

²⁷ Cette analyse tripartite de la connaissance est souvent présentée dans la littérature comme « l'analyse classique » ou « traditionnelle », sans doute en partie parce qu'elle semble remonter au moins au *Théétète* de Platon (201). Cependant, Plantinga (1993, 6–7) fait observer qu'il est en fait assez difficile de trouver des exemples de cette analyse dans la littérature qui précède l'article de Gettier. Il ne faut donc probablement pas voir l'analyse tripartite comme une tradition établie que Gettier serait venu renverser d'un trait de plume, mais plutôt comme une analyse dont l'insuffisance manifeste a fait apparaître « en creux » une condition nécessaire de la connaissance (la *quatrième* condition) que les épistémologues n'avaient jamais aperçue si clairement.

²⁸ *The Importance of Being Ernest*.

autrement, que le hasard épistémique qui nous permet parfois d'avoir des croyances vraies [...] est incompatible avec la connaissance. » (Pritchard 2005, 1)²⁹
Ce consensus universel porte donc sur une *condition nécessaire* de la connaissance, qu'on peut appeler la condition Anti-Hasard :

(AH) S sait que *p* seulement si ce n'est pas par pur hasard que sa croyance que *p* est vraie.

Au-delà de ce consensus, il peut y avoir des désaccords importants sur la bonne manière d'éviter le hasard épistémique (nous aborderons cette question principalement dans la troisième partie de cette étude) ou sur la bonne manière de réviser l'analyse JTB pour prendre en compte (AH).³⁰ Mais mon but ici n'est pas de proposer une analyse *du* concept de « connaissance » (il n'est pas évident, on l'a vu, qu'il y ait un unique concept de connaissance) ; mon but est d'utiliser le consensus anti-hasard pour isoler et caractériser *un* concept de connaissance, un *desideratum* épistémique strictement externaliste que j'appellerai « connaissance ». Je ferai cela en trois étapes.

Premièrement, il faut remarquer que l'intuition (AH) résiste très bien à l'idée qu'il y aurait plusieurs sens de « savoir ». Quand bien même le mot « savoir » aurait plusieurs sens, quand bien même il s'appliquerait à des projets ou des *desiderata* épistémiques différents, il semble clair que *aucun* sens de « savoir » ne peut s'appliquer à la croyance de Gwendolen, à moins de changer totalement de langue. Il peut donc y avoir une certaine souplesse sémantique du mot « savoir » (ou du mot « connaissance »), mais cette souplesse ne peut pas aller jusqu'à accepter des « connaissances par pur hasard ».³¹ La condition (AH) fait partie du

²⁹ Ce qui fait consensus, c'est non seulement l'intuition anti-chance, mais aussi l'idée que cette intuition anti-chance est la *leçon principale* à tirer des cas de Gettier. Voir par exemple (Dancy 1986, 134) et (Pritchard 2005, 148). Une des premières formulations vraiment explicites de l'intuition anti-hasard dans la littérature analytique est celle de (Unger 1968, 158) : « a man knows that *p* if and only if (at *t*) it is not at all accidental that the man is right about its being the case that *p*. » Et Unger interprète clairement cette condition comme une réponse adéquate aux problèmes de Gettier (Unger 1968, 164–165).

³⁰ Faut-il ajouter la condition (AH) comme *quatrième* condition, ou en faire un renforcement de la condition de Justification, ou bien au contraire abandonner totalement l'analyse tripartite comme un mauvais point de départ ?

³¹ Ne peut-on pas trouver des contre-exemples à cette affirmation ? On pourrait penser notamment au parieur d'une lotterie qui, découvrant qu'il a gagné, s'exclame « je le savais ! » alors que c'est par pur hasard qu'il est tombé sur le bon numéro. Dans *Knowledge and Lotteries* (2004, 68–70), John Hawthorne relève également deux cas où des personnes interrogées n'hésitaient pas à utiliser le terme « connaissance » pour des croyances vraies hasardeuses. Je me limiterai au cas 1 : on distribue à six enfants six livres qui contiennent une information sur la capitale de l'Autriche ; cinq contiennent une information fausse, et un seul contient la bonne réponse, mais les enfants croient tous que l'information contenue par leur livre est vraie ; on leur demande « d'apprendre » la capitale de l'Autriche sur la base de ce livre ; puis on demande à un observateur extérieur (qui connaît tout le dispositif) combien d'enfants connaissent (*know*) la capitale de l'Autriche à la fin du scénario. Au lieu de répondre « aucun », les observateurs interrogés ont fortement tendance à répondre « un », en pensant à l'enfant qui, par pur hasard, a reçu la bonne information. Hawthorne souligne en note que « de tels faits à propos de l'usage [du mot connaître] [le] font désespérer du projet épistémologique traditionnel d'extraire une analyse de la connaissance à partir d'une réflexion sur l'usage ». Ce désespoir me semble inutilement pessimiste : on peut assez bien expliquer ces « anomalies » sans réviser pour autant l'analyse selon laquelle le concept de savoir contient la notion anti-hasard. Pour le cas rapporté par Hawthorne, on peut facilement interpréter que les observateurs « traduisent » la question posée en termes de connaissance en une autre question, qui leur semble contextuellement plus pertinente, à savoir la question « combien d'enfants 'ont juste' ? ». Pour le cas du parieur heureux qui s'exclame « je le savais ! », on peut très bien penser que ce parieur s'attribue tout bonnement une croyance *non hasardeuse* (en effet, cette exclamation semble bel et bien manifester la satisfaction d'avoir employé une *compétence*) mais que cette attribution relève d'une prétention erronée. De la même façon, le tireur à l'arc novice qui tremble mais réussit par pur hasard à mettre dans le mille du premier coup s'exclamera peut-

cœur commun à toute notion de « connaissance », au même titre que la condition de « croyance » et que la condition de « vérité ».³² Ainsi, un projet ou *desideratum* épistémique quel qu'il soit ne pourra être considéré comme une notion de connaissance C_n que s'il est constitué de la manière suivante : $C_n = B + T + AH + X_n$. Les différentes notions de connaissance ne différeront que sur la ou les conditions supplémentaires X_n .

Deuxièmement, la condition (AH) est une condition *externaliste* dans le sens évoqué plus haut. On le remarquera aisément en se demandant à quelles conditions la croyance de Gwendolen aurait pu satisfaire (AH). Pour que la croyance de Gwendolen fût non-hasardeuse, il aurait fallu, par exemple, que M. et Mme Worthing soient les vrais parents de John Worthing et lui aient donné le nom de Ernest. Entre ce scénario et le scénario de la pièce de Wilde, il y a quelques différences... mais ces différences se situent soit dans un passé lointain, soit en dehors de Londres (dans la ville de province où habitent M. et Mme Worthing) – puisqu'à Londres, John se fait toujours appeler Ernest, donc on peut imaginer que les scénarios sont entièrement identiques à l'intérieur du périmètre londonien. La pauvre Gwendolen, qui n'est jamais sortie de Londres, n'a donc aucun accès à cet élément de différence entre les deux scénarios. L'élément de hasard épistémique est dû à une rupture de la chaîne d'information, mais le chaînon rompu se situe très loin à l'extérieur du système cognitif de Gwendolen. En réunissant cette remarque avec la précédente, on peut conclure que le cœur de toute notion de « connaissance » est constitué d'une condition clairement externaliste. On comprend mieux alors la remarque faite un peu plus haut, à savoir que le terme de « connaissance » tend vers le pôle externaliste de l'évaluation épistémique.³³

Troisièmement, une fois qu'on a remarqué que la condition anti-hasard constitue le cœur (externaliste) de toute notion de « connaissance », on peut légitimement se demander si elle ne pourrait pas être *suffisante* (avec la vérité et la croyance) pour une notion minimale de connaissance. Evidemment, si on se restreint strictement aux conditions de croyance + vérité + anti-hasard, on abandonne la condition de justification. La question revient donc à se demander si on peut encore parler de « connaissance », en un sens si minimal soit-il, quand on abandonne toute forme de justification. Et il semble bien que la réponse soit oui. On peut utiliser ici un exemple de Richard Foley qui est devenu célèbre, le cas des sexeurs de poussins

être « vous voyez que j'en étais capable », mais il aura tout simplement *tort* ; il aura la croyance *fausse* d'avoir un contrôle et une capacité qu'en réalité il n'a pas.

³² Il est universellement admis que si p n'est pas vraie, on ne peut pas dire que S sait que p , dans quelque sens que ce soit du verbe « savoir ». Toute notion de « savoir » doit être *factive*. Il est (presque) aussi universellement admis que si S ne croit pas que p , on ne peut pas dire que S sait que p . On a proposé comme contre-exemple l'étudiant nerveux qui donne la bonne réponse (parce qu'il l'a apprise et qu'elle lui vient à l'esprit) mais sans y croire. Il y a certainement un sens dans lequel on peut dire que l'étudiant *sait* la réponse, et un sens dans lequel on peut dire qu'il ne la *croit* pas. Mais quand on observe la faiblesse du sens de « savoir » ici employée (une disposition à prononcer les bons mots), il me semble qu'on est aussi amené à reconnaître qu'en un sens aussi faible, l'étudiant croit la réponse (ce serait une forme de croyance tacite : telle réponse est « ce qu'il dirait »). Pour dire les choses autrement : si on supprimait la condition de croyance, il est évident qu'il faudrait la remplacer par une autre condition de la même famille (une attitude positive de S à l'égard de p , comme la disposition à prononcer que p plutôt que $\sim p$). Il serait absurde de s'en tenir à la condition de *vérité*, et on ne peut *formuler* la condition anti-hasard que par rapport à une certaine *attitude* positive susceptible de vérité ou de fausseté (ce n'est pas par pur hasard que l'attitude A de S à l'égard de p est vraie).

³³ N'y a-t-il pas pourtant une forme internaliste de hasard épistémique ? (Pritchard 2005) n'a-t-il pas, justement, distingué entre une forme externaliste (le « hasard épistémique véritéque ») et une forme internaliste (le « hasard épistémique réflexif ») ? Je traiterai abondamment du « hasard réflexif » qui, d'une certaine manière, est au cœur du problème épistémique du libre arbitre. Je n'ai donc nullement l'intention de l'écarter. Seulement nous verrons qu'il est terminologiquement abusif de le considérer comme un autre « type de hasard » à côté du hasard « véritéque » ; ce que Pritchard appelle « hasard réflexif » est en fait la *possibilité épistémique* d'un hasard *véritique*. Quoi qu'il en soit de ce « hasard réflexif », la notion de hasard qui est au cœur de toute notion de connaissance est seulement la notion de hasard véritéque.

« qui, dit-on, sont capables de déterminer de manière fiable le sexe de poussins très jeunes mais ne sont pas conscients de la moindre marque distinctive qui leur permette de faire cela. Bien qu'ils soient capables de distinguer poussins mâles et poussins femelles, ils ne peuvent pas s'expliquer à eux-mêmes ni aux autres comment ils le font. [...] Tout ce qu'ils font, c'est qu'ils attrapent un poussin, et après un petit moment, par un processus qu'ils ne comprennent pas eux-mêmes, ils acquièrent la conviction soit que le poussin est mâle, soit que le poussin est femelle. Et ils ont presque toujours raison. » (Foley 1987, 168)

Les sexes de poussins permettent de faire le tri, non seulement entre poussins mâles et poussins femelles, mais aussi entre notions externalistes et notions internalistes de la connaissance. En effet, ces sexes n'ont (par hypothèse) aucune forme de justification internaliste pour leur croyance (de leur point de vue, tout ce qu'ils peuvent dire, c'est qu'ils « acquièrent la conviction »).³⁴ Donc si on est enclin à dire qu'à cause de cela ils ne savent pas que tel poussin est un mâle, c'est qu'on a recours à une notion riche de « savoir », une notion qui implique la justification. Mais inversement, on peut remarquer, à la suite de Foley, qu'il y a un *sens très naturel* dans lequel on peut dire que les sexes savent que tel poussin est un mâle.³⁵ On utilise alors une notion de « savoir » plus restreinte, qui peut se passer de la justification. Certains auteurs ont présenté ces deux intuitions comme deux thèses incompatibles à propos d'un seul et même concept de connaissance : la thèse internaliste et la thèse externaliste. Mais si on met en doute cette description des choses, si l'on pense qu'il y a peut-être simplement plusieurs notions (plus ou moins fortes) en jeu, alors il devient important de les distinguer clairement et de les nommer. Ernest Sosa (1991, 125–126) a proposé de distinguer ainsi une notion pauvre de connaissance – la « connaissance animale », restreinte à la croyance vraie *plus* la condition de fiabilité externaliste – et une notion plus riche – la « connaissance réflexive ».³⁶ Sosa parle de connaissance *animale* parce que, dans la mesure où l'on peut attribuer des croyances aux animaux, on peut aussi leur attribuer des croyances vraies dont la vérité ne repose pas sur un pur hasard (mais, par exemple, sur la fiabilité de leur système olfactif ou visuel). Bien sûr, ce que n'auront pas les animaux, ce sont des croyances de second-ordre *à propos* de la fiabilité de leur croyance... mais à ce titre, ils ne semblent guère plus dépourvus que nos pauvres sexes de poussins.

Cette notion de « connaissance animale » – la notion minimale externaliste – est celle que j'utiliserai dans cette étude lorsque j'emploierai les termes de « connaissance » et de « savoir » (sauf indication explicite contraire). Encore une fois, je n'entends pas adopter par là une certaine thèse sur le concept de « connaissance », ni sur le bon usage du mot « connaissance » ; j'entends simplement définir la notion précise qui m'intéressera, et la

³⁴ Foley (1987, 169) dit qu'ils n'ont pas de « rationalité épistémique ». Je reviendrai plus loin sur la notion de « rationalité » et son rapport avec la « justification ». Mais il est clair que les raisons qu'il apporte pour défendre l'absence de « rationalité » sont autant de raisons contre la « justification » dans le sens où nous l'entendons ici.

³⁵ Clairement, cette intuition reposera sur le fait que *ce n'est pas par hasard* que les sexes tombent juste. Dans le scénario, l'absence de hasard est assurée par l'expression « capables de déterminer de manière fiable ». C'est donc bien la satisfaction de la condition (AH) qui fait ici le travail essentiel pour fonder l'attribution de connaissance. Il est intéressant de noter que (Unger 1968), dont on a vu qu'il donnait une des premières formulations explicites de (AH), utilise justement un cas similaire de connaissance non réflexive (il s'agit chez lui d'un gitan qui utilise une boule de cristal qui est fiable, mais dont il ne connaît pas la fiabilité) pour susciter l'intuition que l'absence de hasard est *suffisante* pour un certain type de connaissance – qu'il appelle « connaissance factuelle » (Unger 1968, 163–164).

³⁶ Dans sa définition précise de la connaissance animale, Sosa utilise une théorie déjà élaborée de la bonne manière d'éviter le hasard épistémique : « belief that is apt and derives from an intellectual virtue or faculty » (Sosa 1991, 145), une croyance appropriée qui dérive d'une vertu ou d'une faculté intellectuelle. Mais les auteurs qui ont repris son expression par la suite ont libéré la notion des particularités propres à Sosa. C'est également dans ce sens élargi que j'utiliserai la notion de « connaissance animale ».

désigner par un terme courant et simple, pour ne pas alourdir une réflexion qui en fera un usage intensif. Ce choix lexicologique n'est pas totalement arbitraire. D'une part, on l'a vu, il y a des raisons de penser que le terme « connaissance » du langage naturel *peut* désigner le concept qui m'intéresse, donc ce choix lexicologique ne fait pas trop violence au langage ordinaire. Par ailleurs, quitte à utiliser le mot « connaissance » (tout court) pour *une* des notions que ce terme peut désigner, il paraît judicieux de l'utiliser pour la notion la plus élémentaire, car cela permet d'isoler le cœur notionnel par rapport aux autres « éléments » conceptuels (notamment la justification) qui entrent dans les notions plus riches comme la connaissance réflexive. Avant de passer à la notion de justification, concluons ce paragraphe par une définition formelle du concept de « savoir » (ou de « connaissance », les deux termes seront utilisés comme synonymes) que j'utiliserai :

S sait que $p =_{df}$

(i) S croit que p

(ii) p est vraie

(iii) ce n'est pas par pur hasard que la croyance de S que p est vraie.

2. une notion internaliste non déontologique de « justification »

Le terme de « justification », plus encore que celui de « connaissance », a fait l'objet d'analyses extrêmement diverses qui révèlent clairement l'existence de plusieurs notions hétérogènes.

A l'un des extrêmes, on trouve la conception dite « déontologique » ou « responsabiliste » de la justification.³⁷ La conception déontologique suppose l'existence d'un certain nombre de « devoirs » ou obligations épistémiques (« ne croire qu'en présence de raisons suffisantes », ou « rechercher un maximum de preuves avant de trancher », etc.) que j'ai la liberté (et la responsabilité) de suivre ou d'enfreindre. Lorsque j'ai rempli tous mes devoirs épistémiques, lorsque j'ai fait de mon mieux, on peut dire que ma croyance est « justifiée », ou plutôt que *je* suis « justifié », c'est-à-dire inblâmable.³⁸ Cette notion représente un extrême « internaliste » dans la mesure où elle confère au sujet un contrôle maximal sur sa justification : « Dans la conception déontologiste classique, [...] le fait que nous soyons justifiés ou non dans nos croyances dépend de nous » (Plantinga 1993, 15).

A l'autre extrême, on trouve la conception de Goldman, qui part de « l'étroite relation » traditionnellement supposée entre connaissance et justification. La « justification » est alors,

³⁷ Plantinga (1993, 11–13) a certainement raison de présenter cette notion déontologique comme la notion *classique* de justification, et même comme « la source et l'origine » de l'usage du terme « justification » lui-même, qui a clairement une origine juridique ou morale (voire théologique). Ainsi chez Descartes :

« lorsque je ne conçois pas [une chose] avec assez de clarté et de distinction, [...] si je me détermine à la nier, ou assurer, alors je ne me sers plus comme je dois de mon libre arbitre » (Descartes 1953, 307).

Et chez Locke :

« He that believes, without having any reason for believing, [...] neither seeks truth as he ought, nor pays the obedience due his maker » (Locke 1690, IV, xvii, 24).

Pour une version contemporaine du déontologisme, voir « The relation of Epistemic Justification to Moral Duty » (Chisholm 1989, 58–60), « Justified Belief as Responsible Belief » (Foley 2005), ou (BonJour 1985, 8) : « being epistemically responsible in one's believings is the core of the notion of epistemic justification ».

³⁸ Il est important de noter ici que la notion d'inblâmabilité épistémique (*epistemic blamelessness*) n'est pas nécessairement tributaire du « volontarisme doxastique », c'est-à-dire de l'idée (problématique) que nous aurions un contrôle volontaire *direct* sur nos croyances. En effet, la notion de « devoirs épistémiques » peut désigner non seulement des devoirs directs (crois ceci dans telles circonstances), mais aussi des devoirs indirects (recherche des informations pour optimiser tes chances de croire la vérité). Les seconds permettent de donner un contenu réel à la notion d'inblâmabilité épistémique même si nous n'avons pas de contrôle épistémique direct.

par définition, le cœur de la notion de connaissance, l'élément quasiment suffisant pour qu'il y ait connaissance. Aussi n'est-il pas surprenant que Goldman arrive à une analyse externaliste de la « justification », développée en termes de « fiabilité du processus ou des processus qui ont causé [la croyance] » (Goldman 1979, 95). En d'autres termes, comme Goldman veut un concept de justification assez fort pour rendre l'analyse JTB *grosso modo* suffisante, il est obligé de faire jouer à la justification elle-même (et non à une quatrième condition) le rôle anti-hasard qui est au cœur de la notion de connaissance. Or ce rôle, nous l'avons vu, est fondamentalement externaliste dans la mesure où il ne dépend aucunement du sujet, et peut même, dans certains cas, lui être totalement inconnaissable. Cette notion garde un point commun avec la précédente (et avec toute notion de justification), c'est l'idée qu'une croyance « injustifiée » ne remplit pas certains critères normatifs et prête donc le flanc à la critique. Mais il ne s'agit plus en l'occurrence de normes et de critiques *morales* (de blâmes) ; il s'agit de normes et de critiques proprement épistémologiques (et externalistes). J'appellerai donc cette notion « l'impeccabilité doxastique externe ».

La notion de justification qui nous intéressera n'est ni la notion déontologique ni l'impeccabilité externe. La notion qui nous intéressera sera une notion internaliste mais proprement épistémologique (non déontologique). Pour l'introduire, je procéderai en trois étapes, en partant, une nouvelle fois, du consensus post-Gettier.

Premièrement, la « justification », dans le sens important en épistémologie, doit être telle que les cas de Gettier traditionnels sont dits « justifiés » et non pas « injustifiés ». Cette contrainte permet d'exclure la notion goldmanienne : en effet, si on utilise sa notion de justification, alors la plupart des cas de Gettier devront être considérés comme non-justifiés.³⁹ Goldman reconnaît qu'il y a une autre notion de justification, indépendante de la notion de connaissance, et pour laquelle les cas de Gettier seront « justifiés ».⁴⁰ Mais pour lui, cette autre notion semble moins importante. Le « consensus post-Gettier » fait au contraire apparaître le caractère fondamental de cette autre notion : *il y a* une notion épistémique fondamentale qui est satisfaite par les cas de Gettier (bien qu'ils ne soient pas des connaissances) ; et il semble légitime de réserver le terme de « justification » pour cette notion pour laquelle nous n'avons guère d'autre terme.⁴¹ Le consensus post-Gettier, en même temps qu'il fait apparaître le cœur anti-hasard de la notion de connaissance, fait apparaître « en creux » une notion de justification *distincte* de la condition anti-hasard.

Deuxièmement, la notion de « justification » qui est pertinente dans les cas de Gettier est une notion strictement internaliste. Autrement dit, lorsqu'on juge que la croyance de Gwendolen est « justifiée » (bien qu'elle ne soit pas une connaissance), les normes qu'on prend en compte pour porter ce jugement positif sont des normes strictement internalistes. Ce n'est pas une conséquence immédiate de la première remarque : après tout, il se pourrait que la croyance de Gwendolen satisfasse *certaines* normes externalistes (certes, pas la norme anti-hasard, mais d'autres) et que notre jugement positif soit fondé sur une prise en compte de *ces* normes. Bergmann soutient en fait cette thèse ; ou plutôt, il soutient que les cas de Gettier peuvent être réputés « justifiés » selon deux lectures différentes, l'une externaliste, l'autre

³⁹ Ce qui n'est pas surprenant étant donné, comme nous l'avons vu, qu'il part d'une notion de « justification » pour laquelle l'analyse JTB doit être *grosso modo* correcte... c'est-à-dire une notion de justification pour laquelle il ne *doit pas* y avoir de contre-exemples (sauf exceptions rares et périphériques).

⁴⁰ C'est même une notion qu'il a lui-même utilisée avant l'article de 1979 : « In previous papers on knowledge, I have denied that justification is necessary for knowing, but there I had in mind 'Cartesian' accounts of justification » (Goldman 1979, 89-90).

⁴¹ Bien sûr, l'accent de la discussion Gettier est généralement mis sur un autre fait, à savoir : *il y a* une notion épistémique fondamentale qui n'est *pas* satisfaite par les cas de Gettier, et qui empêche qu'ils comptent comme connaissance. Mais pour désigner cette autre notion épistémique (la notion anti-hasard), nous n'avons pas besoin du terme de « justification » (*pace* Goldman) : on peut utiliser le terme de « connaissance » lui-même.

internaliste.⁴² Et quant à lui, il garde le terme « justification » pour la notion externaliste « gettierisable ». Mon choix terminologique est inverse. Pour éviter toute controverse, on peut considérer ce choix comme une simple stipulation, mais cette stipulation n'est pas gratuite. Elle repose sur le fait qu'il y a un *desideratum* épistémique qui est *strictement internaliste*, qu'il est important d'isoler clairement ce *desideratum* de tout *desideratum* externaliste (gettierisable ou non), et que la terminologie la plus commune pour désigner ce *desideratum* semble bien être celle de la « justification ». Un argument récent qui montre cela clairement est le *Nouvel Argument du Malin Génie* (NAMG). Le NAMG réutilise le scénario sceptique traditionnel d'un Malin Génie (qui manipulerait une âme pour lui donner exactement les mêmes expériences que nous avons) mais pour en faire une utilisation totalement différente. L'argument soutient que la « justification » doit être une notion internaliste puisque, intuitivement, une âme manipulée par un Malin Génie serait tout aussi « justifiée » que nous dans ses croyances.⁴³ Bien sûr, les deux âmes ne seraient pas épistémiquement à égalité : l'âme manipulée ne pourrait avoir aucune *connaissance*. Mais il y a *quelque chose* d'épistémiquement important qu'elles ont en commun : elles sont toutes les deux (par hypothèse) impeccables du point de vue internaliste. En ce sens, il semble légitime de dire que l'âme manipulée serait totalement *justifiée* dans ses croyances. Certains épistémologues « internalistes » en tirent la conclusion que « la » notion de justification est internaliste. Cette conclusion est controversée et présuppose l'existence d'une unique notion de justification sur laquelle il y aurait désaccord. Mais ce qui semble indiscutable, c'est que le NAMG fait apparaître *une* caractéristique épistémique fondamentale qui est strictement internaliste (dans le sens où elle résiste au scénario de l'âme manipulée), à savoir l'« impeccabilité doxastique interne ». C'est à cette notion que je choisis de réserver l'appellation de « justification ».

Troisièmement, la notion pertinente d'impeccabilité doxastique interne n'est *pas* la notion déontologique de justification mais une notion proprement épistémologique. On a vu que la notion déontologique de justification était une notion strictement internaliste⁴⁴, mais ce n'est pas la *seule* notion strictement internaliste qui puisse être préservée en cas d'échec externaliste (en cas de manipulation par le Malin Génie, notamment) : dans le NAMG, les croyances de l'âme manipulée ne sont pas seulement *moralement* irréprochables (parce que l'âme a fait de son mieux), elles sont aussi *épistémiquement* irréprochables du point de vue de leurs caractéristiques internes. Et les deux ne vont pas toujours de pair : par exemple, il est certainement possible d'imaginer une personne qui a été si mal éduquée (d'un point de vue épistémique) qu'on ne peut pas lui reprocher *moralement* d'accepter certaines propositions contradictoires. Il n'en demeure pas moins qu'une telle personne n'aurait pas un système de croyances impeccable du point de vue interne. C'est-à-dire que son système de croyances violerait des normes *strictement internalistes* et néanmoins *non déontologiques* (proprement épistémiques). Par conséquent, au sein même des normes strictement internalistes, on peut encore distinguer des normes d'ordre moral et d'autres proprement épistémiques (c'est-à-dire non morales).⁴⁵ Celui qui ne viole aucune de ces normes internalistes proprement

⁴² « One can read Gettier's paper (1963) thinking of justification as subjective deontological justification or as the sort of justification I'm describing here. Either way his conclusion is persuasively established. » (Bergmann 2006, 8)

⁴³ « La vérité de l'hypothèse du Malin Génie implique que nos expériences et nos raisonnements sont exactement tels qu'ils seraient si nos processus cognitifs étaient fiables, et par conséquent, que nous serions tout aussi justifiés à croire ce que nous croyons si l'hypothèse était vraie que si elle était fausse. Contrairement au fiabilisme, nous affirmons que dans les conditions de l'hypothèse du Malin Génie, nos croyances seraient justifiées dans un sens épistémique. » (Lehrer and Cohen 1983, 192–193).

⁴⁴ Au moins à première vue ; Pryor (2001, 115–116) montre que cette thèse n'est pas si évidente. Je n'entrerai pas ici dans cette discussion, dans la mesure où je n'aurai aucun usage de la notion déontologique de justification.

⁴⁵ Cette même distinction est défendue par Pryor (2001, 117) qui au lieu du Malin Génie cartésien, utilise un autre scénario classique, celui des cerveaux manipulés dans des cuves de formol : « Consider three brains in a

épistémiques satisfait un *desideratum* épistémique important, un *desiratum* qui est à la fois strictement internaliste et non déontologique : l'impeccabilité épistémique interne.⁴⁶ Et c'est donc *ce desideratum* important qui m'intéressera particulièrement dans cette étude et que je choisirai d'appeler « justification ».

J'insiste sur le fait que cette terminologie n'engage aucune *thèse* internaliste : il est incontestable qu'il y a des *desiderata* strictement internalistes, et il est aussi incontestable que certains de ces *desiderata* strictement internalistes (par exemple la non-contradiction, ou la non-circularité⁴⁷) sont d'un ordre non moral. Les défenseurs de « thèses externalistes » ne remettent pas en cause l'existence de tels *desiderata*. Ces deux points étant posés, il est donc loisible de s'intéresser au statut épistémique particulier d'une croyance qui « n'a violé aucun de ces *desiderata* strictement internalistes et non déontologiques ». Et il est loisible également de donner un nom à ce statut épistémique. Un nom clair (mais un peu complexe) est « l'impeccabilité épistémique interne ». Comme je ferai un usage abondant de cette notion, je préfère lui donner un nom plus simple : la « justification ». On peut donc donner la définition suivante :

S croit que *p* de manière justifiée =_{df}

(i) S croit que *p*

(ii) S ne viole aucune des normes strictement internalistes proprement épistémiques en croyant que *p*.

Mais alors, dira-t-on, si *cela* n'est pas une thèse internaliste sur la justification, qu'est-ce qu'une thèse internaliste ? Et comment y aurait-il place pour défendre une thèse externaliste avec une telle définition ? Pour une réponse développée à ces questions, je renvoie à l'annexe 1. Dans cette introduction, je soulignerai seulement ceci : un externaliste comme Bergmann ou Plantinga *ne nie pas* l'existence, ni même l'importance, des *desiderata* internalistes non déontologiques que je viens de mentionner, et il pourrait donc accepter de parler le langage que je définis ici.⁴⁸ Seulement, il les considèrera comme des *desiderata* simplement « négatifs », des *defeaters* potentiels de justification (des conditions d'annulation), mais incapables de fournir une notion positive de justification. L'internaliste en revanche, soutient qu'il existe des *desiderata* internalistes positifs, c'est-à-dire des normes qui n'exigent pas seulement l'*absence* d'éléments internes problématiques (l'absence de contradiction ou de circularité), mais la *présence* d'un élément interne (un argument, une raison, des données) contribuant positivement à la justification – un élément qu'on peut appeler un *justifieur* internaliste. Donc le débat substantiel entre Internalistes et Externalistes porte sur la *nature*

vat. The first brain is epistemically quite reckless. He believes whatever he feels like believing, so long as he doesn't have incontrovertible evidence that it's false. The second brain is [...] doing the best he can, epistemically, and the best that can be reasonably expected of him. He's very scrupulous and careful in following the epistemic standards he picked up as a child. Unfortunately, these epistemic standards are defective in a way he's not capable of recognizing. [...] The third brain carefully follows the same epistemic standards we follow. It seems natural to say that, although there's a sense in which the second and third brains are both epistemically blameless, only the third is conducting his epistemic affairs *properly*, and so only he is justified in his beliefs. This indicates that there's a kind of positive epistemic status which goes beyond mere epistemic blamelessness, but which one can possess even if one's beliefs are reliably false, as the brains' beliefs are. »

⁴⁶ Pour compléter la terminologie, on pourrait choisir d'appeler « impeccabilité *doxastique* interne » l'impeccabilité à l'égard de *toutes* les normes internalistes, qu'elles soient déontologiques ou épistémologiques. On aurait alors l'équation suivante : impeccabilité doxastique interne = inblâmabilité + impeccabilité épistémique interne.

⁴⁷ Nous reviendrons abondamment sur la notion de circularité épistémique.

⁴⁸ De fait, Plantinga reconnaît l'existence de cette notion de justification qu'il appelle « *broad justification* » et caractérise comme « the idea that things are going right [...] with respect to everything downstream of experience » (Plantinga 1993, 27).

(positive ou seulement négative) des normes internalistes non-déontologiques et pas sur leur existence. Ainsi, l'une et l'autre thèse peuvent être formulées en employant la notion de « justification » telle que je viens de la définir :

Internalisme :

Pour être justifié à croire que *p*, *S* doit avoir un justifieur (interne) en faveur de *p*.⁴⁹

Externalisme :

S peut être justifié à croire que *p* sans avoir de justifieur (interne) en faveur de *p*.⁵⁰

A ce stade, un point de clarification s'impose sur la notion d'Internalisme ou de « caractéristique interne ». Le sens de « interne » dans ce contexte ne peut pas être trivialement local. Le bon fonctionnement de ma rétine est une caractéristique interne à mon corps, et cette caractéristique est certainement un des *desiderata* que mes croyances visuelles devraient satisfaire... mais ce ne sera pas un *desideratum* internaliste dans le sens qui nous intéresse. Et il est possible d'être internalistiquement impeccable avec une rétine défectueuse. « Interne », dans notre contexte, ne veut donc pas dire « interne au corps du sujet ». Mais alors interne à *quoi* ? Il y a principalement deux réponses, qui ont été clairement distinguées par (Feldman and Conee 2001, 2) : interne à l'*esprit* du sujet (sa « vie mentale »), ou interne à la *perspective du sujet* (ce que le sujet peut dire de son propre point de vue, ce à quoi il a un accès spécial, ou direct, ou privilégié).⁵¹ Feldman et Conee appellent « Mentalisme » le premier type d'Internalisme,⁵² et « Accessibilisme » le second.⁵³ L'Accessibilisme n'implique

⁴⁹ Ce réquisit de « justifieur internaliste » a été très souvent exprimé de la manière suivante : « Pour être justifié à croire que *p*, *S* doit avoir une *raison* ou des *données* (*evidence*) en faveur de *p* ». On comprend alors le lien si étroit entre « Internalisme » et « Evidentialisme » : il s'agit en fait d'une seule et même thèse, désignée de deux points de vue différents. Lorsqu'on part de l'idée que toute croyance doit reposer sur un fondement positif, et qu'on souligne ensuite que ce fondement positif est *interne* et non externe, on parlera d'*Internalisme*. Lorsqu'on se concentre strictement sur les conditions d'impeccabilité du domaine interne, et qu'on souligne que ces conditions d'impeccabilité incluent un réquisit de fondement positif, c'est-à-dire un réquisit d'*evidence*, on parlera d'*Evidentialisme*. Sur le lien entre Internalisme et Evidentialisme, voir (Plantinga 1993, 25–26) et (Conee and Feldman 2004, 48–49). J'y reviendrai au chapitre 2.

⁵⁰ Cette définition de l'externalisme peut surprendre car elle ne pose aucune condition externe particulière. Comment une position serait-elle externaliste si elle ne pose aucune condition externe ? Cette étrangeté est liée au fait que nous ne caractérisons ici qu'une partie de la position de l'externaliste, celle par laquelle il s'oppose à l'internaliste sur la question de l'impeccabilité interne. Sur cette question de l'impeccabilité interne, la position externaliste est strictement négative : elle ne requiert *pas* de justifieur positif. Mais *par ailleurs*, l'externaliste insistera sur l'importance de réquisits (positifs) externalistes. Il faut noter que l'internaliste peut tout à fait accepter également ces réquisits externalistes : dans son débat avec l'internaliste, c'est par défaut que l'externaliste peut être appelé « externaliste » ; ce n'est pas parce qu'il accepte plus de contraintes externes, mais parce qu'ayant refusé les contraintes (positives) internes, il ne peut se rabattre que sur les externes.

⁵¹ Cette même distinction se retrouve de diverses manières chez différents auteurs ; Bergmann (2006, 111) la présente d'abord comme une distinction entre deux types d'*evidence* « One proposal in the literature is that evidence is limited to things of which the subject is aware or potentially aware (i.e. to epistemically accessible things). Another proposal is that evidence is limited to the subject's mental states ». Pryor (2001, 104–105) oppose à l'*Access Internalism* un *Simple Internalism* qui correspond au Mentalisme de Feldmann et Conee, mais son explication insiste moins sur le concept du *mental* que sur le concept de *survenance* qui est nécessaire pour exprimer la thèse Mentaliste (comme on va le voir) : « there are two dominant ways in which epistemologists understand the notion of internalism. One is in terms of a supervenience thesis. The other is in terms of whether the subject has special access to her epistemic condition. » Plantinga, quant à lui, définit tout d'abord l'Internalisme à partir de l'Accès épistémique (Plantinga 1993, 5) : « the internalist holds that a person has some kind of special epistemic access », mais il reconnaît ensuite, à côté de cette notion de « *epistemic internalism* », une autre notion : « 'personal internalism' (calling attention to the way in which my beliefs, desires, experience, and aims are crucial to me as a person. » (Plantinga 1993, 23).

⁵² « [Mentalism] is the view that a person's beliefs are justified only by things that are internal to the person's mental life » (Feldman and Conee 2001, 2).

pas (en tout cas pas trivialement) le Mentalisme : il n'est pas exclu que certains aspects de la vie mentale d'un sujet soient difficiles voire impossibles à ressaisir du point de vue du sujet.⁵⁴ Mais l'inverse semble plus compliqué : un élément ne peut faire partie de la « perspective » ou du « point de vue » du sujet que s'il fait partie de sa vie mentale. La notion la plus générale de « condition interne » est donc la notion « mentaliste », et c'est celle que je retiendrai pour mon propos⁵⁵ :

x est une condition interne pour le sujet $S =_{df}$

x est une condition dont la satisfaction survient sur la vie mentale du sujet S .

A partir de cette notion de « caractéristique interne », on peut reformuler plus précisément la définition de « justification » donnée plus haut :

S croit que p de manière justifiée $=_{df}$

(i) S croit que p

(ii) les faits mentaux de S (ou les faits qui surviennent sur ceux-ci) ne violent aucune des normes (non déontologiques) qui s'appliquent à eux, relativement à la croyance que p .

Nous avons donc à présent les deux notions épistémologiques qui seront utilisées dans cette étude : une notion strictement externaliste de « connaissance » comme « croyance non hasardeusement vraie », et une notion strictement internaliste de « justification » comme « impeccabilité mentale non déontologique ». Malgré la complexité apparente des réflexions qui précèdent, l'intention de cette terminologie est très simple : séparer strictement ce qui relève du bon fonctionnement externe de ce qui relève du bon fonctionnement interne ; ou pour reprendre une métaphore de Plantinga, séparer les réquisits qui sont « en amont de l'expérience » de ceux qui sont « en aval de l'expérience ».⁵⁶ Une distinction parfaitement claire de ces deux ordres deviendra particulièrement précieuse lorsque nous aborderons les questions d'*interaction* entre les deux.

III. Le Problème d'Intégration du Libre arbitre

Les définitions qui précèdent nous apportent une meilleure compréhension des deux questions (J) et (K) concernant le Libre arbitre : la question (J) concerne la place de la

⁵³ « What we shall call 'accessibilism' holds that the epistemic justification of a person's belief is determined by things to which the person has some special sort of access » (Feldman and Conee 2001, 2). Bergmann, quant à lui, parle de la « Subject's Perspective » comme d'une *motivation* particulière pour l'internalisme, mais la distinction reste fondamentalement la même (Bergmann 2006, 12).

⁵⁴ On peut penser, par exemple, à certains aspects fins et non centraux d'une expérience perceptive, comme une nuance fine de couleur aux limites extérieures de notre champ visuel. Ou encore à des aspects de la vie mentale trop abstraits et complexes pour être conceptualisés par le sujet de cette vie mentale, comme par exemple l'unité intermodale de l'expérience des formes entre expérience visuelle et expérience tactile, chez un enfant de 2 ans. Savoir si l'expérience des formes est une unique expérience intermodale ou non, est certainement une question déterminée de manière purement internaliste... mais il est clair que cette question est totalement inaccessible à la perspective d'un enfant de 2 ans (cf. Milhet and Streri 1988).

⁵⁵ Pourquoi retenir la notion la plus générale ? Parce que la seule chose qui soit vraiment admise par tous, c'est l'existence de *desiderata* (négatifs) internalistes... dans le sens le plus général (donc mentalistes). Soutenir l'existence de *desiderata* internalistes *accessibles*, c'est déjà soutenir une thèse controversée, une thèse qui en outre semble amener naturellement l'idée de *desiderata* positifs et non plus seulement négatifs (l'Internalisme).

⁵⁶ « Downstream of experience » (Plantinga 1993, 27). Je n'ai rencontré l'expression inverse (« upstream of experience » pour désigner les considérations externalistes) que dans (Siegel 2011).

croyance libertarienne à l'intérieur même du système de croyances, ou plus généralement de l'esprit : un esprit qui contient cette croyance est-il en lui-même épistémiquement défectueux ? La question (K) en revanche concerne le rapport entre la croyance libertarienne et le *fait* qui est cru par elle : cette croyance peut-elle dépendre du fait, d'une manière ou d'une autre, ou ne pourrait-elle être vraie, au mieux, que *par pur hasard* ?

Laquelle de ces deux questions constitue le cœur du « problème épistémologique » du Libre arbitre ? Les deux sont fondamentales, mais de manière différente. La question (J) est *l'enjeu* du problème épistémologique : ce qui est en question, fondamentalement, c'est qu'on puisse être justifié à tenir une position libertarienne dans le débat philosophique actuel. Mais la question (K) est la *source* du problème épistémologique : la raison pour laquelle une position libertarienne peut sembler injustifiée, c'est précisément parce que nous ne voyons pas comment elle pourrait constituer une *connaissance*.

Avant d'arriver à une formulation définitive du problème abordé dans cette étude, il nous faut encore expliquer deux éléments importants : tout d'abord, le lien entre la « justification » d'une thèse et sa pertinence dans le débat philosophique; et deuxièmement, la raison pour laquelle le problème (K) se répercute en un problème de Justification.

1. Justification et Rationalité dans le débat philosophique

Dans un débat philosophique, la caractéristique fondamentale que doit respecter une position ou une thèse est la *rationalité*. Bien sûr, si la thèse est *en outre* vraie, ça ne gâche rien, et à vrai dire, c'est cela que tout le monde vise. Mais comme la vérité de la thèse est précisément l'enjeu du débat, on ne peut pas exclure une thèse du débat au motif qu'elle n'est pas vraie. En revanche, si une thèse n'est pas *rationnelle*, elle pourra être exclue du débat avant même d'envisager la question de sa vérité. L'enjeu profond de ma recherche est de savoir si le Libertarianisme est ou non une thèse *rationnelle*, une thèse qui mérite de faire partie du débat philosophique sur la liberté humaine.

Mais qu'est-ce que la *rationalité* ou l'*irrationalité* d'une thèse philosophique ? Une forme particulièrement grave d'irrationalité est le mépris volontaire de toute norme de rationalité, c'est-à-dire la sophistique. Quand on dit que les thèses irrationnelles sont exclues du débat philosophique, on peut penser aux thèses purement sophistiques. Mais ce n'est pas notre question ici: il est évident que Thomas Reid, Roderick Chisholm, Timothy O'Connor, etc. n'ont aucun mépris pour les normes de rationalité. La plupart de ces auteurs, en tant qu'épistémologues, montrent même un intérêt inhabituellement développé pour ces normes, et un effort important pour les respecter. Il serait donc absurde de considérer le libertarianisme comme une position qui manifeste un simple mépris des normes rationnelles. Mais on peut violer les normes rationnelles en ayant fait de son mieux pour les respecter. Une position non sophistique (ou « sérieuse ») peut donc, en dépit de son auteur, être irrationnelle. Une telle irrationalité sera assurément moins grave que la première, mais elle sera néanmoins disqualifiante pour le débat philosophique, si du moins elle est établie.

La question de rationalité qui nous intéresse est donc celle de l'éventuelle violation de normes rationnelles par un auteur, quand bien même il aurait « fait de son mieux ». Il est clair que cette notion de rationalité n'est autre que l'impeccabilité interne non déontologique, c'est-à-dire dans mon vocabulaire la « justification ».⁵⁷ Savoir si la position libertarienne est

⁵⁷ Plantinga a donné une excellente classification des différents sens possibles du mot « rationalité » dans (Plantinga 2000, 108–134). Le sens qui nous intéresse ici est un cas particulier de la « rationalité comme bon fonctionnement » (*proper function*), que Plantinga appelle la *rationalité interne*. La définition qu'il en donne montre clairement que cette « rationalité interne » est exactement la même notion que sa « justification au sens large » déjà évoquée en note 48: « We can initially characterize internal rationality as a matter of proper function

rationnelle revient donc à se demander si elle est épistémiquement *justifiée*... mais il faut encore préciser *pour qui* elle doit être justifiée. En effet, la justification comme bon fonctionnement « en aval de l'expérience », ou comme bon traitement de l'information, est une condition assez facile à satisfaire pour des êtres qui ont une *très mauvaise* information, ou de très mauvaises expériences. Par exemple, imaginons un enfant de dix ans qui aurait toujours vécu dans une société archaïque où tout le monde croirait au libertarianisme, et où aucune information contraire ne lui serait parvenue ; un tel enfant serait très probablement justifié à croire au libertarianisme. Mais cela ne permet pas vraiment d'affirmer que le libertarianisme est une position rationnelle dans le débat philosophique actuel, tout simplement parce que les participants à ce débat possèdent bien d'autres informations que cet enfant de dix ans, et ces informations peuvent donc avoir un effet sur leur justification.

Dire que la position libertarienne est rationnelle dans le débat philosophique actuel, c'est donc dire qu'elle est (ou peut être) justifiée pour quelqu'un qui maîtrise l'ensemble des données du problème, l'ensemble des informations aujourd'hui connues.⁵⁸ La rationalité d'une thèse, ce n'est pas la justification *pour n'importe qui*, c'est la justification pour le philosophe informé.

Cette notion de « rationalité » comme « justification pour le philosophe » peut faire naître une confusion : on pourrait croire que ce qui est visé ici est une forme particulière de justification, une justification répondant à des standards plus élevés. L'activité scientifique ou philosophique est parfois conçue comme une manière d'élever les standards épistémiques de nos croyances, c'est-à-dire de les soumettre à des normes d'acceptation plus strictes. La décision cartésienne de révoquer en doute toutes les croyances qui ne sont pas absolument certaines peut être interprétée de cette manière : si l'on s'en tient aux normes épistémiques courantes, la croyance que j'ai deux mains réussit le test de justification, mais si on choisit des normes épistémiques plus rigoureuses, la même croyance échoue.⁵⁹ Ce genre de projet « d'élévation des standards » a un intérêt théorique incontestable lorsqu'il constitue une étape méthodologique.⁶⁰ Mais ce n'est absolument pas la question que je poserai à propos du Libertarianisme. Il y a indéniablement des manières d'élever nos standards épistémiques telles

of all belief-producing processes 'downstream from experience'. » (Plantinga 2000, 110). Or ce sens large de « justification » est précisément celui que j'emploie.

⁵⁸ Ce personnage, « qui maîtrise l'ensemble des informations aujourd'hui connues », implique bien sûr une certaine idéalisation. Après tout, il se pourrait qu'il y ait deux informations, A et B, toutes deux pertinentes sur la question du Libre arbitre, et connues aujourd'hui par de nombreux scientifiques *séparément*, mais connues par personne *conjointement*. Il n'y aurait alors personne dans le monde pour savoir « ce qu'on sait aujourd'hui sur la question du Libre arbitre ». Cette idéalisation n'aura pas d'importance pour ma recherche, qui repose uniquement sur des données *effectivement connues* par tous les philosophes actuels ; la question sera donc simplement de savoir si ces données rendent impossible une position libertarienne justifiée.

⁵⁹ Cette lecture de Descartes n'est pas la plus évidente d'un point de vue historique, ni la plus intéressante pour discuter du problème sceptique. Comme Barry Stroud l'a montré, si le problème sceptique est si important philosophiquement, c'est parce qu'il semble émerger des normes épistémiques *les plus communes*, celles que nous utilisons dans nos « pratiques métacognitives » quotidiennes. Ce que le scepticisme met en cause dans ce cas, ce n'est pas la viabilité d'un projet épistémique utopiste, c'est bien la connaissance quotidienne elle-même (Stroud 1984, chap. 1).

⁶⁰ On peut en effet distinguer une utilisation *méthodologique* et une utilisation *normative* de l'élévation de standards. La première consiste à élever les standards simplement « à titre d'hypothèse », pour se rendre compte de « ce qui reste » lorsqu'on les élève. Par exemple, un historien de l'antiquité pourrait se poser méthodologiquement la question suivante : à quoi ressemblerait un manuel d'histoire de Rome si nous n'avions aucun texte de l'époque, seulement les sources archéologiques qui ont l'avantage d'offrir une certitude absolue. Mais évidemment, si un historien écrivait ce manuel et affirmait que tous les autres manuels sont irrationnels car ils reposent sur des standards de connaissance douteux, la communauté scientifique ne le prendrait pas très au sérieux. Autrement dit, cette élévation de standards ne peut pas avoir valeur *normative* pour déterminer ce qu'il est ultimement rationnel ou irrationnel de croire en histoire romaine.

que la croyance libertarienne ne passera pas le test. Par exemple, il suffit d'adopter comme norme de croyance la certitude mathématique ; un tel standard ne laisse pas passer beaucoup de croyances, et en particulier n'en laisse passer aucune sur les objets contingents tels que le monde extérieur ou ma propre existence, et donc *a fortiori* ma liberté. Mais ce résultat n'est pas très problématique pour le Libertarianisme. Ce qui serait vraiment problématique, ce serait que les *standards épistémiques habituels* rendent impossible une croyance justifiée au libertarianisme *lorsque ces standards sont appliqués aux données dont nous disposons aujourd'hui*. Et c'est bien cela que défendent un certain nombre de philosophes anti-libertariens. La question n'est donc pas de savoir si le libertarianisme peut être « éminemment » justifié, mais s'il peut être (normalement) justifié en présence de données expertes.

Pour résumer, l'enjeu de cette étude est la « question de rationalité » suivante :

(R) Etant donné les informations dont nous disposons aujourd'hui, est-il possible pour un philosophe informé de tenir une position libertarienne *justifiée*, c'est-à-dire une position qui respecte les normes communes de rationalité internaliste non déontologique ?

2. Rationalité et Connaissance : la contrainte de méta-cohérence

Si l'enjeu de cette étude est, fondamentalement, une question de justification, on peut se demander si la question de connaissance garde une pertinence. En effet, j'ai donné des définitions de connaissance et de justification qui en font deux notions profondément hétérogènes. Il semble donc qu'on puisse répondre positivement à la question de justification sans se soucier de la question de connaissance. Et pourtant, le « problème d'intégration » est bel et bien un problème de *justification* (ou de rationalité) qui repose sur un problème de *connaissance*. D'où vient donc ce problème ?

Repartons de la croyance de Gwendolen : sa croyance est vraie par pur hasard et ne constitue donc pas une connaissance, mais nous avons vu que ce hasard (externe) n'avait aucun effet sur sa justification, conçue en termes internalistes. Dans ce cas, le problème de connaissance ne se répercute pas en problème de justification. Supposons à présent le scénario suivant : quelqu'un vient *informer* Gwendolen de ce hasard ; Gwendolen accepte cette nouvelle information, mais comme elle désire profondément que son fiancé s'appelle Ernest, elle n'abandonne pas sa croyance pour autant. Gwendolen se trouve alors dans une situation où elle croit à la fois que son fiancé s'appelle Ernest, et que cette croyance ne peut pas constituer une connaissance. Dira-t-on encore que le problème de connaissance n'a aucun effet sur la justification de Gwendolen ? Intuitivement non : la croyance entêtée de Gwendolen a un vrai problème de justification, c'est-à-dire qu'elle a un défaut internaliste. Comment un problème externe peut-il provoquer un défaut internaliste dans la croyance de Gwendolen ? La réponse est assez évidente à partir de l'exemple : ce qui pose un problème de justification pour Gwendolen, ce n'est pas le défaut externe *en lui-même*, c'est la *croyance* qu'il y a un défaut externe. Cette croyance, étant un élément strictement internaliste, est bien le genre de chose qui peut provoquer un défaut de justification.

C'est le même genre de problème qui se produit pour le Libertarianisme. Un simple défaut de connaissance *s'il est totalement ignoré* ne provoque aucun problème pour la justification de la croyance libertarienne. Mais un philosophe informé est censé être au courant des différentes discussions et informations pertinentes, et donc s'il y a une discussion sur le problème de connaissance, cela devra être pris en compte dans ses réflexions sur la question. C'est-à-dire que l'éventuel problème externe sera « représenté » à l'intérieur même

de son système de croyances, et pourra susciter un problème de justification interne. Supposons par exemple que le philosophe en question soit à la fois libertarien et convaincu qu'il n'y a pas de solution au problème de connaissance. Dans ce cas, il aura, comme Gwendolen dans le scénario proposé plus haut, une croyance dont il considère lui-même qu'elle ne peut pas constituer une connaissance. Et cette situation semble intuitivement violer un principe de rationalité internaliste (non déontologique). Je discuterai abondamment (au chapitre 3, sec. 2) de ce principe, de son importance et de sa bonne formulation. Mais il est utile dès maintenant de lui donner un nom et une formulation préliminaire. Ce principe est généralement appelé « principe de Méta-cohérence ».⁶¹ Et on peut le formuler ainsi :

(Méta-cohérence) Si je crois (internalistiquement) que ma croyance que *p* n'est pas une connaissance (a un défaut externe), alors ma croyance que *p* n'est pas non plus justifiée (internalistiquement).

ou encore :

(Méta-cohérence') Pour être justifié à croire que *p*, je dois être justifié à croire que je sais que *p*.⁶²

Le principe de Méta-cohérence permet donc de comprendre pourquoi un problème externe de connaissance du libre arbitre peut avoir des conséquences directes dans le débat métaphysique – c'est-à-dire dans la question de savoir quelle métaphysique du libre arbitre est la mieux justifiée. C'est lui qui établit le pont entre connaissance et rationalité, et qui se trouve au cœur de tout problème d'intégration : sans principe de méta-cohérence, un philosophe des mathématiques pourrait répondre à Benacerraf : « Effectivement, nous ne pouvons pas *savoir* que $2 + 2 = 4$ puisqu'il s'agit d'un fait concernant des objets abstraits, mais cela n'empêche pas que nous soyons pleinement justifiés à le croire ». On pourrait donc, semble-t-il, éviter le scepticisme sans fournir aucune solution d'intégration épistémique. Le Défi d'Intégration est donc fondamentalement un problème qui concerne la « justification (interne) à croire qu'on a une connaissance (externe) ». On peut donc l'appeler un problème de « méta-justification », ou « méta-épistémique », ou encore un problème « externe-interne », ou « de deuxième ordre ». On pourrait être tenté de dire également qu'il s'agit d'une question de « méta-philosophie » puisqu'elle étudie philosophiquement la manière dont il faut rationnellement procéder dans une question philosophique. De ce point de vue, mon traitement du problème se rapprocherait de celui de Richard Double dans *Metaphilosophy and Free Will* (1996). Mais je ne suis pas favorable à cette approche « méta-philosophique » qui, à mon sens, revient à trancher des questions qui sont fondamentalement *épistémologiques* sans utiliser les ressources conceptuelles de l'épistémologie. Je développe ce point dans l'annexe 2.

Sur la base des considérations qui précèdent, nous sommes donc en mesure de donner une formulation plus précise du défi d'intégration tel qu'il se pose en particulier pour le philosophe libertarien :

⁶¹ C'est un principe de « cohérence », parce qu'il concerne le rapport plus ou moins harmonieux de plusieurs croyances entre elles (la croyance que « *p* » et la croyance que « ma croyance que *p* constitue une connaissance »). Et c'est un principe de niveau « méta », parce qu'il concerne une croyance de second ordre, une croyance à propos de mes croyances.

⁶² Ces deux formulations ne sont pas équivalentes. Nous verrons leurs différences au chapitre 3, et nous aborderons aussi la question de savoir laquelle (ou lesquelles) il faut accepter. L'important ici est seulement l'intuition générale qu'un certain principe dans cette « famille » doit être vrai, et pose un problème de justification interne sur la base d'un problème de connaissance externe.

(Problème) Il n'est pas évident que la position libertarienne soit une position philosophique rationnelle (i.e. justifiée étant donné les informations dont nous disposons aujourd'hui) parce qu'il n'est pas aisé de voir comment elle pourrait satisfaire au principe de Méta-cohérence (i.e. elle ne semble pas pouvoir pas constituer une connaissance).

Et la thèse que je défendrai sera la suivante :

(T) La position libertarienne est une position philosophique rationnelle (i.e. justifiée étant donné les informations dont nous disposons aujourd'hui) parce qu'elle peut satisfaire au principe de Méta-cohérence (i.e. elle peut constituer une connaissance).

IV. Plan général de cette étude

Le plan adopté dans cette étude est assez inhabituel si on le compare à d'autres discussions du problème du libre arbitre. D'une certaine manière, on peut dire que j'aborde un certain nombre de questions dans le sens exactement inverse à l'ordre habituel. D'autre part, on pourrait trouver que j'accorde une place extrêmement importante à des questions strictement épistémologiques. Je répondrai brièvement à ces deux étonnements dans cette introduction. Mais auparavant, je vais présenter le plan pour lui-même.

Cette étude est divisée en deux grandes parties. L'articulation de ces deux parties se fait à partir de la notion de « justification *prima facie* » : en première approche, on peut dire qu'une croyance est justifiée « *prima facie* » si elle est présumée justifiée en l'absence d'argument contraire (ou de réfutateur).⁶³ Une croyance qui est justifiée *prima facie* peut être, ultimement, non justifiée à cause de la présence d'un réfutateur de justification, mais s'il n'y a pas de réfutateur, la justification *prima facie* permettra à la croyance en question d'être justifiée « toutes choses considérées », ou *ultima facie*. Dans cette étude, la première partie est consacrée à montrer que nous avons une justification *prima facie* à faire confiance à notre expérience du libre arbitre – c'est-à-dire à la fois une justification *prima facie* à croire le contenu de cette expérience (« nous sommes libres »), et une justification *prima facie* à considérer cette expérience comme une source fiable de connaissance (« nous savons que nous sommes libres »). La deuxième partie est consacrée à montrer que cette justification *prima facie* ne rencontre pas d'argument contraire décisif, pas de réfutateur. En conclusion, nous sommes absolument justifiés à nous fier à notre expérience du libre arbitre, et il aura été prouvé, chemin faisant, que cette expérience du libre arbitre a des conditions de vérité incompatibilistes ou libertariennes. L'originalité de mon plan tient au fait que cette analyse du contenu incompatibiliste de l'expérience arrive très tard dans la réflexion, soit au début de la deuxième partie. Toute la première partie de l'étude procède avec un concept de « libre arbitre » qui n'est pas analysé – ni dans un sens incompatibiliste, ni dans un sens compatibiliste. Je justifierai un peu plus loin ce choix de plan, qui peut paraître étonnant, mais les premières explications se trouvent dans la réponse que j'apporte au « problème de la définition » que je viens de présenter en section 1.1.

Au chapitre 1, je présente ma solution au problème de la définition. Pour fixer le concept de libre arbitre, on ne peut pas se contenter de « faire confiance » au lecteur, en lui rejetant la responsabilité de comprendre (sans aucune définition) de quoi l'on parle.

⁶³ La notion de justification *prima facie* sera introduite avec précision au chapitre 2 (sec. II.2.c.).

Malheureusement, on ne peut pas non plus, à mon sens, donner une définition qui spécifierait le contenu de ce concept, car une telle définition risquerait d'être controversée et donc de trancher de manière purement stipulative certains débats de fond concernant la nature du libre arbitre. La solution que j'adopte consiste à fixer le concept de libre arbitre non pas en spécifiant son contenu, mais en décrivant le rôle qu'il joue dans la discussion traditionnelle sur le libre arbitre. Or cette discussion repose à mon sens sur un conflit intuitif universel entre la notion de déterminisme et la conception que nous nous faisons de nous-mêmes et de notre action. Le concept de libre arbitre est donc défini comme ce concept que nous avons de nous-mêmes et de notre action, et qui a tendance à entrer en conflit intuitif avec la notion de déterminisme. Cette définition laisse totalement ouverte la question du *contenu* de ce concept (en particulier, ce contenu *pourrait* tout à fait être en fait conceptuellement compatible avec la notion de déterminisme). Mais elle fixe tout de même un point important : elle fixe le fait que nous avons, naturellement, « l'impression » (ou l'expérience) de posséder le libre arbitre, c'est-à-dire que, à tort ou à raison, nous croyons spontanément instancier ce concept. C'est grâce à cette expérience naturelle qu'il sera possible (en première partie) de soutenir un certain nombre de thèses à propos « du libre arbitre » tout en continuant de laisser ouverte la question de son contenu conceptuel déterminé.

Au chapitre 2, je soutiens que notre expérience d'avoir le libre arbitre nous donne une justification *prima facie* à croire que nous avons effectivement le libre arbitre. Cette thèse est la première étape d'une démarche épistémologique dans laquelle *l'expérience* du libre arbitre (l'expérience en première personne selon laquelle « je suis libre ») joue le rôle fondamental de source épistémique de la connaissance du libre arbitre. C'est ce que l'on a appelé traditionnellement une épistémologie du libre arbitre par « introspection ». Cette épistémologie n'est pas la seule qui ait été défendue. Je commencerai donc par quelques réflexions qui présenteront l'épistémologie introspectionniste et ses détracteurs, et qui expliqueront pourquoi cette épistémologie est à mon sens le meilleur candidat pour résoudre le problème d'intégration. Dans le reste du chapitre, je m'attacherai à montrer qu'on ne peut pas refuser à l'expérience du libre arbitre le statut de justification *prima facie* sans être mené à certaines formes radicales de scepticisme à propos de la perception sensible. Dans la mesure même où l'on veut éviter le scepticisme radical, on doit donc admettre la conclusion que l'expérience du libre arbitre peut justifier, *prima facie*, la croyance au libre arbitre.

Au chapitre 3, j'aborderai la question de « méta-justification », c'est-à-dire la question de savoir si nous sommes justifiés à croire que notre expérience du libre arbitre nous donne une connaissance du libre arbitre. Et là encore, je soutiendrai que, avant toute analyse du contenu de l'expérience, nous avons une justification *prima facie* à croire que nous savons (donc une « méta-justification *prima facie* »). J'emploierai ici le même type d'argument « anti-sceptique » que dans le chapitre 2 : je m'attacherai à montrer que si l'on refuse à notre expérience commune du libre arbitre la méta-justification *prima facie*, alors il faudra la refuser également pour les croyances perceptives, et on sera par conséquent conduit à une forme radicale de scepticisme.

La conclusion de la première partie est donc la suivante : si l'on se place en amont de l'analyse du contenu de notre expérience, alors cette expérience offre une justification *prima facie* à croire à son contenu (inanalysé) et nous sommes également justifiés *prima facie* à croire qu'elle apporte une *connaissance* au sens externaliste. En fin de première partie, je montrerai que cette conclusion peut être renforcée : non seulement nous avons une justification et une méta-justification *prima facie*, mais on peut montrer que les seules réfutations possibles de ces justifications seraient des réfutations d'ordre « scientifique » ou « spécialisé », c'est-à-dire soit des informations psychologiques ou neurologiques sur notre expérience du libre arbitre, soit des informations spécialisées sur la bonne analyse conceptuelle du contenu de notre expérience. Autrement dit, il n'existe aucun réfuteur

« préscientifique » de notre justification introspectionniste à croire au libre arbitre. Donc préscientifiquement, notre expérience du libre arbitre est *ultima facie* justifiée. Du point de vue du Défi d'Intégration, cette première partie revient à établir la légitimité de l'intuition de connaissance (IM) : *prima facie*, il semble bien que nous *connaissions* le libre arbitre par expérience.

La deuxième partie consistera à examiner s'il existe des réfutations scientifiques de la justification *prima facie*, et, tout spécialement, une réfutation qui viendrait d'une analyse conceptuelle (IA) qui serait métaphysiquement trop « lourde » pour pouvoir être épistémiquement intégrée.

Le chapitre 4 sera consacré à l'analyse du contenu conceptuel de notre expérience du libre arbitre. Je soutiendrai que notre expérience du libre arbitre a bel et bien un contenu incompatibiliste, c'est-à-dire que l'intuition analytique (IA), le deuxième pilier du Défi d'Intégration, repose sur des arguments solides. Ayant ainsi établi les deux premières thèses du Défi d'Intégration, la troisième (IE) sera examinée (et contestée) au chapitre 6. Pour ce qui est de la conclusion incompatibiliste du chapitre 4, je l'établirai au moyen de deux arguments indépendants, un argument abductif et un argument direct. L'argument abductif consiste à dire que, étant donné l'existence (et l'universalité) du conflit intuitif entre notre conception du libre arbitre et l'idée de déterminisme, l'explication la plus simple de ce conflit est qu'il y ait un conflit conceptuel réel (une incompatibilité) entre les deux concepts. Bien sûr, on pourrait chercher à donner une autre interprétation du conflit intuitif – une explication qui consisterait à dire que le conflit repose sur une *illusion* – mais je montrerai que les théories qui ont été proposées pour rendre compte de l'apparence incompatibiliste en termes d'illusion ne sont pas du tout convaincantes. Cet argument abductif restera, encore une fois, très neutre sur l'origine et la nature exacte de notre expérience du libre arbitre. Le second argument sera plus simple : il consistera à identifier notre expérience du libre arbitre et, une fois localisée, à examiner son contenu par une méthode « phénoménologique ». Je soutiendrai ici que notre expérience du libre arbitre provient de l'expérience en première personne que nous faisons au moment où nous faisons des choix. Et j'apporterai ensuite des éléments phénoménologiques tendant à montrer que cette expérience du choix a des composantes incompatibles avec le libre arbitre. Une fois que l'intuition incompatibiliste (IA) sera établie, il s'agira de voir si ce contenu « lourd » vient réfuter la justification *prima facie* que nous avons à croire au contenu de notre expérience.

Le chapitre 5 examinera une première manière dont le contenu incompatibiliste pourrait venir réfuter la justification *prima facie* : il pourrait se faire que ce contenu incompatibiliste soit logiquement impossible, ou bien que nous ayons des données scientifiques qui prouvent sa fausseté ou le rendent incompréhensible. C'est ce que j'appelle les « réfutations directes » : elles concernent directement la véracité du contenu, et non pas la possibilité d'en donner une épistémologie convaincante (ce qui sera abordé au chapitre suivant). Je tâcherai de montrer qu'aucune des réfutations directes qui ont été proposées par les anti-libertariens n'est convaincante. Parmi les objections directes, il y a l'objection moderne du déterminisme : je montrerai pourquoi cette objection scientifique n'en est plus une aujourd'hui. Mais les objections principales consistent à dire que l'idée même d'un libre arbitre qui contient des indéterminismes est incompréhensible, soit absolument incompréhensible *a priori*, soit inconcevable étant donné ce que nous savons des indéterminismes physiques réels. Pour répondre aux objections d'incompréhensibilité (*a priori* et empirique), j'aurai recours aux deux modèles d'incompréhensibilité libertariens les plus importants aujourd'hui, le modèle événementiel de Robert Kane et le modèle de la causalité agentive. La conclusion de ce chapitre sera donc que si il y a aujourd'hui une objection décisive contre le libertarianisme, cette objection *doit être* de nature

épistémique et non pas métaphysique – c'est-à-dire que le problème d'Intégration est le seul véritable défi que le libertarien ait encore à résoudre aujourd'hui.

Le chapitre 6 abordera enfin la question de l'Intégration proprement dite, en la situant au sein de la théorie des réfutations possibles de la justification *prima facie* : le problème d'Intégration serait une réfutation « fondationnelle », c'est-à-dire une réfutation qui ne concerne pas la plausibilité du contenu de l'expérience, mais la capacité pour l'expérience, de constituer une source épistémique fiable. Pour être complet, ce chapitre commencera par examiner une certaine catégorie de réfutations fondationnelles issues des neurosciences et de la psychologie : ces réfutations considèrent que des données scientifiques (celles de Libet ou de Wegner) discréditent notre expérience du libre arbitre indépendamment de tout contenu incompatibiliste. Ces objections ne sont pas encore le Défi d'Intégration puisqu'elles ne reposent pas sur l'intuition (IA) et la difficulté à intégrer un contenu lourd. Je montrerai que ces objections ne sont pas convaincantes. Enfin j'examinerai la question d'Intégration elle-même, et je montrerai que, si l'on part d'une épistémologie réaliste qui évite le scepticisme pour la perception sensible, il est en fait possible de donner des modèles épistémologiques dans lesquels notre expérience du libre arbitre confère une *connaissance* de l'indéterminisme libertarien. Je proposerai pour cela un modèle d'intégration pour le libertarianisme événementiel de Kane, et un modèle d'intégration pour la causalité agentive. Cette conclusion revient à dire que l'intuition d'inconnaissabilité (IE) n'est pas fondée, si l'on part d'une bonne épistémologie, et l'on résout ainsi le Défi d'Intégration dans un sens libertarien (en conservant (IM) et (IA)).

Pour simplifier, on peut lire le plan de cette étude de deux manières différentes. On peut le lire à partir du trilemme d'Intégration :

- (i) défense de l'intuition de connaissance (IM) (première partie),
- (ii) défense de la conception libertarienne (IA) (chapitres 4 et 5),
- (iii) rejet de l'intuition d'inconnaissabilité du libertarianisme (IE) (chapitre 6).

Ou bien, on peut le lire selon l'articulation « justification *prima facie* / réfutations » :

- (i) l'expérience du libre arbitre donne une justification *prima facie* à croire que *nous sommes libres* (chapitres 1 et 2)
- (ii) elle donne aussi une justification *prima facie* à croire que *nous savons que nous sommes libres* (chapitre 3)
- (iii) la justification *prima facie* à croire que *nous sommes libres* n'est pas réfutée (chapitres 4 et 5)
- (iv) la justification *prima facie* à croire que *nous savons que nous sommes libres* n'est pas réfutée non plus (chapitre 6).

Ce plan inhabituel mérite quelques mots de commentaire. Les deux reproches qu'on peut lui faire sont, premièrement, qu'il prend les questions traditionnelles « à l'envers », et deuxièmement qu'il consacre beaucoup trop de place aux préliminaires strictement épistémologiques (puisque dans les chapitres 2 et 3, nous aborderons des questions d'épistémologie sans avoir encore aucune notion déterminée du libre arbitre). Ces deux caractéristiques problématiques ont fondamentalement la même justification, qui est d'établir comme point de départ la solidité du fondement épistémologique de la croyance au libre arbitre. Mais voyons cela plus en détail.

Le premier reproche part de l'ordre habituel de présentation des différentes « questions ». Si l'on considère en particulier les défenses habituelles du libertarianisme, on trouvera en général un traitement des questions dans l'ordre suivant :

- (i) Question de Compatibilité : démonstration de l'incompatibilité du libre arbitre avec le déterminisme.
- (ii) Question d'Intelligibilité : construction d'un modèle dans lequel l'indéterminisme (requis d'après (i)) peut effectivement aider à comprendre le libre arbitre.
- (iii) Question d'Existence : le libre arbitre indéterminisme étant possible et concevable, recherche des données qui permettraient d'établir que nous l'avons effectivement.

et cette question (iii), dans le cas d'une démarche introspectionniste, se subdivise ensuite en deux questions :

- (iii a) Question Phénoménologique : avons-nous vraiment une expérience du libre arbitre libertarien ? Pouvons-nous la mettre en évidence ?
- (iii b) Question Epistémologique : cette expérience libertarienne permet-elle vraiment de justifier ou de donner une connaissance du libre arbitre indéterministe ?

Cet ordre de traitement des questions correspond à peu près à l'ordre chronologique dans lequel les différents problèmes sont réapparus dans la discussion au cours des cinquante dernières années.⁶⁴ Cet ordre a aussi une logique : à quoi bon donner un modèle du libre arbitre indéterministe (ii) si l'on n'a pas d'abord montré l'importance de l'indéterminisme (i) ? Et comment pourrait-on chercher l'existence du libre arbitre (iii) si l'on n'a pas une idée de ce que l'on cherche (ii) ? Et pourtant, mon plan consiste précisément à prendre ces questions à l'envers : dans la première partie, je commence par aborder la question de l'expérience (chapitre 1) et la question de la justification épistémologique (chapitres 2 et 3) avant toute analyse compatibiliste ou incompatibiliste. Je crois que le chapitre 1 montre comment cela est possible : dans la mesure où le concept de libre arbitre est *défini* à partir de notre expérience commune (et du conflit intuitif qu'elle génère), l'existence de cette expérience peut être établie sans dire « de quoi » elle est l'expérience précisément, et l'on peut également en tirer quelques conclusions épistémologiques à condition d'avoir une bonne épistémologie des expériences. Mais quand bien même cette démarche est *possible*, on peut se demander si elle n'est pas inutilement compliquée. Ne pourrais-je pas en effet commencer par mes chapitres 4 (incompatibiliste) et 5 (donnant les modèles du libertarianisme), et montrer ensuite que nous en avons l'expérience (puisque je prétends pouvoir le montrer) et que cette expérience donne toutes les justifications *prima facie* et *ultima facie* que je défends aux chapitres 2, 3, et 6 ? Cela serait en effet *possible*. Mais, d'après mon expérience des débats sur le libre arbitre, j'ai le fort sentiment que cet ordre de présentation joue, rhétoriquement, en défaveur de la crédibilité du libertarianisme. En effet, lorsqu'on commence par donner des modèles du libertarianisme extrêmement sophistiqués (philosophiquement et scientifiquement), et qu'on se demande *ensuite* s'il existe des données pour affirmer l'existence d'une telle faculté, le recours à l'expérience agentive semble presque immédiatement dérisoire d'un point de vue

⁶⁴ Pour simplifier, la question (i) est réapparue en premier dans les années 60 (notamment grâce aux articles de Ginet (1966) et van Inwagen (1975), comme je le rappelais plus haut), la question (ii) à partir des années 80 (grâce aux contributions de Robert Kane (1985; 1989), puis de Timothy O'Connor (1995)), et la question (iii a) est davantage une question des années 2000 (deux équipes de recherches, celle d'Eddy Nahmias et celle de Shaun Nichols, ont amorcé ce programme de recherche à peu près au même moment : voir (Nahmias et al. 2004) et (Nichols 2004)). Quant à la question (iii b), je milite pour qu'elle prenne plus d'importance dans les débats contemporains.

épistémique. Et tous les efforts qu'on pourra faire ensuite pour montrer que nos expériences donnent généralement une justification au moins *prima facie*, et que certaines expériences fortes doivent être présumées comme des sources de connaissance si l'on veut éviter le scepticisme pour la perception sensible, tous ces efforts quelle que soit leur ingéniosité ressembleront toujours à des tentatives désespérées pour profiter des obscurités du problème sceptique et en tirer, là où ça nous arrange, une épistémologie laxiste et permissive. Autrement dit, quand on commence par faire du libre arbitre une question technique et spécialisée, on se coupe presque définitivement la possibilité de recourir ensuite à une épistémologie de sens commun. Et pourtant, le libre arbitre est *d'abord* une expérience de sens commun, dotée d'une situation épistémologique assez solide, avant d'être analysable en des termes complexes et des modèles scientifiques, tout comme la perception sensible a d'abord une position toute particulière dans notre structure épistémologique de base, avant d'être analysable scientifiquement en termes de rayons lumineux et de projections inversées sur la rétine. Mon intention fondamentale est de replacer le problème du libre arbitre dans sa situation épistémologique originaire, c'est-à-dire repartir de la situation épistémologique privilégiée qu'il a dans notre système de connaissances avant toute réflexion philosophique et scientifique. Or cette situation épistémologique privilégiée n'est pas, à mon sens, un *deus ex machina* ou un dernier recours pour sauver le libertarianisme : c'est quelque chose que *tout* théoricien du libre arbitre (compatibiliste, libertarien, ou même incompatibiliste dur) devrait pouvoir reconnaître de manière non controversée). Cette première partie peut donc sembler inutilement « abstraite », dans son refus de définir le libre arbitre, mais je la crois difficilement contestable dans ses conclusions épistémologiques. Et pourtant, une fois que ses conclusions épistémologiques sont acceptées, on découvrira (je l'espère) de manière naturelle que la position épistémologique du libertarianisme est beaucoup moins menacée que ce que l'on aurait pu croire.

Ce plan révèle également ce qui est, à mes yeux, un ordre de priorité : comme je le défends au chapitre 1, la question du libre arbitre est pour moi, fondamentalement, le problème de la défense de notre expérience de sens commun : notre conception fondamentale de nous-mêmes est-elle illusoire ou non ? S'il s'avérait que cette expérience de sens commun n'est *pas* incompatibiliste, alors toute question épistémologique sur le libertarianisme serait, pour moi, sans grand enjeu. Inversement, si notre expérience est incompatibiliste, toute tentative pour intégrer épistémologiquement une autre notion (compatibiliste) de libre arbitre ne changera rien à la triste conclusion que nous vivons dans une illusion fondamentale. Pour cette raison, toute question phénoménologique *déterminée* (avons-nous oui ou non l'expérience de *tel type* de libre arbitre ?) est pour moi épistémologiquement secondaire : il ne fait aucun doute, par définition, que nous avons une expérience *du libre arbitre*, et c'est bien elle, et pas d'autres expériences éventuelles, qui nous intéresse.

Ainsi, l'ordre étonnant de mon plan, aussi bien que la longueur des considérations strictement épistémologiques, est justifiée par mon intention de caractériser précisément le « point de départ épistémologique » de la question du libre arbitre, c'est-à-dire le fait que, préscientifiquement, nous avons toutes les raisons de croire à notre expérience et de la tenir pour une véritable *connaissance*. La *vraie* question du libre arbitre consiste simplement à savoir si certaines informations au cours de l'histoire (informations scientifiques, philosophiques, ou phénoménologiques) ont changé quelque chose à ce point de départ. Ma réponse est non : la philosophie et les sciences nous ont apporté toutes sortes d'informations passionnantes sur ce qu'est le libre arbitre et comment nous le connaissons, mais d'un point de vue de notre structure épistémologique fondamentale, nous sommes toujours justifiés, comme au premier jour, à nous fier à notre expérience du libre arbitre comme à une source de connaissance. Il en va ici comme de la perception sensible : la philosophie et les sciences nous ont apporté toutes sortes d'informations intéressantes sur ce qu'est la perception et comment

elle fonctionne, mais d'un point de vue d'épistémologie fondamentale, nous avons toujours le même type de justification (non scientifique et non technique) pour nous fier à la perception comme à une source de connaissance.

PREMIERE PARTIE :
PRESOMPTION DE CONNAISSANCE DU LIBRE ARBITRE

Si dans le temps que nous sentons notre liberté à l'égard d'un bien particulier, nous devons douter que nous soyons libres, à cause que nous n'avons point d'idée claire de notre liberté, il faudra aussi douter de notre douleur et de notre existence dans le temps même que nous sommes malheureux, puisque nous n'avons point d'idée claire ni de notre âme, ni de notre douleur, mais seulement sentiment intérieur.

(Malebranche 1979, 809)

Chapitre 1 : L'expérience commune et la définition de « libre arbitre »

The concept of a person [...] can be construed as the concept of a type of entity for whom the freedom of its will may be a problem.
(Frankfurt 1971, 14)

A definitione liberi arbitrii socialismus pendet.
(Jaurès 1891, sec. I)

I.	Difficultés pour définir le concept de « libre arbitre »	41
1.	Problème de la polysémie du mot « libre »	41
2.	La solution « confiante »	42
3.	La solution « stipulative »	44
4.	La solution « descriptiviste »	47
II.	Le conflit intuitif originel : solution descriptiviste pour fixer la référence du terme « libre arbitre »	49
1.	Emergence du problème du libre arbitre : le syndrome d'Eleuthère	49
2.	Du problème traditionnel à la description définie du concept de « libre arbitre »	50
3.	Définitions diverses : « Problème psychologique », « problème philosophique », incompatibilisme, compatibilisme, etc.	53
4.	Les présupposés de cette définition	57
III.	L'existence du conflit intuitif, et l'expérience du libre arbitre	63
1.	L'existence du conflit intuitif naturel	63
2.	Du conflit intuitif à l'expérience du libre arbitre	69
IV.	Conclusion : L'expérience du Libre arbitre comme point de départ	76

Ainsi que je l'ai annoncé en introduction, la définition du concept de « libre arbitre » constitue à elle seule une question très délicate et un vrai problème méthodologique pour pouvoir simplement *poser* la question du libre arbitre – et *a fortiori* la question de la *connaissance* du libre arbitre. On pourrait faire le même genre de remarques à propos de la définition des concepts épistémologiques (connaissance, justification), mais pour ces concepts, j'ai pu recourir à un cadre terminologique qui, s'il n'est pas absolument consensuel, n'engage pas réellement de présupposés importants pour la discussion présente. L'intérêt de la terminologie épistémologique retenue est *méthodologique* (il permet de rendre la discussion présente plus claire) et non pas substantiel. Il n'en va pas de même pour la définition du concept de libre arbitre : la manière dont je résous ce problème définitionnel est à la fois originale et fondamentale pour mon argument d'ensemble. Ma solution consiste à définir le concept de libre arbitre non pas classiquement au moyen d'une définition réelle plus ou moins large (« la faculté telle que ... ») mais par référence à notre *expérience* commune de la liberté, et à son contenu (quel qu'il soit). Cette définition aura des conséquences importantes pour la question épistémologique puisque je défendrai ensuite que la *connaissance* du libre arbitre

repose sur cette même expérience. Si cette définition est inhabituelle dans la lettre,⁶⁵ il me semble qu'elle correspond à l'*usage* implicite de la plupart des philosophes. Après tout, la question principale à propos du libre arbitre est de savoir s'il s'agit ou non d'une *illusion* : mais pour qu'il y ait apparence illusoire ou non illusoire, il faut bien qu'il y ait une apparence, c'est-à-dire une certaine forme d'expérience, au moins dans un sens faible du mot « expérience ».⁶⁶

Ce premier chapitre sera donc consacré à la définition du libre arbitre par l'expérience que nous en avons. Dans une première section, je poserai de manière systématique le problème de la définition du terme « libre arbitre » tel qu'il se présente dans le débat philosophique contemporain, et je proposerai une classification des solutions possibles : solution confiante, solution stipulative et solution descriptiviste. En définitive, je soutiendrai que seules les solutions de type descriptiviste sont satisfaisantes, et dans la deuxième section, je montrerai que la définition par l'expérience commune est le meilleur candidat pour une solution descriptiviste. Enfin, dans la dernière section, je donnerai quelques éléments pour bien comprendre en quoi consiste cette « expérience » commune.

I. Difficultés pour définir le concept de « libre arbitre »

Qu'est-ce que le Libre arbitre ? C'est la faculté qu'on attribue à un agent lorsqu'on dit de lui (à tort ou à raison) qu'il est libre, en général, ou plus spécifiquement qu'il a librement accompli telle ou telle action. Quand je dis que De Gaulle s'est librement engagé dans la résistance, j'attribue (à tort ou à raison) le Libre arbitre à De Gaulle.

Fort bien, me direz-vous, mais quand on a dit cela, on n'a rien dit ! On a fait comme ces dictionnaires qui définissent la diligence comme la « qualité d'une personne diligente ». Cette réaction est partiellement justifiée. Mais seulement partiellement. Car de la même façon que le dictionnaire nous signale que la diligence, dans ce sens, est une *qualité* (et non pas un objet matériel à quatre roues, à la différence de l'*autre* sens du mot « diligence ») et qui plus est une qualité de *personnes*, de la même façon notre première définition du libre arbitre précise de quel genre de chose il s'agit : le libre arbitre est la *faculté* d'un *agent* qui agit librement, et non pas, par exemple, la propriété d'une *action* accomplie librement. Si ce genre de définition (celle de la diligence comme celle du libre arbitre) nous laisse néanmoins insatisfaits, c'est qu'elle repousse à un autre terme (un terme de la même famille lexicale) ce qu'on pourrait appeler le « cœur » de la notion. Dans un dictionnaire, ce genre de problème est réglé par un rapide coup d'œil à la définition suivante : « Diligence... Diligent... Ah, voilà : adjectif qualificatif, etc. ». Jetons donc nous aussi un rapide coup d'œil au mot suivant.

1. Problème de la polysémie du mot « libre »

Qu'est-ce qu'être libre ? Ou encore, qu'est-ce qu'agir librement pour un agent ? C'est malheureusement ici que les choses se compliquent. Les complications viennent de deux directions opposées. D'une part, il est clair qu'on ne peut pas en rester à ce terme de « libre » en le prenant pour terme primitif, comme Euclide peut en rester au terme de ligne ou d'angle

⁶⁵ Inhabituelle, mais pas totalement absente de la littérature. Voir notamment Richard Holton qui affirme également que l'expérience du libre arbitre devrait être le critère fondamental pour savoir si l'on est dans le sujet ou hors sujet : « Once we start to contemplate the experience of free will, much of the literature on it seems beside the point » (Holton 2006, 1).

⁶⁶ Nous verrons plus loin dans ce chapitre de quelle notion d'expérience il est ici précisément question (cf. II.2).

sans avoir besoin de les définir parce qu'ils sont compris par tous de manière univoque et que tout essai de définition devrait avoir recours à des termes moins connus et moins compris. Le problème du mot « libre » à cet égard, c'est qu'il n'est pas du tout univoque. Quand on dit que « depuis hier soir 21h00, Ingrid Betancourt est libre » ou que « depuis sa psychanalyse, Sigmund se sent beaucoup plus libre », on ne parle évidemment pas du même genre de capacités chez ces personnes. Et lorsqu'un métaphysicien parle du « libre arbitre », c'est encore un troisième sens de « libre » qu'il a à l'esprit. On est donc bien obligé de préciser de *quel* sens il s'agit. Mais on rencontre alors la complication inverse : dès qu'on essaye de définir le sens de « libre » qui nous intéresse, on risque d'utiliser des termes ou des caractéristiques qui ne font pas l'accord de tous les philosophes. Par exemple, on pourrait être tenté de définir le libre arbitre comme le sens de liberté « requis par la responsabilité morale » ; mais certains philosophes mettent en doute que la liberté soit requise pour la responsabilité morale. Inversement, Peter van Inwagen stipule que le libre arbitre est « la faculté d'agir autrement que l'on a agi, dans les mêmes circonstances » ; pourtant certains philosophes soutiennent que le libre arbitre peut se concevoir même dans des scénarios où l'agent n'aurait pas pu agir autrement. Par conséquent, dire que le libre arbitre est essentiel à la responsabilité morale, ou qu'il implique la faculté d'agir autrement, semblent être des *thèses* à propos du libre arbitre ; ces thèses sont peut-être initialement intuitives, sans doute défendables, mais l'exigence de clarté philosophique exige de donner des définitions qui soient exemptes de tout parti-pris.

Le problème est donc le suivant : soit on en reste au mot « libre » non défini, mais dans ce cas, on n'a pas précisé de quoi on parle, car ce mot a plusieurs acceptions ; soit on essaye de le définir, mais dans ce cas, il semble qu'on soit obligé de charger notre définition de présupposés qui biaiseront la discussion.

Je vois trois solutions à ce problème.

2. La solution « confiante »

La première solution, ou *solution confiante*, consiste à minimiser la difficulté que pose l'ambiguïté. « Certes, il y a différents usages de la notion de liberté, un usage politique, un usage psychologique, etc. mais personne ne confondrait sérieusement ces genres de liberté entre eux. Et il est clair que nous ne parlons pas de ces genres de liberté-là, mais de la liberté au sens métaphysique. » Le fait d'appeler « métaphysique » le sens que nous cherchons à isoler ne nous apporte évidemment aucune information, puisque nous savions d'entrée de jeu, que nous étions dans une discussion métaphysique. Cela revient donc à dire, assez trivialement, que « le sens de liberté dont parlent les discussions métaphysiques est le sens métaphysique de la liberté ». Mais la première solution que j'envisage ici a conscience de n'apporter aucune aide au lecteur pour distinguer et isoler le sens de « liberté » dont il va être question : cette solution *fait confiance* au lecteur, elle fait appel à sa bonne volonté ou à son bon sens pour distinguer de lui-même un concept qu'on ne parvient pas à distinguer verbalement. Le problème de cette solution confiante, c'est qu'elle se rapproche dangereusement de la paresse spontanée par laquelle les auteurs sans méthode et les auditoires sans exigence s'autorisent à manipuler des grandes notions non définies tout en se donnant l'impression qu'ils « voient bien » de quoi il retourne. Généralement, cette compréhension à demi-mot n'est guère plus qu'une demi-compréhension.

Plus fondamentalement, le problème de la solution confiante est qu'elle laisse assez peu de place à une discussion argumentée, notamment sur la question de savoir si le libre arbitre est compatible ou non avec le déterminisme causal. Si le concept de « libre arbitre » est donné sans aucune analyse conceptuelle, il semble que le philosophe « compatibiliste » et le

philosophe « incompatible » pourront seulement s'opposer l'un à l'autre leur intuition brute de compatibilité ou d'incompatibilité. Historiquement, la discussion philosophique a pourtant réussi à proposer mieux qu'un conflit direct d'intuitions.

En particulier, les auteurs compatibilistes ont traditionnellement proposé des *analyses* de la notion de libre arbitre permettant d'argumenter en faveur du compatibilisme. On peut prendre deux exemples célèbres. Tout d'abord l'analyse conditionnelle « classique » que l'on trouve notamment chez Hume :

Par liberté, nous pouvons seulement entendre un pouvoir d'agir ou de ne pas agir selon les déterminations de la volonté ; c'est-à-dire, si nous choisissons de rester en repos, nous le pouvons, si nous choisissons de nous mouvoir, nous le pouvons aussi.
(Hume 1983, 164)

A partir de cette analyse conditionnelle, on peut aisément argumenter que le libre arbitre est compatible avec le déterminisme : les actes peuvent fort bien dépendre causalement des volontés, même si les volontés dépendent elles-mêmes de facteurs causaux antécédents. Deuxième exemple, l'analyse « hiérarchique » de Harry Frankfurt (1971). Pour Frankfurt, la liberté de la volonté (*freedom of the will*)⁶⁷ est une condition qui porte sur le rapport entre nos volitions de premier ordre et nos volitions de deuxième ordre (les volitions d'avoir telle ou telle volition de premier ordre) :

Dire qu'une personne possède la liberté de la volonté signifie [...] qu'elle est libre d'avoir les volontés qu'elle veut. [...] C'est donc en réalisant la conformité de sa volonté avec ses volitions de deuxième ordre qu'une personne exerce la liberté de la volonté. (Frankfurt 1971, 15)

Et à partir de cette analyse hiérarchique, on peut également argumenter en faveur du compatibilisme : mes volitions de deuxième ordre peuvent fort bien contrôler mes volitions de premier ordre même si elles sont elles-mêmes causalement déterminées par des facteurs antécédents.

Les auteurs incompatibilistes ont également proposé des analyses du libre arbitre permettant d'argumenter en faveur de leur thèse. Là aussi, je ne présenterai pour l'instant que les deux analyses les plus importantes. D'une part, l'analyse en termes de « possibilités alternatives » dont Peter van Inwagen (1983) a été un des défenseurs les plus importants :

La thèse du libre arbitre est la thèse selon laquelle nous sommes parfois dans la position suivante à l'égard d'un acte futur considéré : nous avons simultanément les capacités suivantes : la capacité d'accomplir cet acte et la capacité de nous retenir d'accomplir cet acte. (van Inwagen 2008, 329)

A partir de cette conception du libre arbitre, van Inwagen a défendu que le déterminisme excluait l'existence de possibilités alternatives (c'est le très célèbre « argument de la conséquence » dont nous reparlerons au chapitre 4). De nombreux auteurs incompatibilistes ont utilisé une notion de possibilités alternatives, mais depuis une quinzaine d'années une

⁶⁷ Ce n'est pas un hasard si Frankfurt n'emploie pas l'expression « free will » mais plutôt l'expression « freedom of the will ». Il y aurait donc lieu à discuter sur le rapport qu'entretiennent chez lui les deux expressions. Mais il est clair qu'un grand nombre d'auteurs ont considéré et utilisé son analyse de la « liberté de la volonté » comme une bonne analyse (compatibiliste) du libre arbitre à proprement parler. C'est dans ce rôle particulier que je l'utilise ici.

autre notion a pris au moins autant d'importance, la notion de « source » dont une version importante est la « responsabilité ultime » définie par Robert Kane :

L'idée est que la responsabilité ultime se situe à l'endroit de la cause ultime – là où s'arrête la régression. C'est l'image traditionnelle du libre arbitre. (Kane 1996, 4)

A l'issue de toutes ces discussions historiques, nous nous retrouvons donc avec (au moins) quatre analyses différentes du libre arbitre : le libre arbitre humien (LA1), le libre arbitre frankfurtien (LA2), le libre arbitre des alternatives (LA3) et le libre arbitre de la responsabilité ultime (LA4). Une description plus sophistiquée du débat en donnerait encore un grand nombre d'autres. C'est-à-dire qu'on se retrouve devant un nouveau problème de polysémie, un problème de « polysémie philosophique » pour lequel la solution confiante n'est clairement pas applicable. Autant on peut faire confiance à des auditeurs intelligents et bienveillants pour comprendre que lorsqu'on parle du libre arbitre on ne parle pas du fait de « ne pas être prison » ou « d'avoir le droit politique de circuler sur le territoire national », autant il serait absurde de la part d'un compatibiliste de « faire confiance » à son interlocuteur incompatibiliste pour que celui-ci accepte une analyse conditionnelle du libre arbitre.

3. La solution « stipulative »

On vient de voir que la solution confiante présentait un grand risque de confusion, où personne ne sait si l'on parle bel et bien d'une même chose. La seconde solution, ou *solution stipulative*, cherche à éviter ce risque de confusion en *choisissant* pour le mot « libre » une définition précise, et par exemple l'une des quatre définitions que l'on vient de présenter. Un partisan de cette solution raisonnera de la façon suivante :

« Puisque nous ne pouvons faire de philosophie sérieuse sans donner des définitions précises, il faut bien stipuler un sens pour le mot 'libre' avant de pouvoir l'utiliser. Par exemple, nous définirons le libre arbitre comme 'la faculté d'agir autrement dans les mêmes circonstances'. Bien sûr, c'est là une certaine 'conception' du libre arbitre, si on entend par là que c'est un certain *concept* et qu'on pourrait associer à ce même mot d'autres concepts (en faisant d'autres stipulations). Mais si quelqu'un nous reproche d'avoir ainsi 'présupposé' une thèse lourde 'au sujet du libre arbitre' – en ayant, par exemple, à l'esprit que d'autres philosophes considèrent comme 'libres' des agents qui n'auraient pas pu faire autrement mais ont toutes les caractéristiques requises pour être moralement responsables – la réponse sera très simple : notre définition n'affirme aucune thèse 'au sujet du libre arbitre' (au sujet de la chose libre arbitre), elle stipule simplement ce que le *mot* 'libre arbitre' veut dire dans le cadre de notre discussion. Si d'autres philosophes disent qu'on peut avoir le 'libre arbitre' sans avoir la faculté d'agir autrement, c'est qu'ils utilisent le mot 'libre arbitre' pour désigner autre chose que ce que nous désignons par ce terme, et non pas qu'ils affirment 'au sujet de la même chose, le libre arbitre' une thèse différente. Nous n'avons donc aucun désaccord avec ces philosophes ; nous avons seulement une terminologie différente dans laquelle les mêmes questions seront traduites de manière différente. Par exemple, si un philosophe entend par libre arbitre « la faculté (quelle qu'elle soit) qui est requise par la responsabilité morale », alors notre question 'le libre arbitre est-il requis pour la responsabilité morale ?' devra simplement être retraduite dans leur terminologie : 'la faculté d'agir autrement est-elle requise pour être libre ?'. On peut déplorer que les philosophes utilisent des terminologies aussi différentes (et peut-être déroutantes) pour poser les mêmes questions, mais ce n'est certainement pas en renonçant aux définitions claires et précises qu'on résoudra ce problème d'obscurité. »

Cette solution stipulative lorsqu'elle est mal comprise, peut être la cause de véritables « blocages » dans la discussion. Généralement, ces blocages prennent la forme du petit dialogue suivant entre les philosophes A et B :

A : Vous ne pouvez pourtant pas affirmer comme cela, sans argument, que la liberté requiert le pouvoir d'agir autrement ! Ce serait faire l'impasse sur une tradition considérable d'analyses conditionnalistes, ou hiérarchiques, de la liberté.

B : Je ne l'affirme pas : je le *stipule*.

A : ?!⁶⁸

Et quoi que B puisse ajouter ensuite, A reste en général dans l'impression perplexe que, par cette réponse, B ne suit pas les règles du jeu philosophique. Ce qui est certain, c'est que dans ce genre de dialogue A et B ne suivent pas les *mêmes* règles, sinon pour la philosophie en général, du moins pour l'usage particulier du mot « libre arbitre ». Mais il est important, pour lever l'incompréhension, de bien comprendre quelle règle chacun suit, et pourquoi la décision de B (c'est-à-dire la solution stipulative) est contestable.

La décision de B consiste à stipuler un sens pour le mot « libre », c'est-à-dire à faire abstraction de toute signification que ce mot pourrait avoir antérieurement, pour pouvoir poser ensuite différentes questions concernant le seul sens stipulé. Si A ne peut comprendre ou ne peut accepter cette attitude, c'est qu'il part quant à lui du sens pré-philosophique du mot « libre ». Pour A, le problème du libre arbitre est un problème à propos de ce sens pré-philosophique (ou plutôt, à propos de la *chose* que désigne ce sens-là). Par conséquent, si B donne une nouvelle définition stipulative pour le mot « libre », il pourra peut-être poser certaines questions intéressantes à propos de ce nouveau concept, mais il ne sera pas en train de parler du problème qui intéresse A, du problème qui – pour A – est le *vrai* problème du libre arbitre.

Prenons un autre exemple, celui d'un mathématicien A qui s'interroge sur la possibilité de la quadrature du cercle. Son interlocuteur B, qui sait qu'en mathématique les stipulations sont libres, choisit de stipuler que le mot « cercle » désignera l'eau, et que le mot « quadrature » désignera le phénomène d'ébullition. Une fois ces stipulations posées, B se fait fort de pouvoir résoudre le problème de la « quadrature du cercle » et allume le gaz. Il est clair que A sera profondément insatisfait par une telle stratégie, car elle ne revient pas en réalité à résoudre le *vrai* problème de la quadrature du cercle, mais seulement à utiliser l'expression « quadrature du cercle » pour désigner un *autre* problème. Bien sûr, A conviendra que la résolution du vrai problème supposera une définition rigoureuse de notre notion de « cercle » (par exemple, « une ligne fermée dont tous les points sont équidistants d'un même point »), mais pour que cette définition ne nous fasse pas passer à un autre problème, elle ne devra pas être une pure stipulation : elle devra au contraire donner une analyse fidèle de la signification que ce terme avait *déjà* avant même qu'on entreprenne de le définir.

De même, pour poser le « vrai » problème du libre arbitre, il faut peut-être donner une définition de « libre », mais cette définition ne doit pas être une pure stipulation, elle doit prétendre à être une analyse conceptuelle fidèle du sens pré-philosophique de « libre ». Malheureusement, ces exigences du philosophe A nous ramènent une nouvelle fois à notre problème initial, car il n'y a pas *un* sens pré-philosophique de « libre » : il y en a plusieurs.

Par conséquent, de *quel* sens notre définition devra-t-elle être l'analyse conceptuelle fidèle ? Ne faut-il pas une définition pour déterminer de quel sens on va parler ? Et si cette définition est un *préalable* à l'analyse conceptuelle, elle ne sera pas, elle-même, une analyse

⁶⁸ Ce dialogue est inspiré d'un dialogue réel entre Peter van Inwagen et un philosophe compatibiliste.

conceptuelle. Telle est en tout cas la motivation profonde de la position stipulative : face à la multiplicité des sens pré-philosophiques de « libre », il faut bien stipuler, par une définition claire et précise, de quel sens nous choisissons de parler, quitte à définir *plusieurs* sens précis de « libre arbitre » (LA1, LA2, etc.).

La version extrême de la solution stipulative consiste en effet à subdiviser les problèmes « du » libre arbitre en autant de versions qu'on peut construire de concepts LAn. C'est la solution adoptée par Mark Balaguer (2010) qui la présente comme la seule manière de rendre le débat conceptuellement clair et précis. Cette solution entraîne une reconfiguration assez radicale du problème du libre arbitre. Dans la présentation traditionnelle, on pourrait dire que le problème du libre arbitre engage essentiellement deux questions :

(Existence) Avons-nous le Libre arbitre ?

(Compatibilité) Le libre arbitre est-il compatible avec le déterminisme ?

Si l'on subdivise le libre arbitre en autant de concepts distincts, la question d'existence se transforme, comme le montre Balaguer, en question des « types » de liberté :

(Types de liberté) Quels types de libre arbitre avons-nous : LA1, LA2, etc. ?⁶⁹

Quant à la question de compatibilité, Balaguer fait observer que sa stratégie a pour conséquence de *trivialiser* cette question : en effet, pour chacun des concepts de la liste, ce seront soit des analyses d'auteurs compatibilistes, et elles seront alors *évidemment* compatibles avec le déterminisme, soit des analyses d'auteurs incompatibilistes, et elles seront alors *évidemment* incompatibles avec le déterminisme.⁷⁰ Autrement dit, il n'y a pas lieu de poser sérieusement la question de compatibilité « subdivisée » : quels types de libre arbitre (LA1, LA2, etc.) sont compatibles avec le déterminisme ? Pour chaque type, la réponse sera triviale. S'il reste quelque chose de la question de compatibilité, Balaguer conclut qu'elle se réduira entièrement à la question suivante :

(Question Qu'est-ce-que-le-libre arbitre) Quel type de libre arbitre est vraiment *le* « libre arbitre » : LA1, ou LA2, ou etc. ?⁷¹

C'est la question qui préoccupe le philosophe A et qui l'empêche d'accepter la stipulation de B : pour lui, il y a *un* sens privilégié de « libre arbitre », et quelque intérêt qu'on puisse avoir pour les facultés LA1, LA2, etc. en elles-mêmes, notre intérêt principal est de savoir *laquelle*

⁶⁹ Voir (Balaguer 2010, 27). Cette question sera aisée pour certains types de liberté, et compliquée pour d'autres : il sera aisé d'accorder que nous avons la liberté humienne, et sans doute plus difficile de savoir si nous avons liberté de la responsabilité ultime.

⁷⁰ Plus précisément, Balaguer fait cette remarque à propos des concepts compatibilistes : « It is important to note, however, that like Humean freedom, all of these other compatibilist notions of freedom (or at any rate, all the ones that I'm aware of) are not just compatible with determinism, but *obviously* so. » (Balaguer 2010, 5). On pourrait sans doute être plus réservé sur l'idée que les analyses incompatibilistes sont *trivialement* incompatibles avec le déterminisme. En particulier, certains auteurs compatibilistes ont défendu qu'on pouvait rendre compte d'un certain sens de possibilités alternatives même dans un schéma déterministe. Mais cette objection ne fait que repousser le problème d'un cran si l'on adopte la stratégie de Balaguer : en effet, cela revient à dire que lorsqu'on est conceptuellement précis, il y a en fait plusieurs concepts de « libre arbitre des alternatives », certains LAA1, LAA2, etc. qui sont analytiquement compatibilistes, d'autres LAA3, LAA4, etc. qui sont analytiquement incompatibilistes.

⁷¹ Voir (Balaguer 2010, sec. 2.5) : « Why the Compatibilism Question reduces to the What is Free Will Question ». Il est clair que si l'on a une réponse à cette question – par exemple « *le* libre arbitre est LA3 – on pourra ensuite directement répondre à la question de compatibilité puisque pour chaque sens précis LAn la question de compatibilité est triviale.

de ces analyses capture ce sens privilégié. Balaguer ne rejette pas absolument cette question, et on va voir un peu plus loin le sens qu'il lui donne, mais il est clair que cette question nous replace en face du problème de départ : pour donner un sens à cette question, il faut donner un sens à l'expression « libre arbitre » telle qu'elle apparaît dans la seconde moitié de la question. Or pour lui donner un sens, on ne peut pas recourir simplement à la solution confiante (comme on l'a vu), et on ne peut plus non plus recourir à une solution stipulative puisqu'il s'agit justement de discriminer entre les sens stipulatifs LA1, LA2, etc. Il faut donc trouver une autre solution.

Avant de passer à la dernière solution, je soulignerai un dernier avantage de la solution subdivisée de Balaguer. Cette solution permet de poser de façon précise un grand nombre d'autres questions ayant une certaine pertinence dans le débat sur le libre arbitre. La première de ces questions a été formulée dans sa version subdivisée dès 1984 par Daniel Dennett : « quelles variétés de libre arbitre ont une véritable importance pour nous ? ». ⁷² Balaguer présente quelques autres questions de ce type, et on pourrait aisément allonger sa liste :

- (i) Quels types de libre arbitre ont une importance pour nous : LA1, LA2, etc ?
- (ii) Quels types de libre arbitre sont requis par la responsabilité morale ? ⁷³
- (iii) Quel type de libre arbitre correspond à la bonne sémantique de l'expression « libre arbitre » du langage ordinaire ? ⁷⁴
- (iv) Quels types de libre arbitre sont des concepts logiquement cohérents ? ⁷⁵
- (v) Quels types de libre arbitre font partie de nos croyances de sens commun ?
- (vi) Quels types de libre arbitre font partie de notre *expérience* agentive naturelle ?

Ces questions « subdivisées » ont un avantage considérable en termes de clarification des débats, et elles gardent en elles-mêmes une pertinence *même* si l'on donne par la suite un rôle privilégié à une analyse particulière (qui représentera le « vrai » sens de « libre arbitre »). Elles réapparaîtront donc à différents moments de mon étude.

Mais dans l'immédiat, ce que nous cherchons est une solution pour définir le concept de « libre arbitre » de manière à la fois claire et non stipulative.

4. La solution « descriptiviste »

La troisième solution, ou *solution descriptiviste*, essaye de concilier les exigences de clarté de la solution stipulative et les exigences de fidélité à la discussion originelle de la solution confiante. La solution confiante exige que l'on conserve le sens de « libre » qui est à l'origine de la discussion métaphysique sur le libre arbitre et que toute définition (si l'on en donne) soit proposée comme une analyse de ce concept préalablement isolé et identifié, une analyse à discuter et non à présupposer. La solution stipulative exige que l'on détermine, de manière univoque et préalablement à tout débat, le sens de « libre » qu'on veut utiliser. La solution descriptiviste enfin fait remarquer que ces deux exigences sont conciliables : il existe un moyen de déterminer de manière univoque quelque chose sans en donner une définition. Ce moyen est connu des logiciens sous le nom de « description définie ». Par exemple, si je parle du « premier homme à marcher sur la Lune », je sélectionne par cette description un individu et un seul, à savoir Neil Armstrong. Pourtant, cette description ne renvoie pas à une propriété essentielle de Neil Armstrong (cette personne aurait pu choisir de faire un autre

⁷² « *The varieties of Free Will worth wanting* », sous-titre de (Dennett 1984).

⁷³ Pour cette question et la précédente, voir (Balaguer 2010, sec. 2.4.3).

⁷⁴ Voir (Balaguer 2010, sec. 2.2).

⁷⁵ Voir (Balaguer 2010, sec. 2.4.2)

métier, il aurait pu mourir en bas âge, etc.), et elle ne saurait donc compter comme « définition » (pour autant que l'idée de « définir » Neil Armstrong ait un sens). Mais prenons un exemple plus proche de notre problème : si je parle du « régime politique inventé en Grèce au V^{ème} siècle et qui s'est considérablement étendu à partir des Lumières », je ne donne pas une *définition* de la démocratie, mais j'ai pourtant donné une description qui y fait référence de manière univoque. L'intérêt de cette description, c'est qu'elle laisse ensuite le champ libre aux philosophes pour discuter de ce qu'est, dans son essence, ce régime. Est-il défini uniquement par un rapport d'identité entre justiciables et pouvoir législatif ? L'égalité de parole en est-elle un élément essentiel ? Toutes ces questions de définition peuvent rester ouvertes et être débattues sans que les débats ne virent au dialogue de sourds où chacun parle d'une chose différente (la « démocratie-1 », la « démocratie-2 », etc.). En effet, le concept visé ayant été fixé par une description définie, les désaccords seront simplement des thèses différentes sur la bonne analyse du concept. Ce genre de solution peut être appliqué avec grand profit au problème du libre arbitre.

Quelle sera donc la meilleure description définie du concept de « libre arbitre » dans le sens où l'entend le problème originel du libre arbitre ? Une première réponse possible est proposée par Balaguer. Pour Balaguer (2010, sec. 2.2), la façon la plus simple d'interpréter la question « Qu'est-ce que le libre arbitre ? » est de la considérer comme une question de sémantique du langage ordinaire, c'est-à-dire de la tenir pour équivalente à la question :

(iii) Quel type de libre arbitre correspond à la bonne sémantique de l'expression « libre arbitre » (*free will*) du langage ordinaire ?

Si on adopte cette interprétation, on dispose alors d'une description définie qui permet de fixer la référence de l'expression « libre arbitre » dans le débat philosophique, à savoir :

le concept signifié par l'expression linguistique « libre arbitre » (*free will*) dans le langage ordinaire.

Une fois fixée la référence, ce sont ensuite les linguistes, ou peut-être les « philosophes expérimentaux », qui permettront de savoir quel concept (LA1, LA2, etc.) est désigné par cette description. Cette solution descriptiviste a l'avantage d'être simple, mais à mon sens, elle n'est pas satisfaisante, principalement parce qu'elle rend la détermination du concept de libre arbitre dépendante d'un système linguistique particulier. Là où l'anglais utilise l'expression « *free will* », le français ne parle pas de « *volonté libre* » mais de « libre arbitre » (suivant en cela le latin « *liberum arbitrium* »). Par ailleurs, aucune de ces deux langues ne peut aisément se rapporter au langage ordinaire pour fixer le sens de l'expression. C'est particulièrement vrai pour le français, puisque « libre arbitre » n'appartient *pas* au langage ordinaire, et qu'on ne peut pas non plus le décomposer en deux composants sémantiques appartenant au langage ordinaire. Mais le cas de l'anglais n'est pas clairement meilleur : certes l'expression « *free will* » existe dans le langage ordinaire, mais c'est surtout au sein de l'expression « *doing something of one's own free will* » (faire quelque chose de son plein gré) qui ne semble pas du tout adaptée pour formuler le problème philosophique du libre arbitre. Une solution serait de se libérer du deuxième terme (*will*, *arbitrium*) pour se concentrer sur la sémantique des termes *free*, *libre*, *liberum* qui a plus de chances d'être inter-linguistique. Mais on retombe alors dans le problème de la polysémie linguistique de cet adjectif. Donc si l'on cherche à déterminer « le sens de libre arbitre tel que l'entend le problème originel du libre arbitre » (à travers ses versions latines, grecques, françaises, anglaises), il ne semble pas qu'on puisse s'en remettre à la sémantique du langage ordinaire.

Comment donc trouver une description définie permettant de capturer le sens de libre arbitre dans lequel l'entend le problème originel ? La réponse (le cœur de la réponse en tout cas) est dans la question : la description « le sens de 'libre arbitre' dans lequel l'entend le problème originel du libre arbitre » est déjà une description définie du sens que nous cherchons à isoler. Cependant, il est clair qu'elle ne nous fait pas beaucoup avancer puisque, comme notre investigation prétend contribuer à la discussion qu'on appelle le « problème originel du libre arbitre », cette description définie ne semble pas dire beaucoup plus que ceci : « le sens dans lequel notre discussion entendra le mot 'libre arbitre' est précisément celui dans lequel il est entendu dans notre discussion ». On pourra pourtant utiliser cette première tentative pour faire quelque progrès si l'on parvient à déterminer comment le sens de « libre arbitre » (ou plutôt de « libre ») est introduit à l'origine du problème du libre arbitre. Mais pour cela, il nous faut entrer déjà dans un aperçu des débats qui constituent ce problème philosophique. Je ne suivrai donc pas, à la lettre, le principe qui consiste à donner les définitions complètes de tous les termes avant de présenter les problèmes qui engagent ces termes ; mais j'espère en respecter l'esprit, qui est de donner des définitions logiquement indépendantes de tel ou tel parti pris.

II. Le conflit intuitif originel : solution descriptiviste pour fixer la référence du terme « libre arbitre »

Comment « apparaît » le problème du libre arbitre ? Quelle est son origine ? En répondant à cette question, mon intention ne sera pas de présenter un portrait historique détaillé et absolument fidèle des soubresauts accidentels qui ont accompagné la naissance de ce problème. Une telle tâche historique est un projet à soi seul ; d'autres s'en sont chargé (voir par ex. Adler 1958) ou s'en chargeront. Outre son caractère colossal, ce projet aborderait en outre de nombreuses questions qui n'ont pas d'intérêt direct pour ma recherche : questions d'attribution, d'interprétation, d'influences etc. Mon propos consistera à donner une description (d'ailleurs tout à fait classique) des conditions philosophiques qui font apparaître ce problème, c'est-à-dire des conditions générales qui l'ont fait (bon an mal an) apparaître au cours de l'histoire, mais aussi des conditions qui le font apparaître dans la pensée de tel ou tel individu à n'importe quel moment de l'histoire.

1. Emergence du problème du libre arbitre : le syndrome d'Eleuthère

Voici donc le portrait du « problème du Libre arbitre » tel qu'il surgit dans l'esprit d'un individu idéal, que nous appellerons Eleuthère. Au cours de son existence, Eleuthère considère un jour la possibilité que toutes ses actions soient déterminées par des facteurs qui le dépassent. Cette possibilité peut se présenter à lui de différentes façons et sous l'effet de différents facteurs ; par exemple, ce peut être dans le cadre de réflexions théologiques sur les décisions des Parques, ou au cours de réflexions philosophiques sur le Destin, ou encore au cours de réflexions tirées de l'étude de la nature qui suggèrent un déterminisme (total) par la société ou par les forces atomiques. Quelle que soit l'origine de ces considérations, leur effet est le même : elles plongent Eleuthère dans une forme de perplexité : cette possibilité qui s'ouvre à lui n'est pas en harmonie avec l'image qu'il avait auparavant de lui-même et de ses actions. Par ailleurs, cette possibilité suscite également en lui un trouble concernant la responsabilité morale : cette possibilité, en effet, n'est pas en harmonie avec l'idée très générale que nous sommes moralement responsables, une idée qu'Eleuthère n'aurait autrement jamais pensé à mettre en question, ni même sans doute à formuler. Cette nouvelle

possibilité suscite donc chez Eleuthère une double perplexité, concernant d'une part l'image qu'il avait de lui-même et de ses actions, et d'autre part sa conception de la moralité.

C'est ce syndrome d'Eleuthère, un syndrome de double perplexité, qui définit à mon sens le *problème du libre arbitre*. Et c'est à partir de ce problème que je donnerai une description définie du sens métaphysique de « libre ». En effet, la plupart du temps, Eleuthère trouvera dans sa langue maternelle un terme qui désigne de manière appropriée cette partie de son image de lui-même et de ses actions qui entre en conflit avec ses nouvelles considérations sur un possible déterminisme. C'est ce rôle que joue précisément le terme « libre » dans son usage métaphysique. On peut donc donner une première caractérisation de la manière dont est fixé, en situation, ce sens du terme « libre » : ce sens est fixé par l'image qu'Eleuthère a de lui-même et de ses actions, et sert à désigner une certaine partie ou un certain aspect de cette image. Mais je crois qu'on peut aller plus loin dans la compréhension du rôle linguistique de ce terme, et cela nous permettra de mieux comprendre le lien essentiel entre les deux perplexités d'Eleuthère.

Comment se fait-il qu'Eleuthère, quelle que soit sa langue maternelle, ait de fortes chances de trouver un terme qui désigne précisément cet aspect qui nous intéresse de son image de lui-même et de ses actions ? Il y a après tout une quantité de caractéristiques de notre image de nous-mêmes que nous n'avons pas de raison particulière de désigner par un terme propre ; peut-être ces caractéristiques auront-elles été classifiées et dénommées par certains psychologues patients et méticuleux, mais pour qu'une certaine caractéristique soit dénommée dans la plupart des langues naturelles, et qui plus est dans leur vocabulaire pré-philosophique et pré-scientifique, il faut bien que cette notion joue un rôle important dans toute existence humaine. Une fois que le problème est ainsi posé, il me semble que la réponse s'impose avec évidence : si cette caractéristique de notre image de nous-mêmes se trouve désignée par un terme dans la plupart des langues naturelles, c'est qu'elle joue un rôle fondamental dans un contexte qui est fondamental à toute existence humaine, le contexte de l'imputation morale. C'est donc dans le sous-vocabulaire de l'imputation morale qu'Eleuthère trouvera habituellement le terme convenable pour expliciter sa première perplexité, celle qui tient simplement au rapport entre la possibilité du déterminisme et son image de lui-même. On comprend alors pour quelle raison cette première perplexité amène le plus souvent à une seconde perplexité qui concerne proprement l'imputation morale et sa légitimité dans le scénario déterministe envisagé par Eleuthère.

2. Du problème traditionnel à la description définie du concept de « libre arbitre »

Cette reconstruction du problème du libre arbitre en son origine est, somme toute, très classique. Et pourtant, les métaphysiciens qui discutent de ce problème ne semblent pas en avoir tiré une méthode pour déterminer avec précision le sens des termes fondamentaux du débat. Tous les éléments d'une description définie précise et non triviale y sont pourtant réunis. Que nous dit en effet cette reconstruction à propos du concept de « libre arbitre » ? Premièrement, que c'est un concept qui appartient à l'image qu'Eleuthère a de lui-même pré-philosophiquement (et pré-scientifiquement). Deuxièmement, que ce concept entre en conflit (ou a tendance à entrer en conflit) avec la possibilité envisagée d'un déterminisme total. Troisièmement, que ce concept joue dans l'imputation morale un rôle tel que lorsque le déterminisme entre en conflit avec ce concept, il entre également en conflit avec l'imputation morale. Autrement dit, à tort ou à raison, Eleuthère semble considérer ce concept comme un réquisit de l'imputation morale (légitime). Qu'est-ce donc que le libre arbitre pour Eleuthère ? C'est la faculté qui appartient à son image initiale de lui-même et de ses actions, qui entre en

conflit avec la suggestion d'un déterminisme total et qu'il croit être requise pour l'imputation morale. Nous avons donc à présent une description définie de ce que signifie le mot « libre arbitre » dans le cas d'Eleuthère. Eleuthère n'étant qu'une fiction heuristique, nous allons à présent traduire cette première description définie en termes plus généraux.

Que représente cette fiction heuristique qu'est Eleuthère ? On serait tenté de répondre : Eleuthère représente ici le Sens Commun. Cette réponse, quoique acceptable, n'est pas très satisfaisante parce que l'expression de « Sens Commun » a des acceptions plus ou moins fortes. Chaque acception donne une lecture possible de notre fiction d'Eleuthère, mais il nous faudra bien choisir entre ces différentes lectures. Dans son sens le plus faible, le sens commun est une notion simplement statistique et relative à une société particulière, qui désigne ce que la plupart des hommes de cette société croient être vrai. C'est en ce sens qu'on peut dire que la croyance géocentrique faisait partie du sens commun avant les découvertes de Copernic et Galilée. Dans un second sens, également statistique mais universalisé, le sens commun désigne ce que la plupart des hommes *en général* (de toutes sociétés, de tous continents, et de toutes époques... jusqu'à présent) croient être vrai. En ce sens, l'existence du soleil fait certainement partie du sens commun. Enfin, il existe un sens fort et normatif du terme « sens commun » (ou plutôt une famille de sens normatifs) : le sens commun est alors compris non seulement comme ce que la plupart des hommes croient mais surtout ce qu'ils font bien de croire, soit parce que ne pas croire cela manifesterait une forme de maladie mentale (norme psycho-biologique), soit parce que toute croyance contraire serait auto-réfutante ou inconsistante (norme logique ou épistémologique).

A chacun de ces sens correspondrait une manière différente de déterminer le sens du mot « libre arbitre ». On pourrait en effet choisir de donner de ce terme un sens relatif à telle société ou à telle époque : « libre arbitre » désignerait par exemple une faculté que *les Grecs du V^{ème} siècle* (ou la plupart d'entre eux) croyaient avoir, qui suscitait *en eux* un conflit avec l'idée de déterminisme, et que *ces mêmes Grecs* croyaient être requis pour la responsabilité morale. Et le mot pourrait alors avoir un autre sens à d'autres époques, par exemple à notre époque, en fonction des changements dans la conception commune de l'action humaine et de la responsabilité morale. Peut-être pourrait-on en effet trouver un *autre* concept (un concept propre à notre conception présente de nous-mêmes) qui jouerait un rôle analogue en suscitant également un conflit avec l'idée (que *nous* nous faisons) du déterminisme et serait requis pour (notre concept de) la responsabilité morale. Ou alors, il se pourrait également qu'il n'y ait pour nous aucun concept qui joue ce genre de rôle, auquel cas le problème du libre arbitre ne serait définissable que relativement aux conceptions grecques (et éventuellement à telle ou telle autre époque, par exemple médiévale, mais pas relativement au sens commun de *notre* époque).

Chacune de ces deux hypothèses me semble d'un pessimisme injustifié. D'une part, il est clair que, encore aujourd'hui, l'idée de déterminisme peut susciter chez la plupart d'entre nous une forme de perplexité par rapport à notre représentation de nous-mêmes et de notre responsabilité morale. D'autre part, s'il est vrai que nous trouvons à notre époque et à toute autre des « analogues » de ce problème, pourquoi continuer de supposer qu'il ne s'agit pas du même problème ? La seule raison que je peux imaginer, c'est un argument holiste absolument général qui exclut la possibilité que deux époques aient le moindre concept en commun du fait qu'elles n'ont pas *entièrement* la même vision du monde. Un tel holisme, s'il est conséquent, n'empêche pas seulement que deux sociétés différentes partagent un seul concept, mais elle empêche également que deux *individus* quels qu'ils soient partagent le moindre concept (pourvu du moins qu'ils soient en désaccord sur au moins une thèse, mais cette condition semble trivialement remplie, dans les faits, par n'importe quel ensemble de deux humains, quelle que soit leur proximité de vues). Par conséquent, un tel holisme exclut la possibilité de discussion philosophique, si une telle discussion suppose de parler réellement des mêmes

choses et de s'entendre au moins sur les termes. Quels que soit les mérites d'une telle position, ce n'est pas ici le lieu d'en discuter, et je présumerai donc qu'il est possible, pour deux personnes, d'avoir des concepts en commun. Et partant, je ne vois pas de raison de supposer que le problème du libre arbitre qui préoccupait les Grecs engageait d'autres concepts que celui qui nous préoccupe aujourd'hui. Autrement dit : je ne vois pas de bonne raison de relativiser ma définition du concept de « libre arbitre » au sens commun (statistique) d'une époque ou d'une société particulière.

Eleuthère représente donc, au moins, ce que presque tout homme croit, quelle que soit son époque, quelle que soit sa société. Doit-on aller plus loin et affirmer qu'Eleuthère représente ce que presque tout homme (ou tout homme absolument) *devrait* croire, en un certain sens (psychologique, épistémologique, ...) de « devoir » ? Il est clair que si je place cette contrainte dans la description définie qui me sert à fixer le sens de « libre arbitre », je réponds par avance et par stipulation à une question qui sera au cœur de cette étude : que vaut notre croyance au libre arbitre et peut-elle compter comme connaissance ? Donc le « sens commun » dans son sens le plus fort, et aussi le plus courant en contexte de discussion épistémologique, n'est pas pertinent à ce stade préliminaire de la discussion. Ce qui nous intéresse ici, c'est que tel concept est partagé par presque tous les hommes ; quant à savoir si c'est un bon état de choses ou un mauvais état de choses, nous en discuterons plus tard, abondamment.

Nous avons donc à présent les moyens de supprimer, dans notre première description définie, toute référence à Eleuthère. Le Libre arbitre peut être décrit de manière univoque comme la faculté que presque tout homme pense avoir, qui entre en conflit avec la suggestion d'un déterminisme total, et que presque tout homme pense être requise pour l'imputation morale. Ou pour en donner une version plus formelle :

Libre arbitre =_{df} la faculté *F* telle que :

- presque tout homme pense avoir *F*
- pour presque tout homme, la suggestion d'un déterminisme total génère un conflit d'intuition avec l'idée qu'il a de posséder *F*
- presque tout homme pense que *F* est requise pour l'imputation morale

Cette caractérisation directe du libre arbitre (de la *chose* libre arbitre) ne doit pas faire oublier le cœur de notre démarche, qui consiste en fait à singulariser tout d'abord le *concept* auquel renvoie, dans son sens métaphysique, le mot « libre » :

Concept métaphysique de « libre » =_{df} le concept *C* tel que :

- presque tout homme maîtrise le concept *C* et croit l'instancier
- pour presque tout homme, la suggestion d'un déterminisme total génère un conflit d'intuition avec le concept *C* et l'idée qu'il a de l'instancier
- presque tout homme pense que *C* est impliqué par le concept de « responsabilité morale »

J'aurai quelques remarques fondamentales à faire sur les présupposés de cette terminologie (ceux qu'elle a effectivement, et ceux qu'elle n'a pas), mais puisque nous en sommes à fixer la terminologie du Libre arbitre, réglons rapidement leur sort à quelques autres expressions.

3. Définitions diverses : « Problème psychologique », « problème philosophique », incompatibilisme, compatibilisme, etc.

Bien que nous ayons déjà déterminé un premier sens de l'expression « problème du libre arbitre » à partir de l'histoire d'Eleuthère, il peut être utile de voir un *autre* sens possible de l'expression. Le premier sens que nous avons vu peut être appelé *psychologique* dans la mesure où le « problème » dont il s'agit est une « perplexité », c'est-à-dire un certain état psychologique. Mais il est clair qu'on peut tirer de notre histoire, presque directement, une interprétation *philosophique* de cette perplexité, sous la forme d'un raisonnement qu'Eleuthère pourrait se faire à lui-même sur la base de sa perplexité :

Il semble évident que nous sommes libres et moralement responsables, mais il semble possible que nous soyons déterminés, or le déterminisme semble exclure notre liberté et notre responsabilité morale. L'une de ces thèses doit donc être abandonnée.

Cette formulation interprète les « conflit intuitifs » d'Eleuthère comme des incompatibilités logiques (ou plus précisément des *apparences* d'incompatibilité logique). On peut donc distinguer d'une part :

Le problème psychologique du Libre arbitre =_{df} le double conflit intuitif que provoque la considération d'une possible détermination totale avec :

- l'image que nous avons de nous-mêmes et de nos actions
- l'image que nous avons de la responsabilité morale

et d'autre part :

Le problème philosophique du Libre arbitre =_{df} la question de savoir laquelle (ou lesquelles) des thèses suivantes devrait être considérée comme fausse :

- a. il est bien possible (pour autant que nous le sachions) que nous soyons totalement déterminés
- b. nous sommes libres (au sens métaphysique)
- c. nous sommes moralement responsables
- d. si nous sommes totalement déterminés, nous ne sommes pas libres (au sens métaphysique)
- e. si nous ne sommes pas libres (au sens métaphysique), nous ne sommes pas moralement responsables.

On peut donner de ce problème un aperçu condensé sous la forme du schéma suivant (les lettres **LA** abrègent Libre arbitre ; **RM** est mis pour Responsabilité Morale ; et **Ind** pour Indéterminisme ; les flèches représentent des implications logiques ; le symbole « ✓ » indique que la proposition ou l'implication correspondante est affirmée par la thèse (x) ; enfin le symbole « ?♦* » indique que la proposition **Ind** est réputée douteuse par la thèse (a), ou « menacée » d'être possiblement fausse) :

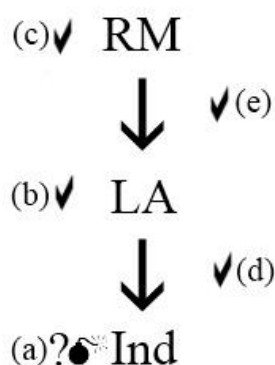


Figure 1 : Le Problème philosophique du Libre arbitre

Est-il vraiment indispensable de refuser l'une de ces cinq thèses ? On pourrait penser qu'il n'y a aucun « problème » (car il n'y a aucune *contradiction logique*) tant qu'on n'a pas *affirmé* le déterminisme total. La simple *possibilité* du déterminisme ne semble pas logiquement incompatible avec les autres thèses (et les autres thèses ne sont pas non plus contradictoires entre elles). Cette objection fait remarquer quelque chose de très juste : que le problème philosophique du Libre arbitre ne repose pas sur une *contradiction logique*. Et elle en tire une conclusion erronée : que dans ces conditions il n'y a pas de problème du tout⁷⁶. Cette conclusion est erronée car il y a bien des cas dans lesquels deux thèses sont incohérentes (ou simultanément inacceptables) sans pour autant être logiquement contradictoires. Un exemple fameux en philosophie est le paradoxe de Moore : je ne peux pas soutenir à la fois que

(1) il pleut

et que

(2) je ne crois pas qu'il pleuve.

Pourtant (1) et (2) ne sont pas logiquement contradictoires. Elles pourraient même très facilement être vraies ensemble. Ce serait seulement un banal cas d'erreur : je crois entendre l'eau tomber, mais c'est un jet d'eau qui vient d'être installé. Ce ne sont donc pas à proprement parler ces deux propositions qui s'excluent l'une l'autre, mais plutôt les deux croyances correspondantes. Prenons un autre exemple, plus proche de notre situation : je ne peux pas soutenir à la fois que

(3) il reste du beurre dans le frigo

et que

(4) (pour autant que je sache) il est bien possible qu'il n'y ait plus rien dans le frigo.

⁷⁶ Ou bien que « dans ces conditions il n'y *aurait pas* de problème du tout » : on peut en effet utiliser cette objection pour montrer que c'est *ma définition* du « problème » qui doit être erronée puisqu'elle arrive à la conclusion qu'il n'y a pas de problème du tout alors qu'il y en a bien un. Le problème du Libre arbitre apparaîtrait par exemple lorsque nous avons des raisons *positives* d'accepter le déterminisme total. Ou bien, indépendamment de la thèse du déterminisme, le problème du Libre arbitre viendrait d'une double incompatibilité du Libre arbitre avec l'indéterminisme et le déterminisme (donc une impossibilité absolue du Libre arbitre), qui serait logiquement contradictoire avec l'existence effective du Libre arbitre et de la responsabilité morale (van Inwagen 2008).

Là encore, (3) et (4) ne sont pas logiquement contradictoires et pourraient fort bien être vraies ensemble. Ce serait un cas banal de manque d'information : c'est la fin de la semaine, il arrive souvent que le frigo soit vide à ce moment de la semaine, et je n'ai pas encore vérifié son contenu, mais en fait, à mon insu, (3) est vraie : il reste effectivement du beurre dans le frigo⁷⁷.

Dans le paradoxe de Moore et dans l'exemple du frigo, l'incompatibilité entre les deux thèses n'est donc pas une incompatibilité logique, mais plutôt une incompatibilité épistémique : il n'est pas possible d'accepter rationnellement l'une et l'autre à cause de ce que veulent dire « je crois que » ou « (pour autant que je sache) il est bien possible que ». C'est ce même genre d'incompatibilité qui est au cœur du problème du Libre arbitre : il ne semble pas rationnel à la fois d'affirmer un ensemble de thèses qui impliquent nécessairement l'indéterminisme et de tenir pour « bien possible (pour autant que je sache) » l'éventualité du déterminisme. J'aurai l'occasion de revenir abondamment sur les questions épistémologiques qui sont ici en jeu. Pour l'instant, je ferai simplement observer que, tel que je le comprends, le problème philosophique traditionnel du Libre arbitre est d'entrée de jeu un problème (en partie) épistémologique.

Il faut donc (pour des raisons de rationalité épistémique) abandonner l'une ou l'autre des thèses présentées dans la définition du problème. Je définirai donc les différentes solutions au problème philosophique en fonction de la proposition (ou des propositions) qu'elle refuse⁷⁸. La première solution consiste à refuser la thèse (a), c'est-à-dire en fait à affirmer l'indéterminisme pour préserver l'ensemble des autres thèses. Cela revient également à répudier comme erronées ou illusoires les différentes suggestions d'un possible déterminisme total. Cette solution reçoit traditionnellement le label suivant :

Solution libertarienne =_{df} la solution au problème philosophique du Libre arbitre qui nie la thèse (a) et préserve les autres thèses.

Toutes les autres solutions auront en commun d'accepter (a). Dès qu'on accepte (a), le premier conflit que l'on rencontre est avec les thèses (b) et (d). En effet, si on affirme à la fois que nous sommes libres (au sens métaphysique) et que cette liberté requiert l'indéterminisme, on ne peut pas continuer à considérer le déterminisme comme une éventualité tout à fait possible. Donc si l'on veut maintenir la possibilité du déterminisme, il faut donc mettre en doute soit la thèse du Libre arbitre **LA** que (b) affirme, soit la thèse de l'incompatibilité entre Libre arbitre et déterminisme, thèse affirmée par (d).

Considérons tout d'abord cette deuxième possibilité, c'est-à-dire le débat concernant la proposition (d). Ce débat a pris une importance prééminente dans le traitement philosophique du Libre arbitre depuis Hobbes jusqu'à une époque très récente (disons jusqu'aux années 1980). Pour cette raison, la terminologie la plus courante aujourd'hui reste fondamentalement structurée par la réponse apportée à ce débat particulier. La thèse (d) a reçu le nom (remarquablement imprécis) de

⁷⁷ Cf. (Engel 2007, 59) : « Dire : 'Je sais que mon auto est parkée dans la rue d'à côté, mais elle pourrait bien avoir été volée il y a cinq minutes' sonne bien comme une phrase à la Moore ».

⁷⁸ On pourrait aussi essayer de répertorier les différentes « solutions », c'est-à-dire en fait les différents « traitements », du problème psychologique : méthode Coué (penser régulièrement à notre liberté et écarter toute occasion de penser à la possibilité du déterminisme), méthode dissolvante (raconter un récit dans lequel toutes ces thèses apparaissent à un moment ou à un autre sans produire aucune perplexité), méthode philosophique (produire dans son esprit la conviction que telle solution philosophique est totalement satisfaisante, de telle sorte que la perplexité disparaisse), ... Une telle classification n'aura pas vraiment d'importance ici, puisqu'il est évident que dans une thèse de philosophie, nous chercherons une solution philosophique, qu'elle soit ou non qualifiée pour ses vertus psychologiques curatives.

Thèse incompatibiliste =_{df} thèse selon laquelle si nous sommes totalement déterminés, alors nous ne sommes pas libres (au sens métaphysique) – i.e. le Libre arbitre est incompatible avec un déterminisme total

tandis que sa négation a pris (fort logiquement) le nom de

Thèse compatibiliste =_{df} thèse selon laquelle le Libre arbitre est compatible avec un déterminisme total.

Il est fondamental de remarquer que, en tant que telles, ces deux thèses ne constituent pas des réponses générales au problème du Libre arbitre. C'est particulièrement évident pour la thèse incompatibiliste : cette thèse (en elle-même) ne résout absolument pas le problème puisqu'elle ne fait qu'*accepter* une des thèses du problème sans indiquer laquelle (ou lesquelles) il faudrait refuser. Il y aura donc différentes solutions incompatibilistes, qui refuseront différentes thèses. Nous venons d'ailleurs d'en voir une : la solution libertarienne. Quant à la thèse compatibiliste, elle permet certes de résoudre le problème – puisqu'elle *refuse* la thèse (d) – mais elle ne donne pas non plus une réponse générale à l'ensemble des questions que soulève le problème puisque, en elle-même, elle reste compatible avec la négation du Libre arbitre par exemple. Il est en effet possible, au moins logiquement, d'affirmer que le Libre arbitre, quoi que compatible avec le déterminisme, est une faculté que nous n'avons pas (pour d'autres raisons que celles tenant au problème traditionnel du Libre arbitre). Ou encore d'affirmer que, quoi que libres, nous ne sommes pas moralement responsables (là encore, pour des raisons indépendantes du problème traditionnel). Je ne donnerai pas des noms à chacune de ces « conceptions générales » ; la seule qui m'intéressera ici est celle qui, s'appuyant sur la thèse compatibiliste, préserve la totalité des autres thèses. Selon cette conception générale, nous sommes libres (b) et moralement responsables (c), et même moralement responsables *parce que* libres (e), mais cela n'exclut pas l'éventualité que nous soyons peut-être (pour autant que nous le sachions) totalement déterminés (a), car cette éventualité n'est nullement incompatible avec l'existence du libre arbitre (thèse compatibiliste, qui refuse (d)). Cet ensemble de thèses constitue une solution générale au problème du Libre arbitre et se situe sur le même plan que la *solution* libertarienne ; je l'appellerai

Solution compatibiliste =_{df} la solution au problème philosophique du Libre arbitre qui nie la thèse (d) et préserve les autres thèses.

Il est donc important de ne pas confondre la *solution* compatibiliste et la *thèse* compatibiliste (la solution implique ou contient la thèse, mais cette dernière est de portée beaucoup plus restreinte).

Considérons à présent les autres solutions possibles. Nous avons dit qu'au lieu de refuser (d), il existait la possibilité de refuser (b), c'est-à-dire la thèse du Libre arbitre **LA**. Mais il n'est pas possible de refuser *seulement* (b). En effet, de la même façon que la mise en doute de **Ind** obligeait à abandonner soit **LA** soit **LA**→**Ind**, de même, abandonner **LA** oblige à réviser encore au moins une autre thèse dans la moitié supérieure de notre figure 1 : soit **RM** soit **RM**→**LA**. Les deux dernières solutions consistent donc à refuser (b) et (c) d'une part, ou à refuser (b) et (e) d'autre part. Refuser (b) et (e) consiste à dire que la possibilité du déterminisme réfute notre croyance initiale au libre arbitre, mais qu'elle ne ruine pas pour autant notre croyance à la responsabilité morale, car cette dernière est compatible avec le déterminisme et l'absence de libre arbitre. Cette thèse qui sépare les deux questions de compatibilité pour accepter une compatibilité mais pas l'autre a été baptisée « semi-compatibilisme » par Fischer qui la défend. Je garderai cette terminologie :

Solution semi-compatibiliste =_{df} la solution au problème philosophique du Libre arbitre qui refuse les thèses (b) et (e) et préserve les autres thèses.

La dernière solution consiste à accepter les deux thèses d'incompatibilité tout en conservant la thèse (a). Il s'ensuit que cette solution doit refuser à la fois (b) et (c). En d'autres termes, cette solution exige de renoncer aux deux « images » de nous-mêmes qui sont à l'origine du problème psychologique du Libre arbitre : l'image que nous avons de nos actions et l'image que nous avons de notre responsabilité morale. Cette solution est classiquement appelée « incompatibilisme dur », par opposition à l'incompatibilisme libertarien dont les conclusions sont tenues pour moins amères⁷⁹.

Solution incompatibiliste dure =_{df} la solution au problème philosophique du Libre arbitre qui refuse les thèses (b) et (c) et préserve les autres thèses.

4. Les présupposés de cette définition

Maintenant que toute la terminologie est en place, je peux revenir à la question importante de savoir si la manière dont j'ai déterminé le concept de libre arbitre tranche par avance certaines des discussions que je viens de présenter. Rappelez-vous en effet que l'intention de la solution « descriptiviste » était d'éviter à la fois la Charybde d'un déni de clarification, et le Scylla d'une définition qui trancherait par stipulation les débats les plus importants. La caractérisation du libre arbitre et du concept métaphysique de « libre » devrait donner quelques éléments de clarification pour aider le lecteur, et ainsi éviter Charybde. Qu'en est-il de Scylla ?

Dans la mesure où ma caractérisation mentionne déjà la moralité et le conflit avec le déterminisme, ne pourrait-on pas me reprocher d'avoir eu recours, en substance, au même procédé qui définit le libre arbitre comme « type de liberté requis par la responsabilité morale » ou « type de liberté incompatible avec le déterminisme » – à ceci près qu'en recourant aux deux à la fois, je combinerais les présupposés de l'une et de l'autre ? C'est-à-dire que je présupposerais à la fois que (e) la responsabilité requiert le libre arbitre, et que (d) le libre arbitre requiert l'indéterminisme, de telle sorte que je trancherais par avance le débat dans le sens d'un double incompatibilisme. Auquel cas, j'exclurais du débat, par stipulation, le compatibiliste et le semi-compatibiliste, en ne gardant que le libertarien et l'incompatibiliste dur. Il n'est pas difficile, je crois, de se convaincre que mes descriptions définies n'ont pas ces présupposés. Comme le point est fondamental pour mon propos, il vaut la peine de s'y arrêter un peu, et de considérer attentivement ce que ces déterminations présupposent effectivement et surtout ce qu'elles ne présupposent pas.

⁷⁹ L'expression « d'incompatibilisme dur » est assez souvent définie dans un contexte où la solution semi-compatibiliste n'est pas explicitement mentionnée. Si bien que, d'après un certain nombre de définitions, le semi-compatibilisme lui-même devrait être tenu pour un incompatibilisme dur. Et en effet, le semi-compatibilisme s'oppose au libertarianisme en acceptant la conséquence « dure » de son (semi-)incompatibilisme *pour la question de la liberté*. Si je choisis de réserver l'expression « d'incompatibilisme dur » pour le double incompatibilisme qui va jusqu'à refuser la responsabilité morale, ce n'est pas pour une raison philosophique profonde (ce n'est pas, par exemple, dans l'idée que « l'essentiel » est préservé quand on a préservé la responsabilité morale, si bien qu'abandonner l'image que nous avons de nos actions ne serait pas vraiment « dur »). C'est seulement pour des raisons de clarté, cette expression étant à la fois classique et très proche dans son usage de la solution que je définis ici. Ces deux dernières solutions sont parfois également appelées « scepticismes » (à l'égard du libre arbitre, ou à l'égard de la responsabilité morale). Compte tenu de l'importance de ce terme dans la discussion épistémologique à venir, il me semble important, dans l'économie de ma thèse, de le conserver pour une classification ultérieure.

J'ai d'abord déterminé le libre arbitre comme une faculté que « presque tout le monde pense avoir ». Personne ne serait tenté, je suppose, de m'accuser d'avoir ainsi présupposé l'existence réelle du libre arbitre. Une telle présupposition serait évidemment désastreuse comme entrée en matière d'une discussion philosophique sur le libre arbitre, mais il est clair que ma caractérisation parle simplement de la *croyance commune* à l'égard de cette thèse. Tout ce que « présuppose » ma caractérisation, c'est que presque tout le monde pense avoir le libre arbitre, mais ce « presque tout le monde » pourrait se tromper. La discussion philosophique pourra chercher à déterminer si la croyance commune en la matière est à suivre ou à abandonner.

Il en va exactement de même pour les deux autres parties de ma caractérisation. Le libre arbitre est déterminé comme une faculté dont l'intuition entre en conflit avec la suggestion d'un possible déterminisme total. Si donc un certain concept est tel qu'il n'a aucune propension à générer un conflit intuitif, dans aucune circonstance, avec la possibilité du déterminisme total, ce ne sera pas, d'après ma caractérisation, le concept de « libre » qui a donné naissance à la discussion métaphysique sur le Libre arbitre. Un auteur qui proposerait un tel concept pour définir son « problème du libre arbitre » aurait simplement changé de sujet par rapport à la discussion traditionnelle. La propension à produire un conflit intuitif avec le déterminisme est donc bien quelque chose que je pose comme *critère d'identité* du concept métaphysique de « libre ». Mais ce « présupposé »-là n'engage pas du tout à l'incompatibilisme ; et il est, de fait, satisfait par les grandes stratégies compatibilistes. Je reviens brièvement sur chacun de ces points. L'idée que le concept de libre arbitre a une propension à entrer en conflit intuitif avec le déterminisme total n'exclut absolument pas que cette propension soit trompeuse ou illusoire. Il y a après tout un grand nombre d'illusions presque universelles dans l'existence humaine, dont les illusions d'optique donnent les exemples les plus frappants. Les lignes de Müller-Lyer ont une propension à produire en nous la représentation que l'une d'entre elles est plus courte. C'est une propension qui concerne presque tout le monde dans des circonstances pré-réflexives. Pourtant, dans des circonstances optimales (après avoir mesuré les lignes, ou avoir tracé un quadrillage régulier en fond d'image), cette propension naturelle peut s'estomper. Bien qu'on puisse arriver à faire s'estomper l'illusion, une bonne théorie de l'illusion se devra de rendre compte de l'existence (et de la fréquence) de l'illusion. De la même façon, on peut très bien accepter que le concept de libre arbitre ait une propension très répandue à produire dans les circonstances communes un conflit intuitif avec le déterminisme, et soutenir néanmoins que cette propension est illusoire et s'estompe lorsque le concept est considéré avec plus d'attention ou de clarté. Mais, après avoir fait s'estomper l'illusion incompatibiliste, une bonne théorie de l'illusion devra aussi rendre compte de l'existence de cette illusion, c'est-à-dire qu'elle devra assumer et rendre compte de la propension de ce concept (dans de « mauvaises » conditions) à produire un conflit avec l'idée de déterminisme. De fait, les grands compatibilistes ont presque tous reconnu l'existence du conflit intuitif, et une composante essentielle de leur compatibilisme consistait à expliquer cette propension à l'illusion incompatibiliste. Voici par exemple ce que Stuart Mill dit de cette illusion :

Beaucoup de gens ne croient pas – et très peu de gens ressentent effectivement – que la causalité n'est rien de plus qu'une consécution invariable, certaine, et inconditionnelle. [...] L'imagination retient toujours, même lorsque la raison le rejette, le sentiment de quelque connexion plus intime, de quelque lien spécial, ou de contrainte mystérieuse exercée par l'antécédent sur le conséquent. Or c'est cela précisément, quand on considère son application à la volonté humaine, qui entre en conflit avec notre conscience, et révolte nos sentiments.

(Mill 1974, 837–838)

Pour Stuart Mill, l'image que nous avons de nos actions a bien une propension à entrer en conflit, chez presque tout le monde, avec l'éventualité d'un déterminisme total. Et cette illusion fréquente est liée à une appréhension incorrecte mais quasi inévitable de ce qu'implique réellement le déterminisme total⁸⁰. Une autre stratégie compatibiliste consiste à reporter le lieu de l'illusion sur le concept de « libre » lui-même, lorsqu'il est appliqué de manière inattentive et abusive. Ces deux types de stratégies compatibilistes, en expliquant l'origine (illusoire) de la propension au conflit entre concept de libre arbitre et concept de déterminisme, prouvent par là-même qu'elles acceptent *l'existence* de cette propension.

Un grand nombre d'auteurs incompatibilistes se sont appuyés sur ce conflit intuitif originaire pour affirmer l'existence d'une *croyance* incompatibiliste qui aurait un statut de position initiale au sens chronologique, voire même un statut de position par défaut au sens normatif. Laura Ekstrom, par exemple, considère que « nous arrivons, presque tous, dans la discussion comme des incompatibilistes préthéoriques » (2002, 310).⁸¹ Mais il faut ici avancer prudemment : un conflit intuitif ne fait pas encore une croyance incompatibiliste naturelle, et une croyance incompatibiliste naturelle ne donne pas nécessairement un statut normatif de « position par défaut ». En tout cas, un certain nombre d'auteurs compatibilistes ont remis en cause ces deux étapes supplémentaires. Nous reviendrons un peu plus loin sur ces deux étapes – sur la question d'une croyance originaire dès ce chapitre (sec. III.1), puis sur la question d'un éventuel statut de position par défaut, ou encore de « charge de la preuve » au chapitre 4 (sec. I). Mais ce qui est important pour l'instant, c'est que les grands auteurs compatibilistes, à ma connaissance, ont toujours reconnu au moins l'existence d'un conflit intuitif originaire. Or la manière dont j'introduis le concept de libre arbitre ne présuppose rien de plus que cela.

Pour les mêmes raisons, ma caractérisation ne présuppose rien concernant le rapport réel entre libre arbitre et responsabilité morale. Elle présuppose seulement l'existence d'une propension commune à lier ces deux notions l'une à l'autre. Là encore, cette propension commune pourrait fort bien être illusoire sans que cela remette en cause ma caractérisation. Je ne m'étendrai pas sur ce deuxième cas qui est exactement parallèle au précédent. D'une manière générale, ma caractérisation du concept métaphysique de « libre » ne présuppose donc aucune thèse sur l'instanciation réelle de ce concept, ni sur son analyse conceptuelle, ni sur ses rapports effectifs avec tel ou tel autre concept. Elle ne fait que « présupposer » certaines propensions communes à lier ou opposer ce concept à d'autres. Par ailleurs, je ne considère pas ces « présupposés » comme des décisions terminologiques qui me seraient propre : je considère (à tort peut-être, mais telle est mon intention) que ces propensions au conflit intuitif sont précisément ce qui permet à différents philosophes, de différentes époques, de reconnaître qu'ils parlent de la même chose : lorsque je lis dans l'autobiographie de Mill que la doctrine du chapitre « On Liberty and Necessity » a été élaborée à la suite de (et pour répondre à) ses « accès d'abattement » où son image de lui-même et de ses actions semblait réfutée par quelque chose que « la science avait prouvé », je sais que ce chapitre va parler du même concept que j'ai à l'esprit, un concept que je ne peux verbalement désigner que par un mot hélas polysémique, le mot « libre ».

⁸⁰ On peut d'ailleurs apprendre dans l'autobiographie de Stuart Mill que cette illusion ne l'a pas épargné lui-même : « dans mes derniers accès d'accablement, la doctrine de ce qu'on appelle la Nécessité Philosophique pesait sur mon existence comme un cauchemar. A ce qu'il me semblait alors, la science avait prouvé que j'étais l'esclave impuissant des circonstances antécédentes : comme si mon caractère avait été formé pour moi par des puissances au-delà de mon contrôle, et qui échappaient à présent à mon pouvoir. Je me suis souvent dit quel soulagement ce serait si je pouvais cesser de croire à la doctrine de la formation du caractère par les circonstances » (Mill 1981, 175).

⁸¹ Voir également (Kane 1999, 217) et les autres auteurs cités ci-dessous sec. III.1.

Un dernier présupposé de ma caractérisation du libre arbitre tient au fait qu'elle repose sur une description définie. En philosophie du langage, il est bien connu qu'une description définie telle que « le roi de France » n'est satisfaite qu'à deux conditions. Premièrement, il doit exister au moins un objet qui satisfait la description : il faut qu'il y ait un roi de France. C'est la condition d'existence. Deuxièmement, il ne doit pas exister *plus* d'un objet qui satisfait la description : s'il y a deux rois de France, la description « le roi de France » n'est pas pertinente non plus. C'est la condition d'unicité.⁸²

Donc, lorsque je définis le concept métaphysique de « libre » comme étant « le concept que presque tout homme croit instancier, qui génère chez presque tout homme un conflit d'intuition avec la notion de déterminisme et que presque tout homme associe à la pratique de l'imputation morale », je m'engage à affirmer qu'il y a un concept effectivement maîtrisé par presque tous les hommes et qui a ces différentes caractéristiques, et qu'il y en a *un seul*. Est-il raisonnable de s'engager d'entrée de jeu à l'une et l'autre de ces thèses ?

Commençons par la condition d'existence. Est-il raisonnable de s'engager à l'existence d'un concept qui joue ces différents rôles ? Rappelons tout d'abord que présupposer l'existence d'un *concept* qui satisfait telle description, ce n'est pas présupposer l'existence d'une *chose* qui instancie ce concept⁸³. Mais peut-être n'y a-t-il même pas un concept que presque tous les hommes maîtriseraient et qui, tout en étant lié à l'activité de l'imputation morale aurait tendance à générer un conflit intuitif avec la possibilité du déterminisme ? On peut distinguer trois cas de figures.

Premièrement, il se pourrait qu'il n'y ait pas en nous de propension à expérimenter un conflit intuitif lorsque nous considérons la possibilité du déterminisme. Si tel était le cas, il n'y aurait évidemment pas lieu de parler du concept *responsable de cette propension*. Est-il indiscutable qu'il y ait une telle propension ? Peu de choses sont « indiscutables », mais ce point, comme nous l'avons dit, fait normalement l'objet d'un accord assez large parmi les philosophes qui traitent du libre arbitre, qu'ils soient compatibilistes ou incompatibilistes. Dans la section suivante, je présenterai des données empiriques récentes qui confortent l'idée d'une propension commune au conflit intuitif. Mais plus fondamentalement, mon propos est de souligner que *si* il n'y avait pas une telle propension, *alors* l'expression « problème du libre arbitre » ne désignerait absolument rien. Pour bien comprendre cette idée, il peut être utile de considérer une situation qui semble en être un contre-exemple.

⁸² Il existe un célèbre débat entre Russelliens et Frégéens sur la question de savoir ce qu'il se passe lorsque l'une ou l'autre de ces conditions n'est pas satisfaite. Les Russelliens soutiennent que lorsque les conditions d'unicité ou d'existence ne sont pas satisfaites, toutes les phrases utilisant la description seront *fausses*. Les Frégéens en revanche soutiennent que les phrases utilisant la description dans ces circonstances ne seront ni vraies ni fausses, parce que l'existence et l'unicité ne font pas partie du contenu *asserté* par la description ou la phrase, mais du contenu *présupposé* – et le mot « présupposé » est alors pris dans son sens linguistique technique, et non plus dans le sens non technique où je l'ai employé pour l'instant. Dans le sens linguistique technique, une présupposition est une proposition qui *doit* être vraie pour qu'une phrase ait une valeur de vérité déterminée, soit vraie soit fausse. Dans le sens non technique, une présupposition est une thèse qui ne fait pas l'objet d'une démonstration propre mais qui est simplement admise comme cadre général d'une réflexion donnée. Dans ce sens non technique, les présupposés d'une discussion peuvent très bien être *assertés* dans des phrases et non pas présupposés au sens linguistique... et une manière de les asserter serait d'asserter une phrase contenant une description définie russellienne ! Autrement dit, quelle que soit la lecture Russellienne ou Frégéenne, des descriptions définies, l'assertion d'une telle phrase à l'entrée d'un raisonnement revient toujours à « présupposer » l'unicité et l'existence, si ce n'est dans le sens linguistique précis, du moins dans le sens non technique qui nous intéresse ici.

⁸³ A ce titre, la caractérisation directe de la *chose* libre arbitre comme étant « la faculté *F* telle que presque tout homme pense avoir *F*, etc. » peut donc prêter à confusion : il est compatible avec mes présupposés qu'il *n'y ait pas* (dans la réalité) une telle faculté *F*, mais seulement le concept illusoire de cette faculté. Je ne considère donc cette caractérisation directe que comme une étape heuristique, ma définition proprement dite étant celle du concept métaphysique de « libre ».

Scénario du « Problème philosophique »

Imaginons que la possibilité du déterminisme n'ait en fait aucune propension naturelle à générer en nous un conflit intuitif. Les différentes discussions philosophiques sur la question seraient simplement liées à un fait de hasard : tel philosophe (Epicure) a considéré la notion de déterminisme un jour où, par hasard, il avait des doutes sur sa liberté à cause d'une météo déprimante ; tel autre (Augustin) s'est trouvé quelque temps plus tard, par hasard, dans la même situation. Et toute la discussion sur la « compatibilité » du déterminisme et du libre arbitre serait issue de ce genre de hasards. Cela n'empêche pas qu'un philosophe, disons le « philosophe subtil », parvienne un jour à défendre (contre toute attente) que la responsabilité morale requiert telle propriété P et que telle propriété P est incompatible avec le déterminisme. Prenons un exemple totalement artificiel pour cette propriété P : la capacité de compter jusqu'à 3.000. Notre philosophe subtil a donc produit des arguments pour montrer qu'un type de liberté à la fois requis pour la responsabilité morale est rendu impossible par le déterminisme.

Dans un tel scénario, devrait-on dire que le philosophe subtil a traité du « problème du libre arbitre » tel que nous l'entendons ? Je soutiens que non : il aurait simplement élaboré un *autre* problème (convaincant ou non) de compatibilité entre déterminisme et responsabilité morale. *Notre* problème était celui qui était censé être à la racine d'un conflit intuitif – et donc si l'on a effectivement prouvé qu'il n'y avait pas un tel conflit, on a du même coup vidé notre problème initial de toute substance, sans qu'il soit ensuite possible de redonner du sens à *ce même* problème. Ce que je viens de dire à propos d'un exemple artificiel s'applique tout autant aux exemples plus réalistes. Je pense en particulier à Peter van Inwagen, qui estime que son incompatibilisme repose exclusivement sur un argument philosophique technique, en laissant ouverte la possibilité que les intuitions incompatibilistes communes reposent sur des confusions.⁸⁴ Dans le même ordre d'idée, certains compatibilistes adoptent une stratégie en deux étapes : premièrement, ils montrent que le conflit intuitif dont part la discussion repose sur une confusion grossière, et qu'une formation philosophique minimale suffit à la dissiper ; puis ils remarquent que les *nouveaux arguments* incompatibilistes (par exemple, celui de van Inwagen) ne reposent *pas*, quant à eux, sur ces confusions grossières, mais qu'à ces arguments subtils, on peut apporter des réponses subtiles.⁸⁵ Si cette stratégie en deux étapes fonctionne, la bonne façon de la décrire est de dire que le « problème du libre arbitre » a été bel et bien résolu, mais qu'un *autre* problème (un problème pour spécialistes dont le commun des mortels n'a pas le début d'une intuition) mérite une discussion plus approfondie. La présente thèse souhaite s'atteler au « problème traditionnel du libre arbitre » ainsi compris, et non pas à un éventuel « problème philosophique » (dont la possibilité me paraît très douteuse) ; que ce problème traditionnel soit résolu ou non, son *existence* suppose qu'il y ait une propension chez la plupart des hommes à éprouver une forme d'inconfort intuitif à la considération de la possibilité du déterminisme.

Deuxièmement, il se pourrait qu'il y ait une propension à éprouver un inconfort à l'idée du déterminisme, mais que cet inconfort n'ait aucune base conceptuelle. Auquel cas, il n'y aurait pas non plus de sens à parler du *concept* (de nous-mêmes et de nos actions) responsable de cette propension. La terminologie établie un peu plus haut permet, je crois, de décrire ce cas de manière satisfaisante. S'il y a un inconfort sans base conceptuelle, alors il y a bien un « problème psychologique » du libre arbitre, mais il n'y a pas de « problème philosophique »

⁸⁴ « I shall not examine arguments that turn on the contention that a belief in the incompatibility of free will and determinism rests on some confusion or other—say, a confusion of "descriptive" with "prescriptive" laws, or a confusion of causation with compulsion. My belief in the incompatibility of free will and determinism rests on the arguments of Chapter III, and these arguments involve no such confusion. » (van Inwagen 1983, 106)

⁸⁵ Cette stratégie en deux étapes est celle, notamment, de Robert Audi (1993).

du libre arbitre. Il n'y a pas de problème philosophique *lié au problème psychologique*, mais il ne peut pas y avoir de problème philosophique du libre arbitre *du tout*, car le problème du libre arbitre est *défini* par référence au problème psychologique. Les mêmes remarques que je viens de faire sur le premier point s'appliquent donc ici également : si l'on parvenait à établir que le problème psychologique n'a aucun aspect conceptuel, alors le problème traditionnel du libre arbitre serait bel et bien dissous, et tout ce que pourrait apporter le philosophe subtil serait un *autre* problème concernant les notions de responsabilité morale et de déterminisme.

Troisièmement, il se pourrait qu'il y ait un conflit intuitif reposant sur un certain concept de nous-mêmes et de nos actions, mais que ce concept n'ait rien à voir avec nos pratiques d'imputation morale. Ce cas est plus délicat. Il est clair, historiquement, que les discussions sur le libre arbitre ont toujours été liées à la question de la responsabilité morale. Mais peut-être cet état de fait repose-t-il simplement sur une *illusion* commune. Par exemple, on pourrait imaginer l'illusion suivante : lorsque, en contexte philosophique, nous prêtons attention à notre concept de nous-mêmes qui entre en conflit intuitif avec l'idée du déterminisme, nous avons une propension à penser également au contexte de l'imputation morale... mais lorsque nous sommes en contexte d'imputation morale, ce concept n'intervient d'aucune manière. Dans cette situation, il me semble que rien ne s'oppose à parler du « problème du libre arbitre » car un certain lien intuitif est tout de même assuré entre la responsabilité morale et le concept en question. En revanche, si un philosophe subtil mettait au jour un concept qui fait partie de notre vision de nous-mêmes et qui *peut* provoquer un conflit intuitif avec l'idée de déterminisme, il me semble qu'il n'aurait pas ainsi déterminé le concept traditionnel de libre arbitre, car ce dernier, d'après ce que nous voyons dans l'histoire de la philosophie, génère typiquement des discussions sur la responsabilité morale. On peut donc supposer qu'il aurait mis en évidence une source de conflit *potentiel* entre notre image de nous-mêmes et le déterminisme, qui n'est pas cependant la source du conflit *réel* qu'est le problème psychologique du libre arbitre.

Pour résumer ces trois premiers points : pour qu'il y ait un « problème (philosophique) du libre arbitre », d'après la caractérisation que je donne, il est nécessaire qu'il y ait un conflit provoqué par la possibilité du déterminisme sur notre image de nous-mêmes, que ce conflit soit de nature conceptuelle, et que le concept qu'il engage ait une propension naturelle à renvoyer (au moins en contexte philosophique) à l'idée de responsabilité morale.

Venons-en à présent à la condition d'unicité : que se passerait-il s'il y avait *plusieurs* concepts qui satisfont toutes ces conditions ? Prenons un exemple inspiré des discussions actuelles. Supposons que notre image de nous-mêmes et de nos actions contienne, entre autres, les deux concepts suivants : le concept « d'être capable d'agir autrement (dans les mêmes circonstances) » et le concept « d'être la source ultime de mes actions ».⁸⁶ Supposons en outre que chacun de ces concepts ait une propension à générer une perplexité intuitive lorsqu'on le confronte à la possibilité du déterminisme. Supposons enfin que chacun de ces concepts soit typiquement utilisé dans les concepts d'imputation morale, par exemple dans deux types d'excuses différentes (l'une consistant à montrer qu'on ne « pouvait pas faire autrement », l'autre consistant à désigner la véritable source ultime de ce qu'on a fait). Pourrait-on alors parler *du* concept lié à la responsabilité et capable de susciter un conflit avec l'idée de déterminisme ? D'un point de vue linguistique, ce serait illégitime puisqu'il y a *deux* concepts qui satisfont ces conditions. Mais le problème ici n'est guère plus que linguistique. En effet, les concepts ont la particularité de pouvoir s'agglomérer. Ainsi, à partir du concept « mammifère » et du concept « herbivore », je peux composer le concept conjonctif « mammifère herbivore ». De la même façon, à partir de mes deux concepts liés à

⁸⁶ Nous retrouverons ces deux concepts au chapitre 4 (sec. IV.2.c).

l'imputation morale et au conflit intuitif avec le déterminisme, je peux former un (unique) concept conjonctif. Et cette procédure peut être répétée quel que soit le nombre de (petits) concepts qui satisfont chacune des conditions de ma définition : il sera toujours possible de composer un unique concept *maximal*. Pour être plus exacte, ma définition devrait donc préciser que « le concept métaphysique de « libre » est le concept *maximal* C tel que etc. ». Ou pour éviter cet outil de vocabulaire, on pourrait choisir de définir « le problème du libre arbitre » à partir du concept *ou de l'ensemble de concepts* qui satisfont les conditions pertinentes. Toutes ces précisions m'ont semblé compliquer inutilement l'expression. Je ne les utiliserai donc pas (sauf exceptions) et je considérerai que la condition d'unicité n'impose aucun présupposé philosophique particulier.

En résumé de cette section, disons que la meilleure façon de définir le concept de « libre arbitre » est de partir de notre vision commune de nous-mêmes, de nos actions et de leur caractère moral, et de la propension de cette vision commune à entrer en conflit intuitif avec le déterminisme. Cette définition permet de fixer la référence de l'expression « libre arbitre » d'une manière qui correspond à l'usage dans la discussion philosophique, et *sans préjuger* des questions philosophiques qu'on veut laisser ouvertes, notamment la question d'existence et la question de compatibilité. En effet, cette définition ne présuppose en rien que nous avons *effectivement* le libre arbitre, ni que le libre arbitre entre *effectivement* en conflit conceptuel avec le déterminisme.

En revanche, cette définition a bel et bien un présupposé important : à savoir l'idée qu'il y a en nous une propension naturelle à expérimenter un conflit intuitif lorsque nous envisageons la possibilité du déterminisme. Ce présupposé ne me semble pas problématique non seulement parce qu'il a de bonnes chances d'être satisfait, mais surtout parce qu'il me semble que ce conflit intuitif est réellement la condition qui permet à deux auteurs de s'assurer qu'ils parlent bien d'une seule et même chose lorsqu'ils discutent « du libre arbitre » : s'il n'y avait pas de concept commun de nous-mêmes générant un tel conflit intuitif, alors il me semble qu'il faudrait en conclure que tout le débat sur « le libre arbitre » repose fondamentalement sur des équivoques. Mais passons à présent aux éléments qui permettent d'affirmer qu'il y a bel et bien un conflit intuitif commun.

III. L'existence du conflit intuitif, et l'expérience du libre arbitre

1. L'existence du conflit intuitif naturel

La plupart des auteurs incompatibilistes commencent leur présentation du problème du libre arbitre en affirmant que notre conception spontanée ou naturelle du libre arbitre non seulement « génère un conflit » avec le déterminisme, mais est même tout bonnement incompatibiliste. Par exemple, Robert Kane écrit ceci :

D'après mon expérience, la plupart des personnes ordinaires sont au départ des incompatibilistes naturels. Elles croient qu'il y a un certain type de conflit entre liberté et déterminisme ; et l'idée que la liberté et la responsabilité pourraient être compatibles leur semble tout d'abord n'être qu'une « dérobade lamentable » (*quagmire of evasion*) (William James) ou un « misérable subterfuge » (Emmanuel Kant). C'est seulement par les arguments subtils des philosophes qu'on peut faire sortir les personnes ordinaires de cet incompatibilisme naturel. (Kane 1999, 217)

On trouvera aisément le même genre de déclarations chez la plupart des auteurs libertariens.⁸⁷ On en trouvera également chez les auteurs indéterministes durs, notamment chez Galen Strawson, un des premiers auteurs à s'être intéressé de près au rôle et au contenu de l'expérience commune du libre arbitre :

[La conception incompatibiliste] est exactement le type de liberté que la plupart des personnes croient posséder de manière ordinaire et non-réflexive. (Strawson 1986, 30). Il fait partie de notre nature de considérer que le déterminisme pose un sérieux problème à nos notions de responsabilité et de liberté. (Strawson 1986, 76).⁸⁸

Jusqu'à une époque récente, les déclarations de ce genre reposaient essentiellement sur l'expérience personnelle de l'auteur – soit auprès de ses étudiants, soit auprès de ses amis non-philosophes. Si un Robert Kane, par exemple, a toujours rencontré chez tous ses étudiants débutants et tous ses amis non philosophes une grande difficulté à prendre au sérieux le compatibilisme, au moins initialement, il était sans doute justifié à extrapoler son expérience et à supposer que les autres professeurs rencontraient le même genre de réactions.⁸⁹ Mais quelle que soit la valeur de cette expérience informelle, elle a fini par attirer le scepticisme de certains auteurs compatibilistes, qui ont fait observer trois biais possibles : d'une part, il est possible qu'un auteur incompatibiliste *présente* le problème à ses étudiants d'une manière qui incite à être incompatibiliste (quand bien même il ferait des efforts sincères pour donner une présentation neutre) ; deuxièmement, il se pourrait qu'il *surinterprète* les réactions de ses étudiants dans le sens de sa thèse ; troisièmement, il se pourrait que ses étudiants soient effectivement incompatibilistes, mais que cela manifeste seulement un *biais sociologique* de tel échantillon de population, et non pas un quelconque jugement naturel. Ce genre de précautions – que les auteurs incompatibilistes trouvent généralement indûment scrupuleuses compte tenu de l'expérience pédagogique largement partagée – ont eu le grand mérite de susciter depuis les années 2000 une série d'enquêtes expérimentales pour tester l'existence et la naturalité des intuitions incompatibilistes chez les gens ordinaires.

Les études de « philosophie expérimentale » (X-Phi) sur le libre arbitre dans les dix dernières années ont essentiellement été le fruit de deux équipes de recherche, l'une constituée d'Eddy Nahmias et divers associés, l'autre de Shaun Nichols, Joshua Knobe, et divers associés. Bien que cette discussion soit encore très vive et largement ouverte, on peut déjà en tirer quelques enseignements particulièrement importants. Le plus simple est de présenter l'histoire de cette discussion en trois grandes étapes.

Les recherches de X-Phi sur le libre arbitre ont été lancées, à l'origine, par l'équipe d'Eddy Nahmias, qui avait pour objectif principal de remettre en question l'idée reçue d'un incompatibilisme spontané. Les résultats de ces premières recherches se sont avérés surprenants ; en effet, ils semblaient montrer qu'en réalité les intuitions communes étaient compatibilistes et non pas incompatibilistes. La première étude importante, (Nahmias et al. 2004), demande aux sujets interrogés de donner une interprétation de l'expression « j'aurais pu faire autrement » (parmi plusieurs interprétations proposées) et semble indiquer que, pour une majorité de sujets, l'interprétation privilégiée est celle qui correspond à la conception

⁸⁷ (Nahmias et al. 2006, 29) donnent une excellente revue de ce genre de déclarations. Ils citent notamment (Campbell 1951, 451), (O'Connor 2000, 4), ou encore (Ekstrom 2002, 310) : « we come to the table, nearly all of us, as pretheoretic incompatibilists ».

⁸⁸ Pour le même genre de déclarations chez d'autres auteurs incompatibilistes durs, voir notamment (Pereboom 2001, xvi) et (Smilansky 2003, 259).

⁸⁹ On peut d'ailleurs imaginer que Robert Kane s'appuyait également sur la confirmation du témoignage de ses collègues

compatibiliste des possibilités alternatives.⁹⁰ Je me concentrerai ici sur la deuxième étude importante de Nahmias *et al.*, à savoir (Nahmias et al. 2005). Dans cette étude, Nahmias *et al.* commencent par décrire aux sujets interrogés un monde déterministe (en prenant soin de le décrire d'une manière qui n'entraîne aucune confusion avec une forme de fatalisme ou de manipulation). Puis ils situent à l'intérieur de ce monde déterministe un agent qui commet une certaine action. Les sujets sont ensuite invités à dire si l'agent en question avait, dans ce monde déterministe, la liberté et la responsabilité morale pour son acte. Le scénario d'origine utilisé par Nahmias *et al.* était le suivant :

Scénario : Imaginez qu'au siècle prochain nous découvrons toutes les lois de la nature, et que nous construisions un superordinateur capable de déduire, à partir de ces lois de la nature et de l'état actuel des choses dans l'univers, ce qui va se produire exactement dans le monde à n'importe quel moment futur. Il peut regarder tout ce qui se passe dans l'univers, et prédire tout ce qui lui arrivera avec une précision de 100%.

Supposez qu'un tel superordinateur existe, et qu'il regarde l'état de l'univers à un certain moment, le 25 mars 2150 ap. J.C., 20 ans avant la naissance de Jeremy Hall. L'ordinateur déduit ensuite à partir de cette information et des lois de la nature que Jeremy va dévaliser la Fidelity Bank à 18h00 le 26 janvier 2195. Comme toujours, la prédiction du superordinateur est exacte ; Jeremy dévalise la Fidelity Bank à 18h00 le 26 janvier 2195. (Nahmias et al. 2005, 566)

Les sujets devaient ensuite répondre aux deux questions suivantes :

FW : Croyez-vous que, en dévalisant la banque, Jeremy a agi « *of his own free will* » (de son plein gré / « par son libre arbitre ») ?

MR : Croyez-vous que Jeremy est moralement blâmable pour avoir dévalisé la banque ?

A chacune de ces questions, les sujets interrogés ont répondu positivement de manière très significative, en dépit du scénario déterministe : 76% de oui pour la question FW, 83% de oui pour la question MR (Nahmias et al. 2005, 566–568). La conclusion de cette étude révolutionnaire semblait sans appel : les intuitions communes sont en réalité compatibilistes ; la majorité d'entre nous peut concevoir le libre arbitre et la responsabilité même dans un scénario déterministe.

Pourtant, une deuxième étape de la discussion a renversé la conclusion au moyen d'études empiriques plus approfondies. Dans leur étude (Nichols and Knobe 2007), Shaun Nichols et Joshua Knobe ont enrichi le dispositif expérimental dans deux directions ; ils ont bien sûr retrouvé le résultat de Nahmias *et al.*, mais ce résultat s'est trouvé placé dans un

⁹⁰ Les descriptions proposées pour l'expression « j'aurais pu faire autrement » étaient les suivantes :
« A. 'I could have chosen to do otherwise even if everything at the moment of choice had been exactly the same'.
B. 'I could have chosen to do otherwise only if something had been different (for instance, different considerations had come to mind as I deliberated or I had experienced different desires at the time)' ». (Nahmias et al. 2004, 174). L'interprétation B est ensuite présentée par Nahmias *et al.* comme « l'interprétation compatibiliste », et le résultat donne une approbation de cette interprétation « compatibiliste » à 62%. On peut contester l'interprétation que Nahmias *et al.* donnent de leur résultat de deux manières : d'une part, il est possible que les sujets considèrent les deux scénarios comme *possibles* et le scénario B (choix différent avec considérations différentes) comme seulement *plus courant* ou plus probable. Deuxièmement, il n'est pas évident que le scénario B soit réellement compatibiliste dans l'esprit des sujets, dans la mesure où les considérations elles-mêmes, qui mènent au choix, peuvent tout à fait être conçues comme étant (au moins dans une certaine mesure) au pouvoir du sujet. Dans ce cas, les gens pourraient fort bien n'avoir pas voulu nier au sujet un véritable contrôle des alternatives, mais seulement lui attribuer un contrôle *indirect* (antérieur) des alternatives.

contexte général qui permet de conclure que les sujets ordinaires ont, en fait, des intuitions incompatibilistes.

La première addition apportée au dispositif de Nahmias *et al.* a été d'ajouter une première question avant la question de libre arbitre ou la question de responsabilité morale, une question qu'on pourrait appeler « question sur le monde actuel » ou « question d'actualité ». Après avoir décrit l'univers déterministe dans des termes semblables à ceux de Nahmias *et al.* (l'Univers A), Nichols et Knobe décrivent un deuxième univers, qui contient certains indéterminismes, situés au moment des décisions humaines (l'Univers B). Puis ils posent aux sujets la question suivante :

QA : Lequel de ces deux univers est, d'après vous, le plus semblable au nôtre ?
(Nichols and Knobe 2007, 669)

A cette question, presque tous les sujets (plus de 90%) ont répondu que notre univers ressemblait davantage à l'Univers indéterministe B. Ce résultat a une conséquence considérable pour l'interprétation de l'expérience de Nahmias *et al.* Il montre en effet que les actions libres telles que se les gens se les représentent *effectivement*, ou dans le monde actuel, supposent des indéterminismes. Et donc, le résultat de Nahmias *et al.* nous informe seulement sur la manière dont varient nos intuitions lorsque nous considérons un monde différent du nôtre : si nous vivions finalement dans un monde déterministe, nous garderions sans doute l'attitude d'attribuer libre arbitre et responsabilité morale. Est-ce là une intuition compatibiliste ? Non, et la meilleure façon de le voir est de constater qu'un auteur aussi incompatibiliste que Peter van Inwagen partage cette intuition : ainsi que nous le verrons au chapitre 3, van Inwagen (1983, 219) soutient que si on lui prouvait scientifiquement l'existence du déterminisme, alors il n'en conclurait pas que nous ne sommes pas libres : il en conclurait plutôt que ses arguments incompatibilistes doivent, finalement, contenir une erreur quelque part. Sa croyance au libre arbitre (et à la responsabilité morale) est *plus forte* que sa croyance incompatibiliste, et l'emporterait donc dans une telle éventualité. Mais en l'état actuel des choses, van Inwagen n'en est pas moins incompatibiliste : il croit que nous sommes libres d'une manière qui introduit des indéterminismes dans la nature. Et le résultat de la question d'actualité QA semble montrer de manière très convaincante que les personnes ordinaires ont le même genre d'attitude : si on les place dans un scénario déterministe, elles continuent à attribuer libre arbitre et responsabilité, mais (pour elles) ce scénario est purement contrefactuel, car dans le monde actuel les actions libres engagent certains indéterminismes. Pourquoi préserverions-nous ces attitudes même dans un scénario déterministe ? Le deuxième apport de Nichols et Knobe a permis d'apporter un élément de réponse à cette question.

Pour enrichir encore le dispositif expérimental de Nahmias *et al.*, Nichols et Knobe ont fait observer que les expériences menées jusque là concernaient seulement des cas *concrets* (comme Jeremy dévalisant une banque) susceptibles de produire en nous des affects, et de biaiser nos jugements conceptuels sur la compatibilité du déterminisme et du libre arbitre. Nichols et Knobe ont donc enrichi le dispositif d'une question *abstraite* sur l'univers déterministe A :

QAbstraite : Dans l'univers A, est-il possible pour une personne d'être pleinement responsable de ses actions ?
(Nichols and Knobe 2007, 670)

Or à cette question abstraite, une très large majorité de sujets (86%) a donné la réponse incompatibiliste.⁹¹ Nichols et Knobe ont également posé une question concrète (comme Nahmias *et al.*) – le cas de Bill qui bat sa femme et ses enfants jusqu’à la mort – et ils ont retrouvé le résultat plutôt compatibiliste dans ce cas concret. Mais ce résultat semble pouvoir s’interpréter comme l’effet des émotions engagées par le cas concret : si l’on veut découvrir la vraie intuition conceptuelle, il faut mettre de côté les émotions, ce que fait la question abstraite. C’est ce que Nichols et Knobe appellent le « modèle de l’erreur de performance affective » (Nichols and Knobe 2007, 671–672) : lorsque nous appliquons nos concepts de libre arbitre ou de responsabilité sans affect, notre performance conceptuelle est bonne (nous appliquons bien ce concept) ; en revanche, lorsque nos émotions sont éveillées par la description d’un acte horrible, ces émotions biaisent l’application de nos concepts et nous poussent à attribuer la responsabilité même si une bonne performance conceptuelle requerrait de retenir cette attribution. La grande force de l’analyse de Nichols et Knobe, c’est qu’elle ne nie pas totalement l’existence des intuitions compatibilistes de sens commun : pour eux, le sens commun permet à la fois de délivrer certaines intuitions compatibilistes et certaines incompatibilistes. Mais lorsqu’on regarde les causes de cette divergence, tout laisse à penser que les secondes représentent plus fidèlement notre compétence conceptuelle en tant que telle.

Pour interpréter cette divergence entre cas concret et cas abstrait, une autre interprétation a été proposée, le « modèle de la compétence concrète » (Nichols and Knobe 2007, 673–674). L’idée générale est la suivante : lorsqu’on interroge les gens sur un cas concret, on les invite à *appliquer* leur concept, par exemple leur concept de libre arbitre, et on obtient ainsi des *jugements* appliquant ce concept. En revanche, lorsqu’on pose la question abstraite, on invite les gens à *réfléchir* à propos de ce concept, et ce qu’on obtient alors, c’est la *théorie* qu’ont les gens à propos de ce concept. Or les gens pourraient fort bien se faire une fausse théorie de leurs propres concepts ; donc la manière la plus sûre et fiable de tester le compatibilisme ou l’incompatibilisme du concept de sens commun, c’est de proposer aux sujets des cas concrets et de leur demander leurs jugements, pas de les inviter à donner leur théorie. Ce « modèle de la compétence concrète » permet de poser une distinction très importante pour mon propos : la distinction entre « concept de Sens Commun du libre arbitre » et « conception de Sens Commun du libre arbitre ». Je reprends ici la terminologie de Dana Nelkin :

Il sera utile de [...] distinguer entre le *concept* de liberté et les *conceptions* particulières de la liberté. Premièrement, il y a un concept de liberté que tout le monde (ou en tout cas beaucoup de gens) possèdent, même s’ils sont en désaccord sur les conditions nécessaires et suffisantes de son instanciation. Il doit y avoir un tel concept si l’on veut pouvoir dire que les compatibilistes, incompatibilistes, et ceux qui croient la liberté impossible, sont en désaccord. Bien sûr, il est possible de défendre que tout le problème du débat sur le libre arbitre est que les participants parlent en fait différents langages [...] Mais je crois que cette idée fait trop peu de crédit aux participants de la discussion. [...] Deuxièmement, il y a aussi les différentes analyses de la liberté proposées [...] Ce sont des *conceptions* de la liberté, des tentatives pour formuler les conditions dans lesquelles les actions sont libres.
(Nelkin 2004, 113–114)

⁹¹ Dans cette expérience, Nichols et Knobe n’ont pas posé de question abstraite à propos du libre arbitre proprement dit. Mais des études postérieures ont retrouvé le même résultat incompatibiliste en posant une telle question ; en particulier, (Roskies and Nichols 2008, 375–376) ont observé une très forte approbation à l’affirmation suivante : « it is impossible for people in Universe A to make truly free choices ».

Le « concept [...] que tout le monde possède », dont parle Nelkin, est clairement le concept de Sens Commun que j'essaie de définir de manière descriptive dans ce chapitre, c'est-à-dire le référent commun (quoique non analysé) des différentes discussions sur le libre arbitre. Quant aux *conceptions* du libre arbitre, Nelkin pense ici principalement aux conceptions *philosophiques* ; mais comme on vient de le voir, il se pourrait aussi que les gens ordinaires aient une tendance naturelle à se former une certaine conception du libre arbitre – une conception qui pourrait être systématiquement erronée, c'est-à-dire infidèle à la définition véritable du concept qu'ils tentent d'appliquer. Par conséquent, lorsqu'on se demande si le Sens Commun est compatibiliste ou incompatibiliste, on peut vouloir poser deux questions différentes :

(Question du concept de Sens Commun) : le concept commun de « libre arbitre » est-il logiquement compatible avec le concept de déterminisme ?

(Question de la conception de Sens Commun) : selon la conception que nous nous faisons naturellement du « libre arbitre », celui-ci est-il compatible avec le déterminisme ?

Il se pourrait très bien que « le Sens Commun » soit incompatibiliste dans le sens de la question 2, tout en ayant un concept compatibiliste dans le sens de la question 1. J'ai déjà évoqué ce genre de possibilité lorsque j'ai montré que la définition descriptive du libre arbitre par le conflit intuitif ne préjugait pas de la question de compatibilité : si elle n'en préjuge pas, c'est justement parce qu'une conception incompatibiliste (un conflit intuitif) peut reposer sur un concept qui, en réalité, est compatibiliste.

La conclusion à tirer pour ce chapitre, c'est que le « modèle de la compétence concrète » ne remet pas en cause l'existence du conflit intuitif que nous recherchons dans ce chapitre. Ce que nous recherchons ici, ce n'est pas une preuve que le concept de Sens Commun est *effectivement* incompatibiliste, mais seulement qu'il a une propension à générer un conflit intuitif. Cette propension pourrait être liée à une incompatibilité réelle entre les concepts de libre arbitre et de déterminisme (ainsi que le défendent Nichols et Knobe) ou bien à une erreur systématique de théorisation sur ce concept (ainsi que le défend le modèle de la compétence concrète). Nichols et Knobe ont apporté des réponses expérimentales solides pour rejeter le modèle de la compétence concrète et sa conclusion compatibiliste.⁹² Mais je n'envisagerai cette question (du compatibilisme ou non du *concept*) qu'au chapitre 4.⁹³ Ce qui nous intéresse ici, c'est que Nichols et Knobe ont établi de manière incontestable (et incontestée) les deux résultats suivants :

⁹² La stratégie employée dans (Nichols and Knobe 2007, sec. 6 sqq) a consisté à distinguer le caractère *concret* du scénario et sa dimension *affective*. La question sous-jacente était la suivante : la différence de résultat entre question concrète et question abstraite est-elle due au potentiel *affectif* de la question concrète (comme le prédit le modèle de l'erreur de performance affective) ou bien spécifiquement au caractère *concret* de la question (comme le prédit le modèle de la compétence concrète). Pour répondre à cette question, Nichols et Knobe ont posé deux questions concrètes, l'une à fort potentiel affectif (*Bill stalks and rapes a stranger*), l'autre sans potentiel affectif (*Mark arranges to cheat on his taxes*). L'expérimentation a révélé qu'à la question concrète sans potentiel affectif, les sujets avaient une très forte tendance à répondre comme à la question abstraite (pas de responsabilité morale), et que c'est seulement la question concrète à *fort potentiel affectif* qui faisait apparaître le résultat compatibiliste. Autrement dit, l'intuition compatibiliste ne vient pas du caractère concret ou non de la question envisagée, mais bel et bien d'un investissement affectif, comme le prédit le modèle de l'erreur de performance affective.

⁹³ Dans le chapitre 4, je ne m'appuierai pas sur la (très bonne) discussion de (Nichols and Knobe 2007), qui défendent l'incompatibilisme contre les objections de Nahmias *et al.* Je m'appuierai plutôt (voir p.342 sqq) sur une autre étude plus récente, (Deery, Bedke, and Nichols 2013), qui soutient l'incompatibilisme d'une manière nouvelle, qui me semble plus directe et plus simple que la discussion entre Nichols *et al.* et Nahmias *et al.*

1. la conception que nous nous faisons spontanément de nos actions est qu'elle introduit des indéterminismes dans l'univers (c'est l'apport de la « question d'actualité »)

2. nous avons une tendance naturelle à concevoir (au niveau abstrait) le libre arbitre comme incompatible avec le déterminisme (c'est l'apport de la « question abstraite »)

Aucune des discussions qui ont suivi dans la littérature de X-Phi n'ont permis de mettre en cause significativement ces deux résultats.⁹⁴ Et quelle que soit l'importance qu'on accorde aux intuitions « compatibilistes » exprimées sur les cas concrets, elles n'altèrent en rien ces deux faits. Or ces deux faits sont suffisants pour établir qu'il y a, bel et bien, une tendance naturelle à expérimenter un certain conflit intuitif entre l'image d'un monde déterministe et l'image que nous nous faisons naturellement de nous-mêmes et de nos actions libres.

Pour être plus exact, il a fallu attendre la troisième phase d'études, et en particulier (Sarkissian et al. 2010), pour arriver véritablement à la conclusion d'un conflit *naturel* et non pas seulement *culturel*. En effet, les premières études ont toutes été menées sur des sujets américains. Leur résultat aurait donc pu manifester un biais de sélection sociologique. Sarkissian *et al.* ont donc réalisé une étude plus large, à partir de quatre échantillons d'étudiants : 66 étudiants des Etats-Unis, 55 étudiants indiens, 40 de Hong Kong, et 70 de Colombie. A ces étudiants, ils ont posé la question d'actualité et la question abstraite de compatibilité (donc les deux questions qui délivraient un résultat incompatibiliste dans l'étude de Nichols et Knobe). Le résultat est très clair : dans chacun des échantillons, on retrouve le résultat très majoritairement incompatibiliste, aussi bien pour la question d'actualité (notre univers ressemble plus à l'univers B) et pour la question abstraite (dans l'univers A, il n'est pas possible pour une personne d'être pleinement responsable moralement de ses actions). Il y semble donc bien y avoir un conflit intuitif *universel* (inter-culturel) entre l'image d'un univers déterministe et notre conception de nous-mêmes et de nos actions.

2. Du conflit intuitif à l'expérience du libre arbitre

Reste à savoir d'où vient ce conflit intuitif naturel que révèle la X-Phi. D'où vient que l'univers indéterministe B nous paraît, naturellement, plus semblable au nôtre que l'univers déterministe A ? Une réponse séduisante est la suivante : ce qui fait que nous jugeons spontanément être dans un univers tel que B – un univers où il y a certains indéterminismes situés aux endroits des actions humaines – c'est *l'expérience* que nous faisons de nos propres actions au moment où nous agissons : au moment où je choisis de prendre une glace à la vanille plutôt qu'à la fraise, j'ai une certaine expérience de mon action, et c'est sur la base de cette expérience que je juge ensuite (à tort ou à raison) qu'un univers tel que B doit être le cas. Dans la première section de ce chapitre, j'ai défini le libre arbitre à partir de « l'image que nous nous faisons de nous-mêmes et de nos actions » ; la question est à présent de savoir quel est le lien entre cette « image » et « l'expérience que nous faisons de nous-mêmes et de nos

⁹⁴ Les objections qui ont été faites contre le résultat de Nichols et Knobe – outre l'objection « jugement » vs « théorie » qu'on vient de voir – avaient toutes à voir avec la description des univers ou la formulation des scénarios. La description de l'univers déterministe aurait employé des termes qui portaient à une confusion avec le fatalisme ; ou la question de responsabilité morale était posée en des termes trop forts. Le principal problème de ce genre d'objections, c'est que Nichols et Knobe ont utilisé les *mêmes* formulations pour la question concrète et pour la question abstraite, et que dans le cas concret, ces formulations ne semblent pas avoir eu d'effet de biais puisque le résultat « compatibiliste » de Nahmias *et al.* a pu être retrouvé. Donc le résultat incompatibiliste du cas abstrait (et surtout l'écart entre les deux) ne peut pas être attribué à un problème de formulation. Pour répondre à ces problèmes de formulations, plusieurs études ont par ailleurs été effectuées avec des formulations non controversées, et les mêmes résultats ont été trouvés. Pour un résumé de ces discussions et études, voir notamment (Sarkissian et al. 2010, 349).

actions » au moment où nous agissons. Dans le modèle épistémologique que je défendrai finalement, l'expérience agentive (commune) jouera un rôle fondamental ; mais pour bien comprendre ce rôle, il me faudra apporter quelques précisions conceptuelles sur la notion d'expérience. Tout d'abord, il faudra distinguer entre un usage subjectif (*experience as of X*) et un usage objectif du terme « expérience » (*experience of X*). Ensuite, j'introduirai la distinction entre « expérience phénoménale » et « expérience doxastique » (ou « impression que *p* »). Ces deux distinctions serviront à établir le point suivant : le conflit intuitif ne repose peut-être pas sur l'expérience que nous faisons de nous-mêmes au moment où nous agissons (nous aborderons cette question au chapitre 4, sec. IV.1.), mais elle repose certainement sur une certaine sorte « d'expérience » dans un sens plus faible. Et c'est cette « expérience » (dans le sens faible) qui sera au cœur de la discussion dans toute la première partie de notre étude.

La distinction entre sens objectif et sens subjectif de « expérience » est fondamentale lorsque l'on parle de « l'expérience de l'action » (ou expérience agentive). Cette expression servira souvent à désigner l'expérience que nous faisons *lorsque nous sommes en train d'agir* – ou « ce que cela fait que d'être en train d'agir », ou « à quoi cela ressemble que d'être en train d'agir ». Mais la même expression pourra aussi servir à désigner l'expérience *qui nous représente en train d'agir*. Bien sûr, si mon système cognitif est fiable, alors l'expérience que je ferai au moment où j'agis sera une expérience qui me représente en train d'agir. De même, l'expérience que je ferai au moment où je mange une glace à la vanille sera sans doute une expérience qui me représente en train de manger une glace à la vanille. Mais il existera aussi des cas où les deux usages de l'expression différeront. C'est le cas des illusions. Par exemple, certains vols paraboliques permettent de donner une illusion d'apesanteur. Cette illusion pourra être décrite comme une « expérience de vol parabolique » si on prend l'expression au sens objectif ; mais elle pourra aussi être décrite comme une « expérience d'apesanteur » si on la prend au sens subjectif. La différence fondamentale entre ces deux usages, c'est que le sens objectif est *factif*, tandis que le sens subjectif ne l'est pas : personne ne peut avoir une « expérience du vol parabolique » (au sens objectif) en l'absence de vol parabolique ; mais on pourrait très bien avoir une « expérience d'éléphant rose » (au sens subjectif) même s'il n'a jamais existé d'éléphants roses.

Y a-t-il vraiment là deux sens différents du mot « expérience » ? Sans doute pas : dans l'un et l'autre sens, l'expérience reste un « vécu ». Ce qui diffère, c'est plutôt la nature du complément. On pourrait dire que dans l'usage objectif, on donne au mot « expérience » un complément de circonstance : l'expérience du vol parabolique, c'est le vécu d'un sujet qui se trouve dans les circonstances (effectives) d'un vol parabolique. Tandis que dans l'usage subjectif, on donne au mot « expérience » un complément de contenu : l'expérience de l'apesanteur, c'est un vécu dont le *contenu* est d'être en apesanteur. Puisqu'il y a deux compléments différents, on pourrait même imaginer des expressions qui attribuent l'un et l'autre. On pourrait, par exemple, parler de l'expérience (i) comme d'être en apesanteur (ii) que l'on a à l'occasion d'un vol parabolique, ou en version abrégée (mais obscure), l'expérience d'apesanteur d'un vol parabolique.⁹⁵

Si on pense à « l'expérience de l'action », il faut donc savoir clairement si l'on veut parler de l'expérience comme d'être un agent ou de l'expérience que l'on a au moment où l'on agit. Il se pourrait que nous n'ayons pas toujours une expérience comme d'être en train d'agir au moment où nous agissons (par exemple, un violoniste à la fois très doué et passionné par l'œuvre qu'il joue pourrait très bien avoir une conscience phénoménale *entièrement* constituée du morceau qu'il joue et non pas du fait qu'il le joue) ; mais il se pourrait surtout

⁹⁵ Cette interprétation de l'ambiguïté entre sens objectif et sens subjectif comme un problème d'ambiguïté du complément d'objet est proposée notamment par (Bayne 2011, 363).

que nous ayons une expérience comme d'être un agent qui excède les vécus des moments où nous agissons. Dans la suite du texte, j'aurai donc soin de distinguer entre « notre expérience agentive (au sens objectif) » et « notre expérience agentive (au sens subjectif) ». La distinction est encore plus importante si l'on veut parler de « l'expérience du libre arbitre » : en effet, si l'on prend l'expression au sens objectif (et donc factif), parler d'une expérience du libre arbitre reviendra à présupposer l'existence effective du libre arbitre. Pour cette raison, l'expression « expérience du libre arbitre » sera toujours entendue ici au sens subjectif.

La deuxième distinction fondamentale qu'il nous faut introduire concernera le contenu de nos vécus : ce contenu pourra être présenté soit de manière phénoménale soit de manière simplement doxastique. Pour introduire cette distinction, quelques préliminaires sont nécessaires concernant le contenu des expériences.

Tout d'abord, la distinction qui m'intéresse ne concernera que les contenus *intentionnels* (et pas les « purs *qualia* »). J'ai dit un peu plus haut que pour une même expérience on pouvait la décrire soit au sens objectif selon son occasion, soit au sens subjectif selon son contenu. Mais tous les vécus n'ont pas le même genre de contenu : certains vécus pourraient même n'avoir pas un contenu suffisamment riche pour pouvoir être exprimé de manière conceptuelle sous la forme d'un complément de contenu. Ce serait le cas d'un vécu qui serait un « pur *quale* » : imaginons par exemple qu'un neuro-scientifique, en stimulant la zone ABC de mon cerveau, parvienne à me donner une expérience qualitative telle que je n'en ai jamais eue, incomparable avec aucune autre, et qui n'ait aucun contenu conceptuel. Je pourrai alors décrire ce vécu comme « l'expérience de stimulation de la zone ABC (au sens objectif) », mais je n'aurai aucun moyen de décrire cette même expérience au sens subjectif, de la décrire comme expérience « comme de quelque chose » (*as of X*). Pour pouvoir donner une description subjective de nos expériences, il faut donc que celles-ci aient un contenu *intentionnel* (ou représentationnel).

Parmi les vécus intentionnels, nous ne serons concernés ici que par les vécus intentionnels « *véridicaux* » (et pas les vécus *téliques*). Dans l'expérience agentive d'un potier, on pourra trouver peut-être la représentation du vase achevé, même si celui-ci n'est pas encore achevé (et ne le sera jamais). Ce vécu est clairement « intentionnel » ou représentationnel, mais ce qu'il représente, il ne le représente pas « comme étant le cas ». En reprenant l'exemple bien connu de la liste de course d'Anscombe (1957, para. 32), une liste de course aura normalement une « direction d'ajustement » allant du monde à l'esprit (c'est au monde de correspondre à ce qui est écrit sur la liste) ; en revanche, si un espion fait la liste de ce que j'achète, sa liste aura une direction d'ajustement de l'esprit au monde (c'est à lui de s'adapter à ce que, dans la réalité, j'achète). De la même façon, la représentation du vase achevé aura chez le potier une direction d'ajustement du monde à l'esprit, et sera appelé un vécu intentionnel *télique* (ou « intentionnel » dans le sens étroit de l'intention de l'action) : ce vécu (cette intention) sera satisfait si le monde s'y adapte. En cela, le vécu du potier se distingue, par exemple, de l'expérience visuelle d'un vase : si j'ai une telle expérience visuelle, elle ne sera correcte ou satisfaite que si elle correspond ou s'adapte à la manière dont le monde est réellement. Tous les vécus qui ont une direction d'ajustement de l'esprit au monde seront appelés *véridicaux*⁹⁶ : ils ont la particularité de pouvoir être *vrais* ou *faux*.

Parmi les vécus véridicaux, il nous faut encore éliminer une certaine catégorie ; celle des vécus véridicaux *sans engagement*. Dans cette catégorie entrent notamment tous les actes de pure imagination : si je choisis de me représenter la Tour Eiffel sous un ciel nuageux, je peux sans doute produire en moi un tel vécu. Et mon vécu aura des conditions de véridicité : il pourra ou non correspondre à la réalité, être vrai ou faux. Mais, à moins d'avoir une

⁹⁶ L'expression est empruntée à (Horgan, Tienson, and Graham 2003, 333) qui étudient les « conditions de véridicalité » de l'expérience agentive.

pathologie grave, j'aurai pleinement conscience que ce contenu peut aussi bien être faux que vrai, c'est-à-dire que ce vécu ne se présentera pas à moi avec le moindre « engagement » (ou la moindre « inclination ») à accepter la vérité de son contenu intentionnel. À l'inverse, lorsque je me trouve devant un arbre, mon vécu visuel « comme d'un arbre » m'engage ou m'incline à croire qu'il y a un arbre en face de moi. Les vécus qui nous intéresseront au cours de cette étude seront seulement les vécus véridicaux qui comportent une forme d'engagement ou d'inclination à la vérité.

Or parmi les vécus véridicaux « engageants », il faut encore distinguer deux grandes catégories, qu'on peut appeler « expérience phénoménale » et « expérience doxastique ». Dans la mesure où la notion « d'expérience doxastique » n'est pas universellement admise (ou en tout cas universellement utilisée) le plus simple est de partir de la notion d'expérience phénoménale pour étendre ensuite à l'expérience doxastique. C'est ainsi que procède notamment Alvin Plantinga :

On peut commencer par noter qu'il y a plusieurs sortes d'expériences. Premièrement, il y a l'*imagerie sensible*, le genre d'expérience que vous avez tout particulièrement dans la vision, mais aussi dans l'audition, l'odorat, le goût, et aussi le toucher. Pour employer la terminologie de Chisholm, dans ce genre d'expérience nous avons une *apparence* de telle ou telle sorte. L'imagerie sensible joue un rôle considérable dans la perception ; les croyances perceptives sont formées *en réponse* à l'imagerie sensible et *sur la base* de cette imagerie.

Pourtant, ce n'est pas le seul type d'expérience qui accompagne la formation de croyance. [...] la formation de souvenirs *mémoriels* n'est généralement pas accompagnée d'expérience phénoménale, ou alors seulement d'une imagerie sensible fragmentaire, fugitive, indistincte, difficile à observer. Vous vous souvenez que vous êtes allés à une soirée à Novosibirsk ; il y a un peu d'imagerie, certes, cependant, elle est fugitive, indistincte, et telle que lorsque vous essayez de l'observer en y appliquant votre attention, elle disparaît. Mais il y a un autre type d'expérience qui est présente : la croyance que c'était à Novosibirsk (et pas, par exemple, à Cleveland) semble *vraie, acceptable, naturelle* ; elle s'impose à vous ; elle semble d'une certaine manière inévitable (il est difficile de trouver les bons termes). La croyance *a l'air* d'être correcte, acceptable, naturelle. Elle ne vous fait pas le même effet qu'une croyance que vous pensez être fausse. La même chose s'applique à la croyance *a priori*.

(Plantinga 2000, 110–111)

C'est ce type d'expérience que Plantinga appelle « expérience doxastique » (*ibid.*) : le fait qu'une croyance « semble vraie, acceptable, naturelle » qu'elle « ait l'air d'être correcte ». Plantinga l'applique d'abord au cas des croyances mémorielles, mais comme il le mentionne, elle s'applique aussi au cas de la croyance *a priori*, et à vrai dire c'est *surtout* dans les discussions épistémologiques sur les croyances *a priori* (mathématiques ou philosophiques) que cette notion est apparue. Trois épistémologues ont accordé une importance particulière à ce type d'expériences, George Bealer (1996), Joel Pust (2000), et Michael Huemer (2007). Malgré quelques différences terminologiques, leurs raisonnements suivent exactement les mêmes étapes. Premièrement, ils font observer que les expressions du type « il me semble que *p* » ou « il m'apparaît que *p* » désignent un certain type d'attitude propositionnelle qu'on peut appeler une « apparence » : (Huemer 2007, 30) parle d'*appearance*, (Bealer 1996, 123) parle de *seeming*. Deuxièmement, cette attitude propositionnelle correspond à un « type d'état mental » (Huemer, loc. cit.) au sens fort, c'est-à-dire à un « un type véritable d'épisode conscient » (Bealer, loc. cit.), dont on peut chercher à donner une « analyse phénoménologique », c'est-à-dire à décrire « ce que cela fait d'un point de vue en première

personne » (Pust 2000, 31). C'est en ce sens que les « apparences » peuvent être comptées comme un type particulier « d'expériences », et pour ma part je garderai donc la terminologie de Plantinga, celle des « expériences doxastiques ». Dans un troisième temps, nos trois auteurs s'efforcent de montrer que ce type d'expérience est *sui generis*, en le distinguant d'autres expériences avec lesquelles on le confond parfois. Les « apparences » ou expériences doxastiques sont d'abord à distinguer des *croyances* (Bealer 1996, 123; Huemer 2007, 31) : par exemple, le principe de compréhension non restreinte (dont Russell a montré qu'en dépit de son caractère intuitif il menait à des contradictions logiques) peut continuer de me *sembler vrai* même lorsque j'ai reconnu qu'il était faux, exactement de la même façon que les lignes de Müller-Lyer peuvent continuer de me *sembler* de tailles différentes même lorsque j'ai reconnu qu'elles étaient de la même taille. Les expériences doxastiques sont aussi à distinguer des simples pressentiments (*hunch*), qui habituellement ne demeurent pas lorsque la preuve du contraire a été apportée (Bealer 1996, 124). Enfin la notion d'expérience doxastique est à distinguer de la notion de « disposition à croire » (Huemer 2007, 31) : d'une part, je peux être disposé à croire que *p* dans certaines circonstances sans que cette disposition se manifeste à moi phénoménologiquement de quelque façon que ce soit, ou je peux être disposé à croire que je vais gagner au jeu par simple vœu pieux, sans que cela me *semble vrai*. Certes, les expériences doxastiques constitueront souvent une bonne *explication* de nos dispositions à croire ceci ou cela, mais ce sont des catégories ontologiques distinctes.

A partir de ce qui précède, on peut donner des expériences doxastiques la définition suivante :

L'expérience doxastique que *p* (pour un sujet *S*) est l'attitude propositionnelle consciente qui est désignée par les expressions « il semble à *S* que *p* ».

A partir de cette notion générale, nos auteurs font ensuite remarquer qu'il existe plusieurs sortes d'expériences doxastiques. Il est important de noter, en particulier, que les expériences sensorielles, avec leur richesse phénoménale, font aussi *partie* de cette catégorie générale des expériences doxastiques : en effet, lorsque j'ai l'expérience visuelle comme d'un arbre en face de moi, il est certainement vrai « qu'il me semble (visuellement) qu'il y a un arbre en face de moi ». De même, les expériences mémorielles sont un certain type d'expérience doxastique, bien qu'elles soient phénoménologiquement très différentes des expériences doxastiques que l'on a pour les propositions *a priori* telles que « $2+2=4$ » : dans le cas d'une expérience doxastique mémorielle, il y a un certain « sentiment de passé » (Pust 2000, 34) qui ne se retrouve pas dans l'expérience doxastique *a priori*. On pourrait dire que les expériences doxastiques perceptives engagent quand à elles un « sentiment de présent » qui les distingue à la fois des expériences mémorielles et des expériences *a priori*. Compte tenu de leur intérêt pour l'épistémologie des croyances *a priori*, Bealer et Pust se sont particulièrement attachés à décrire et caractériser la sous-catégorie des expériences doxastiques *a priori*, qu'ils appellent « apparences intellectuelles » (*intellectual seemings*), ou plus couramment « intuitions » (*intuitions*) : en particulier, ces expériences sont généralement associées à une phénoménologie de la « nécessité ».⁹⁷ Huemer, quant à lui, s'intéresse plutôt à la catégorie générale des expériences doxastiques. C'est cette seconde démarche qui sera pertinente pour nous, dans la mesure où l'expérience doxastique associée à notre action (ou au libre arbitre) n'est certainement pas marquée par la nécessité au même titre que les intuitions mathématiques ou que les intuitions philosophiques telles que « si *S* sait que *p*, alors sa croyance que *p* n'est pas hasardeuse ».

⁹⁷ Lorsqu'il me semble que « $2+2=4$ », il fait sans doute partie de cette apparence qu'il me semble *nécessaire* que « $2+2=4$ ». Voir (Bealer 2000, 3).

Nous pouvons à présent appliquer les notions « d'expérience doxastique » et « d'expérience phénoménale » au cas de l'expérience agentive. Il est certain que lorsque nous agissons (lorsque nous faisons l'expérience objective de l'action), nous avons dans les cas normaux au moins une certaine expérience comme d'être en train d'agir (une expérience agentive subjective). Cette expérience agentive aura ordinairement un contenu phénoménal relativement riche : expérience de l'effort, expérience d'être *soi-même* en train d'agir, expérience d'agir d'après une intention, etc. (J'aborderai la question de ce contenu phénoménal au chapitre 4) Mais surtout, cette expérience agentive aura (normalement) un caractère « indexical » : ce sera l'expérience que *je* suis en train d'agir *maintenant*. Par ces deux caractéristiques, cette expérience agentive se rapproche des expériences perceptives. C'est cette expérience que j'appellerai « l'expérience agentive phénoménale » ou « expérience phénoménale de l'action ».

Par ailleurs, nous avons vu dans la section précédente qu'il existe une tendance naturelle à juger que nous sommes « libres » (dans le sens qui entre en conflit au moins apparent avec le déterminisme). Autrement dit, lorsqu'on propose à un individu humain la description de l'univers déterministe A, cet univers lui *semblera* (normalement) n'être pas semblable au nôtre. Nous avons donc, normalement, une expérience doxastique selon laquelle « nous avons le libre arbitre ».

La question posée au début de cette section peut donc être reformulée ainsi : quel est le rapport entre notre expérience phénoménale de l'action d'une part et notre expérience doxastique d'avoir le libre arbitre ? Et la réponse la plus simple semble être la suivante : notre expérience doxastique d'avoir le libre arbitre est *fondée* sur notre expérience agentive phénoménale. C'est ce que j'appellerai le « modèle phénoménal » de l'expérience du libre arbitre.

Modèle Phénoménal de l'expérience du Libre arbitre

L'expérience phénoménale que nous avons (normalement) au moment où nous agissons a un certain contenu conceptuel XYZ (par ex. « je suis en train d'agir intentionnellement, par moi-même, etc. »). Dans les circonstances normales, ce contenu XYZ n'est pas contredit par d'autres informations, et nous *croyons* donc ce contenu conceptuel. Ayant cru ce genre de contenu par expérience phénoménale, nous le conservons ensuite sous forme mémorielle en le généralisant (par ex. « j'ai parfois agi intentionnellement, par moi-même, etc. »). Et c'est cette croyance mémorisée qui suscite l'expérience doxastique du type « l'univers déterministe A n'est pas semblable au nôtre » ou « nous avons le libre arbitre ». C'est-à-dire que le concept central de notre expérience doxastique du libre arbitre *provient* de notre expérience phénoménale agentive.

Ce type de modèle est accepté, de manière plus ou moins implicite, par la grande majorité des auteurs incompatibilistes. On en trouve également une défense explicite chez Galen Strawson, qui soutient que « le vrai centre de notre engagement en faveur de la notion de liberté humaine » se situe dans notre expérience agentive en première personne (1986, sec. 6.5). Il peut être utile de comparer ce modèle phénoménal avec d'autres modèles de l'expérience du libre arbitre. Dans les autres modèles, le concept qui entre en conflit intuitif avec le déterminisme ne proviendrait pas de notre expérience agentive en première personne. D'où proviendrait-il alors ? Une alternative vient à l'esprit si l'on remarque que nous avons, outre des expériences d'action en première personne, des expériences d'action en troisième personne⁹⁸ : lorsque je vois un corps humain en train de couper un tronc d'arbre, j'ai non

⁹⁸ Ou éventuellement en deuxième personne ; mais la distinction entre deuxième et troisième personne ne semble pas ici pertinente.

seulement une certaine expérience visuelle des mouvements qui se produisent, mais aussi une certaine expérience que « cette personne est en train d'agir ». Il n'est pas absolument évident que le contenu de cette expérience de l'action en troisième personne soit *le même* (à la personne près) que le contenu de notre expérience agentive en première personne. Il est probable, par exemple, que notre expérience en première personne contient parfois un aspect d'*effort* qui n'est pas présent dans l'expérience en troisième personne. A cause de ce genre d'exemple, on aura tendance à penser que notre expérience en première personne est probablement *plus riche* que notre expérience en troisième personne, mais là encore, ce n'est pas absolument évident. Et on pourrait imaginer que le composant de « libre arbitre » – ce composant conceptuel qui entrera ensuite en conflit avec l'idée de déterminisme – ne soit présent que dans notre expérience en troisième personne. Autrement dit : notre conviction de la liberté d'autrui serait plus directe et fondamentale que notre conviction de notre propre liberté. Cette hypothèse fait penser à la « preuve de Duns Scot » en faveur du libre arbitre. Au lieu d'en appeler à notre expérience individuelle de l'action, Duns Scot s'est rendu célèbre par la « preuve » suivante :

Ceux qui nient que certaines choses sont contingentes devraient être soumis à la torture jusqu'à ce qu'ils concèdent qu'il est possible qu'ils ne soient pas torturés.
(Duns Scotus 1987, 9)

On peut interpréter cette preuve de la contingence de la manière suivante : un sujet torturé sera toujours bien assez convaincu du libre arbitre *de son bourreau* – plus facilement peut-être que de sa propre liberté. J'appellerai donc « modèle scotiste » ce modèle en troisième personne :

Modèle Scotiste de l'expérience du Libre arbitre

L'expérience phénoménale que nous avons (normalement) au moment où nous voyons *quelqu'un d'autre* agir a un certain contenu conceptuel XYZ. Dans les circonstances normales, nous *croyons* ce contenu XYZ, puis nous le conservons sous forme mémorielle. Et c'est cette croyance mémorisée qui suscite l'expérience doxastique du type « l'univers déterministe A n'est pas semblable au nôtre » ou « nous avons le libre arbitre ».

Dans le modèle « scotiste », notre croyance au libre arbitre provient encore d'une certaine expérience phénoménale. Mais il existe un autre modèle dans lequel notre concept de libre arbitre vient non pas d'une certaine expérience phénoménale mais plutôt du réseau conceptuel de la morale. C'est ce que j'appelle le « modèle moral ».

Modèle moral de l'expérience du Libre arbitre

Lorsqu'il est éduqué en société, un être humain apprend normalement à utiliser (et à approuver) les concepts moraux, tels que « S doit faire X » ou « S ne doit pas faire X ». Dans ce réseau conceptuel est engagé une certaine conception UVW des agents et de leurs actions – une conception beaucoup plus riche que le donné XYZ de notre expérience agentive. Et c'est cette conception UVW qui suscite l'expérience doxastique du type « l'univers déterministe A n'est pas semblable au nôtre » ou « nous avons le libre arbitre ».

Je suppose qu'on pourrait encore imaginer d'autres modèles, mais ces deux exemples seront suffisants pour mon propos.

La question principale de cette section était de savoir dans quelle mesure le « conflit intuitif » (qui permet de définir le concept de « libre arbitre ») repose sur une « expérience » de nous-mêmes et de nos actions. Si l'on prend « expérience » au sens de l'expérience phénoménale de l'action, alors la question n'est pas triviale : il semble au moins initialement *probable* que le conflit intuitif repose ultimement sur le contenu conceptuel de notre expérience agentive en première personne, mais d'autres modèles sont peut-être envisageables. Dans la deuxième partie (chapitre 4), je présenterai des résultats récents de X-Phi qui suggèrent fortement la validité du « modèle phénoménal ». Mais pour la première partie de mon étude, ce qui sera important (épistémologiquement), c'est seulement que nous ayons une « expérience du libre arbitre »... une expérience dans le sens le plus large qui soit. Or il ne fait pas de doute que le conflit intuitif révélé par la X-Phi repose sur une expérience *doxastique* d'avoir le libre arbitre. Ou plus précisément, il ne « révèle » pas une telle expérience, il *est* en lui-même l'expérience doxastique, le fait qu'il ne nous *semble pas* que notre univers soit semblable à l'univers déterministe A.

Par conséquent, la définition du concept de « libre arbitre » à partir du conflit intuitif est également (de manière équivalente) une définition du concept de « libre arbitre » à partir d'une certaine expérience (doxastique) que nous avons de nous-mêmes et de nos actions. On peut donc douter que nous ayons une expérience phénoménale d'être libres, mais on ne peut pas douter que nous ayons une expérience doxastique d'être libres, car c'est cette expérience même qui permet de fixer le sens du concept de « libre arbitre ». Cette expérience (doxastique) d'être libres sera au cœur de la discussion épistémologique des deux chapitres suivants.

IV. Conclusion : L'expérience du Libre arbitre comme point de départ

Dans ce chapitre, je suis parti du problème de la définition du concept de « libre arbitre » : avant même de proposer une *analyse* conceptuelle de tel ou tel concept, comment s'assurer que nous parlons bien du *même* concept ? Ce problème est récurrent dans le débat sur le libre arbitre et j'ai essayé de classer les différents types de solutions. La solution « confiante » n'est pas satisfaisante car elle ne donne aucune garantie que nous parlons bel et bien d'un même concept. La solution « stipulative », à l'inverse, s'assure que nous parlons bien du même concept... mais au risque de changer de sujet par rapport à la discussion traditionnelle. La seule solution qui permet de s'assurer que nous parlons bien d'un *seul* concept, et que ce concept est bien *celui qui est à l'origine* du débat philosophique, c'est de donner une définition « descriptiviste » du concept : le libre arbitre est le concept (quel qu'il soit) qui est à l'origine du conflit intuitif (illusoire ou profond) avec l'idée de déterminisme.

Cette définition descriptiviste a, clairement, des conséquences importantes sur la suite de la discussion. Encore faut-il être précis au sujet de ces conséquences. Une telle définition implique que notre concept de libre arbitre a une propension naturelle à entrer en conflit intuitif avec le déterminisme, mais elle *n'implique pas* qu'il y ait une incompatibilité conceptuelle entre libre arbitre et déterminisme. De même, une telle définition implique que nous avons une expérience doxastique d'être libres, mais elle *n'implique pas* que nous soyons effectivement libres, ni même que nous soyons justifiés à le croire, ni enfin que cette expérience doxastique repose sur notre expérience phénoménale agentive. En fin de compte, je défendrai chacun de ces points : c'est-à-dire, je défendrai que le concept qui entre en conflit doxastique avec le déterminisme provient probablement de notre expérience phénoménale agentive et qu'il est probablement un concept intrinsèquement incompatibiliste (chapitre 4), et je défendrai que cette expérience nous justifie à croire que nous sommes effectivement libres (chapitres 2 et 3) et nous permet même de le *savoir* (chapitres 3 et 6). Mais ces thèses ne sont

pas engagées par la définition descriptiviste que j'ai donnée. Cette définition serait compatible avec un ensemble de thèses exactement inverse, par exemple : notre expérience doxastique d'êtres libres ne repose sur aucune expérience phénoménale, mais seulement sur le vocabulaire conceptuel des normes morales, et le concept de libre arbitre qui est tiré de la moralité n'est pas en lui-même incompatible avec le déterminisme, mais nous avons seulement une propension naturelle à mal comprendre le rapport entre ces deux concepts. La définition descriptiviste pourrait donc être admise aussi bien par un compatibiliste tel que John Stuart Mill, ou par un incompatibiliste dur tel que Derk Pereboom ; bien plus, il me semble que la plupart des auteurs, même lorsqu'ils n'utilisent pas une définition explicitement descriptiviste, reposent implicitement sur une telle démarche pour déterminer « de quel type de liberté il s'agit » dans la discussion sur le libre arbitre.

Bien qu'elle n'engage pas de « parti pris », il ne faut pas sous-estimer les conséquences de cette définition. Comme on vient de le voir, cette définition implique logiquement que les hommes ont tous (ou presque tous) l'impression (l'expérience doxastique) d'avoir le libre arbitre. Ce résultat peut sembler étonnant : comment pourrions-nous trancher une telle question empirique *simplement par définition* ?

Pour dissiper cet étonnement, il convient d'une part de se rappeler les données de X-Phi qui ont été présentées dans ce chapitre. Mais d'autre part, et peut-être plus fondamentalement, il faut se poser la question suivante : que pourrait *vouloir dire* l'idée que « les hommes n'ont pas naturellement l'impression d'avoir le libre arbitre » ? Quel *sens* pourrait-on donner à « libre arbitre » dans une telle affirmation ? On pourrait certainement lui donner un sens stipulatif (comme « la capacité de faire autrement », ou « pouvoir faire ce que l'on veut », etc.) et on aurait alors formé une thèse intelligible et requérant une investigation empirique. Mais on prendrait en même temps le risque de ne pas parler de la même chose que d'autres personnes prenant part à la « discussion métaphysique sur le libre arbitre », car dans cette discussion historique, le mot « libre arbitre » n'est pas employé de cette manière-là, avec un sens stipulatif ; ce n'est pas en se référant à une même définition que les membres de la discussion s'assurent qu'ils utilisent un seul et même concept, mais c'est en partant d'un seul et même conflit intuitif qu'ils reconnaissent tous en eux-mêmes et tiennent pour l'enjeu de la discussion. Certes, c'est une question « empirique » (une question réfutable par observation) que de savoir si les intuitions des uns et des autres utilisent bel et bien le *même* concept ; mais si ce n'est pas le cas, alors la conclusion à tirer est qu'il n'y a pas d'objet commun de la discussion, pas de sens au mot « libre arbitre », et non pas que certains ont l'intuition d'avoir le libre arbitre tandis que d'autres ne l'ont pas.

Pour les lecteurs qui auraient des réticences envers la définition descriptiviste du libre arbitre, je soulignerai enfin que la majeure partie de mes réflexions dans cette étude peut être lue en faisant abstraction de cette définition. On pourrait en effet lire mes réflexions en prenant ma définition du libre arbitre comme une définition elle-même stipulative : « quel que soit le sens habituel de l'expression 'libre arbitre', elle sera ici employée pour signifier le concept – s'il y en a – qu'à peu près tout le monde possède, etc. ». En prenant acte de cette stipulation terminologique, le lecteur ne devrait pas trouver étonnante l'affirmation que « presque tout homme a spontanément l'impression d'avoir le 'libre arbitre' ».⁹⁹ Et il devrait également pouvoir s'intéresser aux questions qui m'occuperont dans les chapitres suivants, à savoir : (chapitre 2) cette expérience doxastique confère-t-elle une justification en faveur de son contenu ? (chapitre 3) pouvons-nous considérer cette expérience doxastique comme une source de connaissance de son contenu ? (chapitre 4) le contenu de cette expérience est-il compatible ou incompatible avec le concept de déterminisme ? (chapitre 5) le contenu de cette expérience est-il démontrablement faux ? (chapitre 6) le contenu de cette expérience est-il

⁹⁹ Au besoin, il pourra compléter ce genre d'affirmation en restituant mentalement la conditionnelle implicite : « ... s'il y a bien un tel concept de libre arbitre ».

inconnaissable ? Toutes ces questions peuvent être comprises et considérées avec intérêt, même par un auteur qui estimerait qu'elles n'ont rien à voir avec le « libre arbitre » lorsque cette expression est définie correctement. Par exemple, on pourrait imaginer un auteur déterministe dur qui dirait la chose suivante : « Le libre arbitre doit être défini stipulativement comme la faculté d'agir autrement dans les mêmes circonstances, cette faculté est très exigeante métaphysiquement : elle est incompatible avec le déterminisme, et nous n'avons *pas* une telle faculté ; *par ailleurs*, nous avons une certaine expérience de nos actions (par ex. l'expérience d'être la source personnelle de nos actions) qui est très peu exigeante métaphysiquement, mais qui (à cause de certaines confusions) est à l'origine du conflit intuitif récurrent avec l'idée de déterminisme ; cette expérience constitue une source de justification et de connaissance tout à fait légitime en faveur de son contenu modeste, mais elle n'a rien à voir avec la question du libre arbitre ». J'ai deux types de désaccord avec cet auteur, mais les deux n'ont pas la même importance. Il y a tout d'abord un désaccord sur le sens des mots : pour moi, un tel auteur devrait être décrit comme un compatibiliste (tout le problème traditionnel reposait sur une *illusion* d'incompatibilité), un compatibiliste qui aurait par ailleurs des intérêts pour une autre faculté, incompatible quant à elle avec le déterminisme. Mais ce désaccord est mineur à mes yeux ; le désaccord de fond concerne l'expérience de nos actions. Cet auteur pense (comme moi) que cette expérience est une source de justification et de connaissance, mais il affirme par ailleurs (contrairement à ce que je défendrai dans le chapitre 4) que cette expérience a un contenu strictement compatibiliste. C'est ce genre de désaccord de fond – des désaccords à propos de notre expérience, de son contenu et de sa valeur épistémique – qui m'intéresseront vraiment au cours de cette étude.

Chapitre 2 : Justification *prima facie* à croire que nous sommes libres

The misguided, and as a rule quite uncritical, belittlement, of the evidence offered by inner experience has, I am convinced, been responsible for more bad argument by the opponents of Free Will than has any other single factor.
(Campbell 1957, 168)

I Introspectionnisme vs Anti-Introspectionnisme	81
I.1) Descartes, Leibniz, Spinoza	81
I.2) Introspectionnistes et Anti-Introspectionnistes dans le débat analytique	82
I.3) Comment s'orienter dans le débat : démarche par modèles et démarche par construction	87
I.4) Comment s'orienter dans le débat : l'introspectionnisme a-compatibiliste	90
I.5) Argument Pro : la motivation doxastique en faveur de l'Introspectionnisme ..	93
I.6) Arguments Contra	96
II Trois interprétations de l'objection anti-introspection	99
II.1) L'objection de la pure impression	99
II.1.a. L'objection et son principe inférentialiste	99
II.1.b. Portée sceptique de l'inférentialisme	106
II.1.c. Solution : évidentialisme faible et justification basique	110
II.1.d. Non pertinence des solutions cohérentiste et infinitiste	117
II.2) L'objection de l'impression faillible	123
II.2.a. L'objection et son principe infaillibiliste	124
II.2.b. Portée sceptique de l'infaillibilisme	128
II.2.c. Solution : sécurité faillibiliste de la connaissance, et justification <i>prima facie</i>	131
II.3) L'objection de l'impression lourde	147
II.3.a. L'objection et son principe de clôture	149
II.3.b. L'argument sceptique de la clôture	153
II.3.c. Solution : transmission de justification <i>prima facie</i> par déduction compétente	156
III Conclusion : la justification indexicale de premier ordre en faveur du Libre Arbitre	170

Dans le premier chapitre, nous avons défini le concept de libre arbitre comme étant le concept responsable du conflit intuitif que génère, chez presque tout homme, la possibilité du déterminisme. Comme on l'a vu, cette définition permet de répondre directement à la question suivante : les hommes ont-ils tous (ou presque tous) l'impression d'avoir le libre arbitre ? Par définition, la réponse est oui. Et par ailleurs, l'observation courante et les données empiriques de la X-Phi confirment que cette définition n'est pas vide : il y a bien une impression commune d'être libres.

Je prendrai donc comme point de départ de ma réflexion épistémologique le principe suivant:

(a) Tout le monde ou presque a une expérience (doxastique) E d'avoir le libre arbitre.¹⁰⁰

Tout l'enjeu du débat sur le Libre arbitre est de savoir si cette expérience est *véridique* ou si elle est illusoire. En d'autres termes, l'enjeu est de savoir si le contenu conceptuel de E (« que nous sommes libres ») est vrai ou non. Il est important à ce stade de ne pas confondre la question de la véridicité avec celle de la fiabilité : si je suis en train de rêver que je suis dans un palais, et qu'en réalité j'ai été transporté dans un palais pendant mon sommeil, mon expérience est véridique, mais elle n'est évidemment pas fiable. L'*enjeu* de la discussion du libre arbitre n'est pas de savoir si notre expérience est fiable ou non, certaine ou non, source de connaissance ou non, etc. Toutes ces questions apparaissent dans la discussion comme des *étapes* du raisonnement. Mais l'enjeu lui-même, c'est de savoir si, oui ou non, nous avons le libre arbitre – si le *contenu* de cette expérience est vrai ou faux.

C'est également l'enjeu de ma thèse. Mais plus précisément, ma thèse s'intéresse à la question épistémologique suivante :

(b) Comment déterminer d'une manière épistémologiquement adéquate si le contenu de E est vrai ou faux ?

D'après la terminologie définie en introduction, cette question se subdivise en une question sur la *connaissance* et une question sur la *justification*.

(K) Question de Connaissance : Comment pourrions-nous *savoir* que le contenu de E est vrai ou qu'il est faux ?

(J) Question de Justification : Comment pourrions-nous être *justifiés à croire* que le contenu de E est vrai ou qu'il est faux ?

Dans ce chapitre je m'occuperai principalement de la Question de Justification, et la Question de Connaissance n'apparaîtra que dans la mesure où un défaut de connaissance (réel ou supposé) peut avoir des effets sur la Justification. Qu'est-ce qui pourrait donc nous donner une justification à croire que nous sommes libres, dans le sens fixé par l'expérience E ?

Une réponse naïve vient spontanément à l'esprit si on compare l'expérience E avec d'autres expériences. Supposons par exemple que j'aie l'expérience visuelle d'une table en face de moi. Qu'est-ce qui pourrait donc, dans un tel contexte, me fournir, une justification à croire qu'il y a une table en face de moi (i.e. à croire le contenu de cette expérience) ? La question devrait sembler curieuse à tout lecteur non philosophe tant sa réponse peut sembler évidente : c'est évidemment mon expérience visuelle qui peut me justifier à croire qu'il y a une table en face de moi ! De la même façon, si nous cherchons comment nous pourrions être justifiés à croire au libre arbitre et s'il est établi que nous avons une expérience d'avoir le libre arbitre, cette expérience semble être un bon candidat pour fournir la justification recherchée. Cela n'exclut pas qu'il puisse y avoir d'autres justifications pour la même conclusion (et des justifications pour la conclusion inverse), mais au moins, il semble naturel de penser que notre expérience E d'avoir le libre arbitre représente au moins un des candidats pertinents pour répondre à la Question de la Justification.

¹⁰⁰ Dans les deux chapitres qui viennent, je ne rappellerai pas systématiquement que l'expérience E est une expérience « doxastique » plutôt qu'une expérience « phénoménale ». L'expérience phénoménale de l'action ne réapparaîtra qu'au chapitre 4.

Mais peut-on vraiment faire de l'expérience commune E une source de justification pour la croyance au Libre arbitre ? Plusieurs philosophes ont considéré ce genre de justification comme totalement irrecevable : pour justifier une croyance, on ne peut pas s'appuyer sur une simple impression subjective ; une impression subjective peut très bien être trompeuse ; il faut donc produire une raison partageable, un argument. Pour ces auteurs, il est totalement illégitime de faire jouer le moindre rôle à « l'introspection » de nos expériences dans le débat sur le Libre arbitre. Pour simplifier, on pourrait classer les différentes épistémologies du Libre arbitre entre Introspectionnistes et Anti-Introspectionnistes, les secondes souscrivant à la thèse suivante :

(Anti-Introspection) L'expérience E d'avoir le libre arbitre n'apporte aucune justification à croire au libre arbitre.

La position que je vais défendre appartient à la famille Introspectionniste. L'introspectionnisme est une position qui concerne à la fois la *justification* et la *connaissance* du Libre arbitre, mais pour bien la comprendre il faut commencer par la justification. J'aborderai la question de connaissance au chapitre 3 (et plus encore dans la troisième partie de la thèse) et j'envisagerai ensuite, au chapitre 4, les théories rivales.

Dans ce chapitre, je présenterai brièvement l'opposition entre introspectionnisme et anti-introspectionnisme telle qu'elle apparaît dans la littérature classique et contemporaine (I). Pour répondre à l'objection anti-introspection, je considérerai trois interprétations possibles, c'est-à-dire trois arguments qui permettent d'établir la thèse (Anti-Introspection). Et je montrerai que chacun de ces arguments est, fondamentalement, un argument sceptique (II). Par conséquent, quiconque veut éviter de tomber dans le scepticisme est obligé d'accepter une conception de la justification plus tolérante que celle qui sous-tend l'anti-introspectionnisme : c'est la notion de justification *prima facie* que je développerai et appliquerai à l'expérience du Libre arbitre (III).

I. Introspectionnisme vs Anti-Introspectionnisme

I.1) Descartes, Leibniz, Spinoza

Historiquement, le *locus classicus* de cette question est l'article 39 des *Principes de la Philosophie* de Descartes, et l'ensemble des critiques qu'il a suscitées. Descartes, qui est d'ailleurs un des premiers auteurs de la tradition à proposer une épistémologie explicite pour la croyance au Libre arbitre, affirme que

« la liberté de notre volonté se connaît sans preuve, par la seule expérience que nous en avons. » (Descartes 1953, 588)

Cette théorie cartésienne de la connaissance par « expérience »,¹⁰¹ reprise et développée par Malebranche,¹⁰² a rapidement fait l'objet de nombreuses critiques, dont les plus célèbres sont les « parodies » de Spinoza et de Leibniz.¹⁰³

¹⁰¹ Descartes présente sa théorie en termes de « connaissance » plutôt qu'en termes de « justification », mais il est clair que l'expérience d'être libre est également, pour lui, ce qui nous justifie à croire au libre arbitre en

« La raison que M. Descartes a alléguée, pour prouver l'indépendance de nos actions libres par un prétendu sentiment vif interne n'a point de force. [...] C'est comme si l'aiguille aimantée prenait plaisir de se tourner vers le nord ; car elle croirait tourner indépendamment de quelque autre cause, ne s'apercevant pas des mouvements insensibles de la matière magnétique. » (Leibniz 1969, 131–132)

La parodie choisie par Spinoza est celle d'une pierre douée de conscience, qui se croirait libre simplement parce qu'elle aurait conscience de son « effort » vers le bas.¹⁰⁴ Qu'il s'agisse de pierre ou d'aiguille aimantée, l'objection est fondamentalement la même : notre expérience de la liberté ne peut aucunement être une bonne raison de penser que nous sommes libres puisque, en tant que simple expérience, elle est parfaitement compatible avec un déterminisme complet du porteur de cette expérience.

Bien sûr, Spinoza et Leibniz divergent ensuite sur le verdict final concernant le libre arbitre : pour Leibniz, il y a une *autre* raison de croire à la liberté humaine, tandis que pour Spinoza, toute justification de croire au libre arbitre s'effondre avec l'échec de cette justification par l'expérience. Mais ce qui nous intéresse ici, c'est leur accord sur l'affirmation suivante :

(Anti-Introspection) L'expérience E d'avoir le libre arbitre n'apporte aucune justification à croire au libre arbitre.

I.2) Introspectionnistes et Anti-Introspectionnistes dans le débat analytique

Dans le débat analytique contemporain, les grandes options philosophiques restent fondamentalement les mêmes.

Un certain nombre d'auteurs défendent le libre arbitre en s'appuyant sur notre expérience commune d'être libres :

même temps qu'elle en fonde la connaissance. Leibniz, dans la citation suivante, attaque plus clairement le caractère « justifiant » de l'expérience en parlant d'une « *raison* [qui] n'a point de force ».

¹⁰² Malebranche parle de « sentiment intérieur » plutôt que d'expérience : « Lorsqu'un bien particulier nous est présenté, nous avons sentiment intérieur de liberté à son égard, comme nous en avons de notre plaisir et de notre douleur, lorsque nous sentons. [...] Ainsi ayant sentiment intérieur de notre liberté dans le temps qu'un bien particulier se présente à notre esprit, nous ne devons point douter que nous ne soyons libres à l'égard de ce bien. » (Malebranche 1979, 807–808).

¹⁰³ Le mot « parodie » est pris ici dans son sens philosophique précis : la parodie d'un argument A consiste à montrer que A n'est pas acceptable en produisant un argument A*, qui repose sur le même principe que A, et dont la conclusion est manifestement absurde. Un argument qui a fait l'objet de réfutations parodiques célèbres est l'argument ontologique en faveur de l'existence de Dieu. On pense notamment à la parodie de « l'île parfaite », par Gaunilon, qui semble être la toute première réponse parodique à l'argument d'Anselme. Pour un traitement contemporain des parodies de l'argument ontologique, voir (Oppy 1995, chap. 11).

¹⁰⁴ « Concevez maintenant, si vous voulez bien, que la pierre, tandis qu'elle continue de se mouvoir, pense et sache qu'elle fait effort, autant qu'elle peut, pour se mouvoir. Cette pierre assurément, puisqu'elle a conscience de son effort seulement et qu'elle n'est en aucune façon indifférente, croira qu'elle est très libre et qu'elle ne persévère dans son mouvement que parce qu'elle le veut. Telle est cette liberté humaine que tous se vantent de posséder et qui consiste en cela seul que les hommes ont conscience de leurs appétits et ignorent les causes qui les déterminent. » (Spinoza 1986, 303–304)

« Ce qui plaide en faveur de la théorie [de la causalité agentive], c'est d'abord et avant tout qu'elle correspond à la manière dont nous expérimentons notre propre activité. »
(O'Connor 1995, 199)

« Au cours de notre vie, nous avons toutes sortes d'expériences au cours desquelles il apparaît comme un simple fait d'expérience que, même si nous avons fait une chose, nous sentons que nous savons bien que nous aurions pu faire autre chose. »
(Searle 1984, 87)

Ce que Searle et O'Connor défendent ici, à partir de notre expérience, c'est non seulement le libre arbitre, mais une certaine forme *libertarienne* du libre arbitre.¹⁰⁵ De fait, les défenses explicites de la position Introspectionniste sont généralement le fait de philosophes libertariens, se réclamant soit d'un lointain héritage cartésien,¹⁰⁶ soit plus fréquemment de l'héritage de Thomas Reid, dont l'épistémologie du libre arbitre repose fondamentalement sur la conscience que nous avons de notre propre activité en tant que cause agentive :

« Dans certains mouvements de mon corps, et certaines directions de ma pensée, je sais non seulement qu'il doit y avoir une cause ayant le pouvoir de produire ces effets, mais aussi que je suis cette cause ; et je suis conscient de ce que je fais pour les produire. C'est de la conscience de notre propre activité que provient, manifestement, le concept le plus clair, et même l'unique concept que nous puissions former d'une activité, ou d'un exercice de pouvoir actif. »
(Reid 1788, 36)

Entre Reid et O'Connor se trouve toute une tradition assez repérable (sinon absolument continue) de *libertariens* défenseurs de l'introspectionnisme.¹⁰⁷

De fait, on pourrait facilement croire que libertarianisme et introspectionnisme sont indissociables ou au moins de proches parents. Pourtant, nous allons voir qu'il y a de

¹⁰⁵ Searle soutient ici le pouvoir d'agir autrement, et O'Connor soutient une théorie libertarienne plus précise encore, celle de la causalité agentive [*agent causation*] que nous étudierons en détail au chapitre 5 (sec.

¹⁰⁶ Descartes lui-même peut-il être considéré comme un auteur « libertarien » ? C'est en tout cas l'interprétation la plus traditionnelle, qui repose essentiellement sur la première définition de la volonté libre dans la quatrième méditation, volonté libre qui « consiste seulement en ce que nous pouvons faire une chose, ou ne la faire pas » (Descartes 1953, 305). Cette lecture libertarienne de Descartes a fait l'objet de vifs débats, notamment à cause de la *seconde* définition proposée par Descartes juste après la première, et qui a des accents beaucoup plus compatibilistes : « ou plutôt seulement en ce que, pour affirmer ou nier [...] nous agissons en telle sorte que nous ne sentons point qu'aucune force extérieure nous y contraigne. » (*ibid.*). Le plus grand problème concerne les cas où nous reconnaissons de manière claire et distincte le vrai et le bien : Descartes semble parfois dire que, dans ces cas au moins, nous inclinons de manière *nécessaire* et néanmoins *libre*. Ces débats interprétatifs ont été bien résumés par (Ragland 2006). Ragland tranche finalement dans le sens d'une interprétation libertarienne, sur la base d'une lettre à Mesland qui ne semble pas laisser place au doute : « cette faculté positive que nous avons de nous déterminer à l'un ou à l'autre de deux contraires [...] j'estime qu'elle se rencontre, non seulement toutes les fois que [la volonté] se détermine à ces sortes d'actions, où elle n'est point emportée par le poids d'aucune raison vers un côté plutôt que vers un autre ; mais même qu'elle se trouve mêlée dans toutes ses autres actions, [...] lors même qu'une raison fort évidente nous porte à une chose. » (Descartes 1964, vol. 3, p.379)

¹⁰⁷ On peut mentionner au moins trois jalons importants de cette tradition : C.A. Campbell, Roderick Chisholm et Richard Taylor. Campbell (1957, 168–178) s'attache particulièrement à défendre le rôle épistémique de « l'expérience intérieure » (*inner experience* – voir la citation en exergue)... mais comme nous le verrons son épistémologie est très ambiguë et contient une défense quasi kantienne des réquisits de notre « être pratique ». Chisholm (1964, sec. 9) et (Taylor 1963, chap. 5) reprennent assez fidèlement une épistémologie reidienne, qui oppose la force des « données » du sens commun à toute forme de « théorisation ». Keith Lehrer, élève de Taylor, développe également ce type de modèle dans (Lehrer 1960). C'est au chapitre 3 que je présenterai en détail ces modèles, car en dépit d'une certaine confusion chez ces auteurs, il faut plutôt les comprendre comme des modèles de justification *méta-épistémique*.

nombreux auteurs libertariens qui attaquent l'introspectionnisme ; et inversement, dans la catégorie des Introspectionnistes, qui nous intéresse pour l'instant, on trouve également des auteurs compatibilistes, comme par exemple Christopher Peacocke :

« Dans les délibérations pratiques ou théoriques de la vie ordinaire, il semble que plusieurs possibilités nous soient ouvertes. Nous avons ordinairement l'impression subjective que nous sommes libres de choisir entre plusieurs options. [...] Puisqu'il est plausible que ces impressions jouent ordinairement un rôle causal et rationnel dans les actes de pensée par lesquels un individu s'attribue à lui-même la liberté, toute tentative pour relever le Défi de l'Intégration [...] devra en tenir compte, et expliquer leur rôle. » (Peacocke 1999, 336)

Pour Peacocke, le défi de l'Intégration s'adresse au compatibiliste tout autant qu'à l'incompatibiliste (même s'il est plus difficile pour ce dernier) ; et le compatibiliste *doit* donc (pour relever le défi) accorder un rôle justificationnel à notre expérience commune. Non seulement il le doit mais, d'après Peacocke, il le peut (au moins dans une certaine mesure) :

« Il y a incontestablement des raisons *prima facie* de sens commun à croire que nous sommes libres à l'égard de certains facteurs. » (Peacocke 1999, 338)

Je commenterai plus loin le rapport complexe entre les deux débats (sur l'introspectionnisme et sur le compatibilisme). Continuons pour l'instant notre tour d'horizon.

La deuxième catégorie est constituée de philosophes qui, à l'instar de Leibniz, défendent le libre arbitre tout en critiquant le recours à l'introspection. Parmi ceux-ci se trouvent, bien sûr, des compatibilistes qui se réclament généralement de John Stuart Mill :

« La volition, un état de notre esprit, est l'antécédent ; le mouvement de nos membres en conformité à cette volition, est le conséquent. Cette séquence, telle que je la conçois, n'est pas un objet de conscience directe [...] Je ne peux pas admettre que notre conscience de la volition contienne en elle-même une connaissance *a priori* que le mouvement musculaire s'ensuivra. » (Mill 1973, 355)

Cette critique de Mill est évidemment une reprise des objections Humiennes à l'introspectionnisme de Reid.¹⁰⁸

À côté de l'anti-introspectionnisme compatibiliste se trouve également une forte tradition de libertariens qui critiquent l'introspectionnisme, parfois avec une certaine virulence, pour prendre leurs distances avec une épistémologie qui aurait fait, d'après eux, plus de mal que de bien à la crédibilité du libertarianisme. On peut mentionner, parmi les représentants les plus importants de cette tradition, William James, Alan Donagan, et Peter van Inwagen.

« [Notre choix] nous apparaît certainement comme indéterminé [...]. S'il est réellement indéterminé, nos actes futurs sont incertains ou non-prédestinés : en langage courant, *nos volontés sont libres*. [...] Mais, d'un autre côté, il y a la certitude que toutes nos

¹⁰⁸ Voir en particulier la septième section de l'*Enquête* : « The motion of our body follows upon the command of our will. Of this we are every moment conscious. But the means, by which this is effected; the energy, by which the will performs so extraordinary an operation; of this we are so far from being immediately conscious, that it must for ever escape our most diligent enquiry. » (Hume 2007, 47).

volitions *sans effort* sont les résultantes d'intérêts et d'associations dont la force et l'enchaînement sont déterminés mécaniquement [... Or] même pour la volition sans effort, nous avons la conscience de la possibilité de l'alternative. C'est indéniablement une illusion dans ce cas-là ; pourquoi ne serait-ce pas une illusion dans tous les cas ? » (James 1983, 1175–1176)

« Malheureusement, la liberté du choix telle qu'on l'entend habituellement est parfois défendue [...] comme quelque chose que l'on peut connaître par expérience. [...] Mais] lorsque j'exerce ma capacité à choisir, je n'expérimente pas mon pouvoir d'exercer cette capacité autrement, comme je pourrais expérimenter mon choix comme plaisant ou difficile. [...] Si l'expérience de pouvoir choisir autrement que je ne choisis est quoi que ce soit d'autre que la conscience de faire un choix et la croyance que j'aurais pu me retenir de le faire, je n'ai aucune idée de ce qu'elle peut être. » (Donagan 1987, 180–182)

« Certains philosophes croient qu'il est possible de découvrir si nous avons le libre arbitre par *introspection*. Mais cela semble trivialement faux, puisque si cela était vrai, nous pourrions découvrir par introspection si nous sommes équipés de dispositifs martiens [i.e. des dispositifs que des Martiens auraient implantés dans nos cerveaux de manière totalement indétectable, pour manipuler chacune de nos décisions, tout en simulant en même temps les « expériences d'agir » que nous pouvons avoir normalement]. Et, évidemment, nous ne le pouvons pas. » (van Inwagen 1983, 204)

Comme ces auteurs affirment l'existence du libre arbitre, ils doivent évidemment produire une épistémologie de substitution, et c'est là que nos trois auteurs divergent. James adopte une épistémologie pragmatiste ou volontariste qu'on peut rattacher à la tradition fideïste.¹⁰⁹ Pour Donagan, la croyance au libre arbitre est une présupposition qui doit être acceptée comme condition de possibilité fondamentale de notre manière de vivre.¹¹⁰ Enfin, van Inwagen fonde la croyance au libre arbitre sur la croyance en la responsabilité morale.¹¹¹ Je discuterai plus loin ces types d'épistémologie – volontariste, transcendantale, par la moralité. Ce qui m'intéresse ici, c'est seulement les objections de ces auteurs à l'introspectionnisme.

¹⁰⁹ « When [...] objective proof is not to be had, the only course is voluntary choice. [...] If [...] the will be undetermined, it would seem only fitting that the belief in its indetermination should be voluntarily chosen from amongst other possible beliefs. Freedom's first deed should be to affirm itself. We ought never to hope for any other method of getting at the truth if indeterminism be a fact. » (James 1983, 1177) L'épistémologie sous-jacente est celle que James a développée dans sa célèbre controverse avec Clifford (James 1897) ; dans ce contexte, il se rattache explicitement à l'héritage fideïste du *Pari de Pascal*.

¹¹⁰ Ce type d'épistémologie peut être rattaché à une tradition kantienne, avec une différence cependant dans le cas de Donagan. Là où Kant prétend mettre au jour des conditions de possibilité *universelles*, Donagan considère la croyance au libre arbitre comme une condition de possibilité relative à une certaine culture : « When Descartes wrote that 'the will simply consists in our ability to do or not do something (that is, to affirm or deny, to pursue or avoid)' (Descartes, VII, 57), and when Dr Johnson declared to Boswell, 'Sir, the will is free, there's an end on't,' they were expressing something they took their contemporaries to presuppose in choosing what to do. Making that presupposition was part of the culture they inherited, and which is still being transmitted. Inventing arguments for it misrepresents its place in the thinking of those who share that culture. » (Donagan 1987, 180).

¹¹¹ Ce type d'argument en faveur du libre arbitre est à la fois très classique et présenté par van Inwagen dans une épistémologie contemporaine. L'argument moral est présenté, par exemple, par Malebranche lui-même, en renfort du sentiment intérieur : « Quand le sentiment intérieur que nous avons de nous-mêmes ne suffirait pas pour nous convaincre que nous sommes libres, nous pourrions nous en persuader par raison. [...] On voit assez que cette erreur détruit de fond en comble toutes sortes de religions et de morales, la justice de Dieu et celle des hommes, preuve incontestable que nous sommes libres. » (Malebranche 1979, 808–810). Mais la manière dont van Inwagen (1983, 219–221) présente cet argument traditionnel est clairement inspirée de la méta-philosophie Rawlsienne de « l'équilibre réflexif » des intuitions (*reflective equilibrium*).

La troisième et dernière catégorie de philosophes – les héritiers de Spinoza, pour ainsi dire – refuse non seulement que l'introspection puisse être une bonne raison de croire au libre arbitre, mais refuse également qu'il y ait *d'autres* raisons de croire au libre arbitre. Et partant, ces auteurs ne croient pas au libre arbitre. Il leur reste cependant deux options possibles. Soit ils croient avoir des raisons dans le sens inverse – auquel cas ils croiront positivement que nous ne sommes pas libres. Soit ils croient n'avoir pas non plus de raisons de nier le libre arbitre, auquel cas ils restent dans une situation de doute ou d'attente d'informations. On trouve des critiques explicites de l'introspectionnisme chez les premiers comme chez les seconds.

Dans le premier cas, on trouve par exemple Galen Strawson, un auteur qui a tout particulièrement étudié le contenu et l'importance de notre expérience commune :

« Quant aux appels à l'expérience de la liberté ou de la vraie responsabilité – eh bien, nous avons indéniablement une telle expérience [...] Mais [...] le fait que nous ayons, et peut-être ne puissions pas ne pas avoir, une telle expérience ne nous permet pas d'éviter la question de savoir si elle est effectivement véridique. » (Strawson 1986, 48)

Ce que G. Strawson dit ici, ce n'est pas seulement l'existence d'une différence (incontestable) entre la question de la possession d'une expérience, et la question de la véridicité de cette expérience. Ce qu'il dit surtout, c'est que la première question laisse totalement ouverte la seconde question (n'apporte aucun élément de justification pour y répondre) : le fait que nous ayons une expérience d'être libres n'apporte aucune sorte de contribution lorsqu'il s'agit de savoir si nous le sommes vraiment. Pour G. Strawson, la réponse à cette deuxième question reposera uniquement sur son argument « impossibiliste » : les conditions de vérité de notre expérience du libre arbitre sont insatisfaisables, donc nous n'avons pas de libre arbitre.

Saul Smilanski relève également de ce premier cas : non seulement il refuse tout rôle épistémique à l'intuition, mais il considère en outre que cette expérience est positivement illusoire (que son contenu est faux).¹¹² Dans sa réfutation de l'introspectionnisme, Smilansky cite (et approuve) l'argument de Frederick Vivian¹¹³ :

« Il est important de se rendre compte que l'indéterministe affirme bien plus que ne peut justifier un acte d'introspection. Il ne rapporte pas seulement une expérience, un sentiment de liberté, qui peut exister ou ne pas exister, mais il affirme une proposition à propos de cette expérience, à savoir que la volonté n'est pas déterminée causalement, une proposition qui peut être vraie ou fausse. [...] Dire qu'un certain acte de volonté n'est pas déterminé causalement, ce n'est pas faire état d'un donné d'expérience, mais réfléchir sur l'expérience et soutenir une proposition à son sujet. » (Vivian 1964, 95)

Mais le deuxième cas est sans doute plus intéressant : il s'agit de philosophes qui considèrent que le libre arbitre est une capacité que les humains *pourraient* avoir, mais que nous manquons pour l'instant des informations pertinentes pour savoir si les conditions nécessaires sont réunies. Bien sûr, si la question n'est pas tranchée pour eux, c'est que ces conditions nécessaires sont assez exigeantes ; autrement dit, ces auteurs sont généralement des

¹¹² La particularité de la position de Saul Smilanski tient seulement à la *valeur* qu'il attribue à cette illusion : qu'une croyance soit illusoire est normalement une bonne raison de la rejeter... mais Smilansky (2000) défend que cette *illusion* doit être préservée car son abandon aurait des conséquences morales et sociales trop graves.

¹¹³ Vivian lui-même n'est pas un philosophe de premier plan dans le débat, et défend en fait une thèse compatibiliste. Je cite son argument non seulement parce que Smilansky en donne une très longue citation qui occupe le cœur de son propre propos, mais parce qu'il représente, comme nous le verrons, un cas tout à fait typique d'objection infondée à l'introspectionnisme.

incompatibilistes. Il est même fréquent qu'on les considère tout bonnement comme libertariens, parce que (à la différence des compatibilistes) ils adoptent une *conception* de la liberté qui est celle des libertariens, tout en rejetant les objections des incompatibilistes durs. Les deux représentants principaux de cette tendance sont Robert Kane et Mark Balaguer.

« Je rejette l'idée (largement réfutée par John Stuart Mill et d'autres) que nous pouvons connaître que nous avons le libre arbitre par introspection ou par simple spéculation *a priori*. Il y a des aspects empiriques à la question du libre arbitre qui ne peuvent pas faire partie d'une pure spéculation philosophique. » (Kane 1996, 17)

« Tout ce que peut faire un philosophe [...], c'est de proposer des modèles de la manière dont le libre arbitre pourrait fonctionner dans le cerveau, pour que les chercheurs à venir aient une idée de ce qu'ils doivent chercher dans leurs expériences. » (Kane 1996, 184)

« La question métaphysiquement intéressante dans le problème du Libre arbitre se ramène à une question empirique très simple (et absolument pas tranchée) sur l'histoire causale de certains événements neuronaux. » (Balaguer 2010, 1)

Les auteurs « libertariens empiriques » comme Kane et Balaguer adressent aux libertariens anti-introspectionnistes (James, Donagan, van Inwagen) le même genre de critique que ces derniers adressent aux Introspectionnistes. Pour un libertarien empirique, les épistémologies *a priori* (transcendantales, volontaristes, ou fondées sur la moralité) ont fait autant de mal à la *notion* libertarienne du libre arbitre que l'épistémologie introspectionniste ; toutes ces épistémologies ont donné l'impression que cette notion libertarienne était essentiellement associée à une méthodologie philosophique douteuse, qui prend ses libertés avec les sciences empiriques en se permettant de se prononcer *a priori* sur des questions qui relèvent de leur compétence.¹¹⁴

Après ce rapide tour d'horizon de la question introspectionniste, quelques remarques sont nécessaires pour s'orienter dans le débat.

I.3) Comment s'orienter dans le débat : démarche par modèles et démarche par construction

L'impression générale que donne ce tour d'horizon, me semble-t-il, c'est la grande diversité d'interprétations de ce qu'est « l'introspection » du libre arbitre.

Cette diversité porte tout d'abord sur le *contenu* dont l'introspection serait possible ou impossible : un certain nombre d'auteurs (notamment Kane et Balaguer) semblent penser que l'impossibilité d'introspecter est liée à un contenu *incompatibiliste* trop exigeant, mais Mill et van Inwagen, par exemple, critiquent l'introspectibilité de conditions *compatibilistes* (la dépendance volonté-mouvement ; la non-manipulation martienne).

Mais il y a aussi des divergences importantes sur le *type* d'expérience auquel s'intéressent les différents auteurs. Ces divergences se reflètent notamment dans le vocabulaire employé : certains auteurs (Malebranche, Leibniz, Vivian) parlent de « sentiment » (ou « feeling »), ce qui suggère un vécu passif ou réceptif. La plupart des auteurs utilisent le vocabulaire plus neutre et général de « l'expérience », ce qui permet

¹¹⁴ Pour reprendre l'expression de van Inwagen, qui s'adresse à lui-même cette objection avant d'y répondre : « your arguments represent just one more attempt by a philosopher to settle by intellectual intuition and pure reason a question that should be left to empirical science » (van Inwagen 1983, 210).

d'inclure les vécus conscients essentiellement actifs. D'autres enfin (notamment Peacocke, mais voir aussi Donagan) semblent parler d'une expérience d'ordre essentiellement épistémique, soit immédiatement une croyance, soit une impression qui incline à croire.

Enfin, il y a un désaccord sur la *manière* dont l'expérience en question peut ou ne peut pas apporter une justification : certains auteurs (van Inwagen, Vivian, Kane) entendent apparemment le terme « introspection » dans un sens fort où la justification viendrait d'un acte réfléchi d'attention à nos expériences internes ; d'autres auteurs envisagent au contraire la possibilité (ou l'impossibilité) que l'expérience nous apporte une justification simplement en tant qu'expérience *vécue* et non en tant qu'expérience *observée*.

En combinant chacune de ces variables, on aboutit à une variété assez importante dans les *arguments* présentés comme des réfutations de l'introspectionnisme. Ce simple constat de variété n'est pas sans importance dialectique : il contraste avec l'idée reçue, parmi les anti-introspectionnistes, que l'introspectionnisme constituerait une position bien identifiée dont une réfutation évidente et définitive aurait été (presque) unanimement acceptée.

Comment donc s'y retrouver devant une telle diversité ? Je vois principalement deux démarches possibles.

Une première démarche consisterait à partir d'un modèle complet de l'introspection du libre arbitre, c'est-à-dire une théorie prenant parti explicitement sur chacune des questions évoquées plus haut. Par exemple, on pourrait partir de la théorie introspectionniste de Thomas Reid, qui nous donne une description précise du contenu de notre expérience commune (essentiellement, la causalité agentive), du type d'expérience dont il s'agit (une opération de l'esprit), et de la manière dont cette expérience nous justifie à croire au libre arbitre (par simple conscience des opérations de l'esprit, sans qu'une *attention* particulière soit requise). On pourrait également partir de la théorie de Malebranche qui soutient que le contenu de notre expérience, quoique partiellement flou, implique la possibilité d'agir autrement, que cette expérience est un sentiment intérieur, et que, comme tel, elle offre une justification infaillible. A partir de ces théories complètes, on pourrait alors poser deux types de question :

- 1) avons-nous réellement une expérience du type décrit et ayant le contenu décrit ?
- 2) une expérience de ce type ayant ce contenu pourrait-elle apporter une justification en faveur de son contenu ?

On pourrait penser qu'une troisième question devrait apparaître dans le débat, à savoir la question de compatibilité :

- 3) le contenu décrit est-il compatible avec le déterminisme ?

Mais normalement, si le modèle évalué est vraiment « complet », la réponse sera déjà trivialement incluse dans le modèle. On pourrait appeler cette première démarche « démarche modélisation et tests » (modelling and testing), ou « démarche par modèles ».

La démarche inverse consiste à envisager chacun des paramètres séparément et à construire progressivement, à partir de zéro, un modèle censé répondre aux différentes objections. On pourrait l'appeler la « démarche par construction ». C'est cette deuxième démarche que je suivrai. Mon choix repose en partie sur de simples considérations d'ordre d'exposition : la démarche par modèles supposerait de consacrer beaucoup de temps à reconstruire et commenter les détails de telle ou telle solution historiquement importante ; mais comme chacune de ces solutions présente, à mon sens, des idiosyncrasies qui prêtent le flanc à la critique, il me faudrait de toutes les façons proposer un nouveau modèle qui extraie des modèles anciens les éléments intéressants en les isolant des éléments contestables avec lesquels ils ont été accidentellement associés.

Mais il y a une deuxième raison beaucoup plus fondamentale de rejeter la démarche par modèles. Cette démarche part en effet d'un contenu d'expérience déjà analysé (c'est-à-dire

qu'elle part d'un sens précis et particulier de « libre arbitre »). Elle incite donc fortement à ne poser la question de justification que par rapport à un tel contenu analysé. Par exemple, si on part de Reid, la question de justification qu'on se posera sera la suivante : « Mon expérience agentive commune peut-elle me justifier à croire ceci : 'je suis en train de causer tel effet (tel mouvement de bras)' ? ». On ne se posera *pas* en revanche la question suivante :

(J) Mon expérience agentive commune peut-elle me justifier à croire son contenu ?

Ou plutôt, il faudrait dire que la question (J) est ambiguë entre un sens « indexical » et un sens « descriptif » :

(JI) Mon expérience agentive commune peut-elle me justifier à croire que 'ceci est le cas' ?

[et en disant 'ceci', je tourne mon attention vers le contenu non encore analysé de mon expérience agentive.]

(JD) Mon expérience agentive commune peut-elle me justifier à croire que 'je suis en train de causer le mouvement de mon bras' ?

[où 'je suis en train de causer le mouvement de mon bras' est le contenu analysé de mon expérience agentive.]¹¹⁵

On pourrait penser que cette distinction est inutile : puisque 'ceci est le cas' signifie la même chose que 'je suis en train de causer le mouvement de mon bras', la justification à croire l'un et la justification croire l'autre sont une seule et même justification. Il n'y a donc qu'une seule question. Cette supposition apparemment innocente repose sur un principe erroné. En termes techniques, on peut l'appeler le principe de *Transparence de la Justification* :

(TJ) Si p est logiquement équivalent à q (l'analyse conceptuelle révèle que p et q ont les mêmes conditions de vérité), alors S est justifié à croire que p ssi S est justifié à croire que q .

Le principe de transparence de la justification reçoit une infinité de contre-exemples dès qu'on s'aperçoit que l'analyse logique est une activité non triviale sur laquelle un agent parfaitement rationnel peut se tromper. Imaginons par exemple deux auteurs, Hume* et Reid*, qui analysent le concept de « cause » de manière différente ; l'un donne l'analyse Ch, et l'autre l'analyse Cr. Nécessairement, l'une des deux analyses est fausse ; disons que Cr est fausse. L'autre analyse, Ch, pourrait être fausse également, mais supposons qu'elle soit vraie. Considérons maintenant les deux énoncés suivants :

(i) l'impact de la balle a causé le bris de la vitre.

et

(ii) l'impact de la balle a été un Ch par rapport au bris de la vitre.

Il est clair que Reid* peut être justifié à croire que (i) est vraie, mais il ne s'ensuit pas qu'il soit justifié à accepter la proposition (ii) car, de son point de vue, il s'agit de deux propositions tout à fait différentes.

¹¹⁵ Bien évidemment, cette question ne représente qu'un *exemple* de justification descriptive, correspondant à l'analyse de Reid. Il y a autant de formulations de la question (JD) qu'il y a d'analyses différentes.

Pour reprendre un exemple traditionnel, le principe (TJ) correspond au sophisme de l'homme masqué : Platon voit un homme masqué qui (à son insu) est son frère, et qui (manifestement) est assis. On peut dire, si l'on veut, que Platon est justifié à croire que son frère est assis, mais seulement si on veut dire ceci : Platon est justifié à accepter 'cet homme masqué est assis', et par ailleurs cet homme masqué est son frère. Mais il est évident que Platon n'est *pas* justifié à accepter 'mon frère est assis'. Donc après une proposition principale « S est justifié à croire que ... » on ne peut pas remplacer tout bonnement une phrase par une autre phrase ayant les mêmes conditions de vérité. Pour le dire dans les termes de (Quine 1960), le contexte « S est justifié à croire que ... » n'est pas transparent, c'est un contexte opaque.

Pour revenir à l'expérience commune du libre arbitre, il y a donc une différence entre la *justification indexicale* et la *justification descriptive*. Non seulement on peut avoir l'une sans avoir l'autre, mais en outre il n'est pas impossible que l'une constitue un fondement justificationnel pour l'autre. En tout cas, le grand risque de la démarche par modèles, c'est qu'en partant d'une analyse précise du contenu de l'expérience commune, elle invite à poser directement la question de justification descriptive, et donc d'oublier l'apport potentiel que pourrait représenter la justification indexicale. Si l'on suppose que le sens commun n'a pas d'analyse particulière de l'expérience du libre arbitre¹¹⁶, alors la justification indexicale constitue sans doute le cœur de la justification de sens commun pour croire au libre arbitre ; en oubliant la justification indexicale, on risque donc de perdre un élément fondamental de justification qui se situe en amont de toute analyse de notre expérience commune... et en particulier en amont de la distinction entre analyses compatibilistes et analyses incompatibilistes.

Pour toutes ces raisons, j'adopterai une démarche par construction progressive et non une démarche par modèles et objections : j'envisagerai séparément les différents aspects que doit contenir un modèle introspectionniste, en partant de la question de la justification indexicale, qui peut être posée indépendamment de toute théorie du contenu de l'expérience commune. Cela ne m'empêchera pas d'envisager les différents modèles complets qui ont été proposés au cours de l'histoire, mais je les envisagerai chemin faisant, question par question, et non pas chacun individuellement dans sa cohérence globale.

1.4) Comment s'orienter dans le débat : l'introspectionnisme a-compatibiliste

Je viens de dire que la question de la justification indexicale apportée par l'expérience commune pouvait être posée en amont et indépendamment de la question de compatibilité. En d'autres termes, on peut très bien adopter une position fondamentalement introspectionniste (une réponse positive à la question II) sans être libertarien. J'ai déjà mentionné plus haut l'exemple de Peacocke, mais cette thèse est suffisamment inhabituelle pour mériter qu'on s'y attarde un peu. L'introspectionnisme et le libertarianisme ont été si étroitement mêlés historiquement qu'on pourrait facilement croire qu'il y a entre eux un lien logique. Sans doute la communauté philosophique s'est-elle habituée au fait que le libertarianisme n'impliquait pas l'introspectionnisme (grâce à plusieurs générations de grands libertariens anti-introspectionnistes) : on peut donc être libertarien sans être introspectionniste, mais inversement peut-on avoir des raisons d'être introspectionniste si l'on n'est pas libertarien ?

¹¹⁶ C'est-à-dire, si toute analyse descriptive de l'expérience agentive de sens commun suppose une réflexion consciente, spécialisée ou du moins attentive. Cette supposition n'est pas évidente comme nous le verrons. Peut-être est-il impossible d'avoir une expérience agentive sans du même coup être disposé à dire « c'est moi qui l'ai fait », auquel cas une analyse (partielle) en termes d'engagement du soi pourrait être elle-même de sens commun.

Ma réponse est qu'un philosophe compatibiliste (ou un philosophe qui n'a pas d'opinion particulière sur la question de compatibilité) a tout autant de raisons qu'un libertarien d'adopter une position introspectionniste.

Un compatibiliste qui croit au libre arbitre doit bien, lui aussi, avoir une justification ou une autre pour y croire. Et pour lui également, il semble que la justification qui vienne le plus naturellement à l'esprit soit l'expérience que nous faisons d'être libres. Pourquoi trouve-t-on alors beaucoup moins de plaidoyers introspectionnistes chez les auteurs compatibilistes ? La réponse, à mon sens, est la suivante : dans le débat entre compatibilistes et libertariens, ce sont les libertariens qui exigent des conditions métaphysiques supplémentaires pour la liberté ; c'est donc à eux qu'il revient d'avancer des raisons de croire que ces conditions sont satisfaites dans la réalité. Les conditions compatibilistes, quant à elles, sont non seulement faciles à satisfaire, mais en outre il semble qu'elles soient déjà supposées satisfaites par tous les membres de la discussion (le libertarien les accepte généralement comme conditions *nécessaires* de la liberté, et même l'incompatibiliste dur ne doute de la liberté qu'à cause des conditions incompatibilistes exigeantes, pas à cause des conditions compatibilistes). Donc il existe bien, dans l'absolu, une question de justification pour le compatibiliste (qu'est-ce qui nous justifie à croire que les conditions faciles à satisfaire sont effectivement satisfaites ?) ; mais cette question n'a pas beaucoup d'enjeu dialectique dans le débat traditionnel, puisque tout le monde admet que, *d'une manière ou d'une autre*, cette question est résolue. Autrement dit, c'est une question épistémologique qui n'a pas d'enjeu direct sur la question métaphysique de l'existence du libre arbitre. Un compatibiliste n'a donc pas vraiment de raisons de soulever cette question, à moins d'avoir des intérêts proprement épistémologiques. C'est pour cette raison, et pour cette raison uniquement à mon sens, que l'introspectionnisme s'est trouvé associé dans les esprits à l'incompatibilisme libertarien.

Prenons l'exemple des théories compatibilistes dites de la « concordance » (*mesh theories*), héritières de (Frankfurt 1971). Pour ces théories, la condition essentielle du libre arbitre est la concordance de nos désirs, d'après lesquels nous agissons, avec un niveau supérieur de nous-mêmes, soit avec l'image que nous avons profondément de nous-mêmes (*real self view*), soit avec des valeurs, soit avec des désirs de deuxième ordre. Frankfurt adopte cette dernière solution : nous sommes libres si nos actions ont été accomplies d'après des désirs que nous avons par ailleurs le désir d'avoir (et non pas d'après des désirs que nous avons le désir de ne pas avoir – par exemple, le désir irrépressible de manger une troisième plaque de chocolat, un désir dont nous nous serions bien passés). Cette concordance entre les deux niveaux de désirs est précisément, d'après Frankfurt, ce qu'on appelle libre arbitre (*freedom of the will*). Par ailleurs, Frankfurt affirme que cette capacité est effectivement possédée par les êtres humains. Bien qu'il n'avance aucune justification particulière pour cette affirmation générale, on peut facilement imaginer ce qui peut justifier une telle croyance chez un agent donné : les désirs d'un agent, quel que soit leur niveau, font tous partie de sa vie mentale, et il semble donc que la manière normale et simple de les connaître (et de connaître leurs relations de concordance ou discordance) soit par introspection. On voit ainsi que la théorie de Frankfurt (ou les autres théories de la concordance pour les mêmes raisons) amène très naturellement à une théorie de la justification introspectionniste : si cette théorie introspectionniste est absente (ou reste implicite) chez Frankfurt, c'est seulement parce qu'elle ne pose aucun problème particulier.

L'exemple de Frankfurt montre qu'on peut très naturellement avoir un introspectionnisme compatibiliste, c'est-à-dire un introspectionnisme « en aval » d'une théorie compatibiliste. Mais pour ma démarche, il est encore plus important de présenter l'introspectionnisme « a-compatibiliste », c'est-à-dire un introspectionnisme « en amont » de toute réponse quelle qu'elle soit à la question de compatibilité. On peut le définir de la manière suivante :

(Introspectionnisme A-compatibiliste) L'expérience agentive commune d'un agent S peut apporter à S une justification à croire que 'ceci est le cas' ?
[où 'ceci' désigne le contenu non encore analysé de l'expérience agentive de S, sans préjuger du résultat compatibiliste ou incompatibiliste de l'analyse de ce contenu.]

C'est précisément ce que fournit Christopher Peacocke. Bien que Peacocke défende *en dernière analyse* une position compatibiliste, sa défense de l'introspectionnisme ne repose pas sur cette analyse.

Plusieurs introspectionnistes libertariens ont également souligné, de différentes manières, l'indépendance de la justification introspectionniste par rapport à une analyse détaillée du contenu de l'expérience. C'est le cas notamment de Malebranche, lorsqu'il souligne que notre idée de liberté n'est pas analysable pour nous de manière claire et distincte, mais que cela ne l'empêche pas de fournir une justification :

Si dans le temps que nous sentons notre liberté à l'égard d'un bien particulier, nous devons douter que nous soyons libres, à cause que nous n'avons point d'idée claire de notre liberté, il faudra aussi douter de notre douleur et de notre existence dans le temps même que nous sommes malheureux, puisque nous n'avons point d'idée claire ni de notre âme, ni de notre douleur, mais seulement sentiment intérieur.
(Malebranche 1979, 809)¹¹⁷

C'est également la démarche de Richard Taylor qui établit clairement le caractère justifiant de notre expérience commune *avant* d'exposer son analyse incompatibiliste agent-causaliste :

Est-ce que ces données [les croyances qui caractérisent notre expérience du libre arbitre] sont compatibles avec la théorie métaphysique du déterminisme ? Nous ne le savons pas encore. Nous allons essayer de le découvrir. [...] Les philosophes peuvent avoir différents conceptions de ce qu'impliquent exactement de telles données, mais en tout cas il n'y a pas plus de raisons, il y a même en fait beaucoup moins de raisons, de rejeter les données que de rejeter telle théorie philosophique.
(Taylor 1991, 42)

Pour bien comprendre ce type d'introspectionnisme a-compatibiliste, il faut expliquer sur quelle motivation il peut bien reposer, s'il ne repose pas sur une analyse particulière du libre arbitre.

¹¹⁷ Pour Malebranche, cette absence de clarté est une conséquence logique du fait qu'il s'agit d'une connaissance par sentiment intérieur, c'est-à-dire d'une connaissance de notre âme : « Comme nous ne connaissons point notre âme par une idée claire, ainsi que je l'ai expliqué ailleurs, c'est en vain que nous faisons effort pour découvrir ce qui est en nous qui termine l'action [...]. Car enfin nous n'avons point d'idée claire d'aucune modification de notre âme. » (Malebranche 1979, 810). Cette théorie ingénieuse apporte par avance un élément de réponse à une objection classique que nous verrons dans la deuxième partie de cette étude, à savoir que l'on a du mal à voir en quoi les conditions *négatives* libertariennes (ne pas être déterminé) apportent une contribution *positive* à l'idée de contrôle de nos actions.

I.5) Argument Pro : la motivation doxastique en faveur de l'Introspectionnisme

La motivation fondamentale en faveur de l'introspectionnisme a-compatibiliste est donnée clairement par Peacocke dans l'extrait cité plus haut :

« Dans les délibérations pratiques ou théoriques de la vie ordinaire, il semble que plusieurs possibilités nous soient ouvertes. Nous avons ordinairement l'impression subjective que nous sommes libres de choisir entre plusieurs options. [...] Puisqu'il est plausible que ces impressions jouent ordinairement un rôle causal et rationnel dans les actes de pensée par lesquels un individu s'attribue à lui-même la liberté, toute tentative pour relever le Défi de l'Intégration [...] devra en tenir compte, et expliquer leur rôle. » (Peacocke 1999, 336)

Le recours à l'expérience agentive repose donc ici sur deux choses : premièrement sur l'objectif de relever le Défi de l'Intégration, et deuxièmement sur le fait que notre expérience agentive joue un « rôle doxastique » effectif (un rôle causal ou au moins rationnel dans la formation effective de la croyance au libre arbitre). Le lien entre ces deux éléments est le suivant : étant donné le rôle doxastique de notre expérience agentive, on ne peut relever le défi de l'intégration sans accorder à cette expérience un rôle justificationnel. C'est ce raisonnement que j'appellerai la *motivation doxastique* en faveur de l'introspectionnisme (a-compatibiliste).

Mais pour quelle raison le Défi de l'Intégration requiert-il d'accorder un rôle justificationnel à l'expérience agentive ? Le principe qui sous-tend ce raisonnement n'est pas vraiment controversé (sinon dans sa formulation exacte) mais il mérite d'être présenté explicitement. L'idée centrale est exprimée par Pollock et Cruz de la manière suivante :

« Pour être justifié à croire quelque chose, il ne suffit pas d'*avoir* simplement une bonne raison de le croire », il faut en outre « croire la conclusion *sur la base de* ces raisons » (Pollock and Cruz 1999, 35)

Ce réquisit est généralement appelé « relation de fondement » (*basing relation*). Il permet de faire la différence entre une simple « justification propositionnelle » (le fait d'*avoir* une bonne raison de croire) et « justification doxastique » (le fait que la *croyance* elle-même – 'doxa' en grec – soit effectivement fondée sur cette raison).¹¹⁸ On peut donner toutes sortes d'exemples. En voici un inspiré de (Turri 2010) :

Imaginez deux jurés, Mlle Dox et Mlle Prop, qui délibèrent sur le cas de M. Mansour. Les deux jurés ont reçu exactement les mêmes informations pendant le procès, et ces informations donnent d'excellentes raisons de penser que M. Mansour est coupable. Les deux jurés arrivent à la conclusion que M. Mansour est coupable, mais tandis que Mlle Dox forme cette croyance sur la base des bonnes informations reçues, Mlle Prop adopte cette croyance simplement parce que M. Mansour a un visage qui ne lui revient pas.

¹¹⁸ Cette terminologie (justification doxastique vs propositionnelle) vient de (Firth 1978), mais elle est aujourd'hui très largement répandue. Feldman et Conee, quant à eux, réservent le terme de « justification » à la justification propositionnelle, et parlent de « croyance bien fondée » (*well founded belief*) pour la justification doxastique. Une version du Principe qui nous intéresse est d'ailleurs leur principe WF (Feldman and Conee 1985, 24).

Dans ce scénario, Mlle Prop a une justification propositionnelle (elle a de bonnes raisons de croire à la culpabilité de M. Mansour) mais pas de justification doxastique (parce qu'elle ne « se sert pas » de ces bonnes raisons), tandis que Mlle Dox a en outre la justification doxastique (parce qu'elle fonde sa croyance sur les bonnes raisons). Pour dire les choses autrement, les deux jurés ont un fondement justificationnel à leur disposition, mais elles n'en font pas le même usage. Chez Mlle Dox, ce fondement justificationnel constitue également son fondement doxastique (ce sur la base de quoi elle croit) ; Mlle Prop, en revanche, au lieu de fonder sa croyance sur le bon fondement justificationnel, adopte un *autre* fondement doxastique (le visage de M. Mansour) qui n'apporte aucune justification. On peut alors reformuler le principe de Pollock et Cruz de la façon suivante :

(Principe du Bon Fondement)

Pour être (doxastiquement) justifiée, une croyance doit non seulement avoir une justification propositionnelle, mais aussi être effectivement fondée sur cette justification.

Ce principe a un corollaire qui nous intéresse plus directement :

(Principe du Mauvais Fondement)

Si le fondement doxastique d'une croyance C ne constitue pas un fondement justificationnel, alors la croyance C en tant que telle n'est pas justifiée.

[*quand bien même* l'agent aurait par ailleurs un fondement justificationnel (inutilisé) pour le même contenu]

C'est ce corollaire que Peacocke utilise comme prémisse implicite dans son raisonnement : pour relever le Défi de l'Intégration, il ne s'agit donc pas de trouver un fondement justificationnel *possible* ; il faut défendre notre fondement doxastique effectif comme fondement justificationnel.¹¹⁹

La deuxième étape du raisonnement consiste à appliquer ce principe au cas de la croyance commune au libre arbitre en identifiant, factuellement, quel est son fondement doxastique. Comment procéder à cette identification ? Comment identifier la raison sur laquelle notre croyance commune est *effectivement* fondée ou basée ? Une réponse parfaitement précise requerrait une théorie de la « relation de fondement » entre une croyance et son fondement doxastique, mais il y a au moins un point qui fait largement consensus sur cette relation : « il s'agit, dans un sens large, d'une relation causale » (Pollock and Cruz 1999, 35). Par conséquent, la deuxième étape du raisonnement de Peacocke est d'identifier les facteurs causaux à l'origine de notre croyance commune au libre arbitre. Cette étape est explicite : « [les impressions d'être libre] jouent ordinairement un rôle causal et rationnel dans les actes de pensée par lesquels un individu s'attribue à lui-même la liberté ». Autrement dit :

¹¹⁹ J'ai gardé ici une formulation assez traditionnelle de la *basing relation*. Il y a un inconvénient à cela : c'est que ce genre de formulation repose sur un vocabulaire fondationnaliste ou évidentialiste, et semble donc présupposer trop de choses à ce stade de la discussion. En réalité, la distinction entre justification doxastique et justification propositionnelle peut très bien être formulée dans des termes neutres, même si cela donne une formulation plus compliquée. L'idée est alors de considérer la justification propositionnelle non pas comme un justifieur, existant quelque part dans le cerveau ou dans l'esprit et qui serait disponible pour fonder la croyance, mais simplement comme la *disposition* à former la croyance d'une *manière* qui la rendrait justifiée (impeccable) (voir notamment la terminologie de (Audi 1998, 3) qui parle de « situational justification », ou celle de (Korcz 2000, 525) qui distingue entre *justified beliefs* – justification doxastique – et *justifiable beliefs* – justification propositionnelle). Tout épistémologue (externaliste, anti-évidentialiste, etc.) peut accepter la distinction entre une croyance que *p* qui a été effectivement formée d'une manière impeccable (justification doxastique), et une croyance que *p* telle que *S* avait à sa disposition une manière de la former qui l'aurait rendue impeccable (justification propositionnelle). Et tout épistémologue reconnaîtra que, quel que soit l'intérêt philosophique du deuxième statut épistémique, le premier est indispensable pour assurer la *rationalité* de nos croyances.

(Introspectionnisme Doxastique)

L'expérience commune du libre arbitre constitue le fondement doxastique de la croyance commune au libre arbitre.

A partir de ces deux prémisses, on peut formuler l'argument suivant :

- (1) Si le fondement doxastique d'une croyance C ne constitue pas un fondement justificationnel, alors la croyance C en tant que telle n'est pas justifiée.
- (2) Le fondement doxastique de la croyance commune est constitué par notre expérience commune d'être libre.
- (C) Si l'expérience commune d'être libre n'apporte pas de justification épistémique, alors la croyance commune au libre arbitre est injustifiée.

ou en d'autres termes :

(C') Si l'introspectionnisme est faux, alors le Défi de l'Intégration échoue pour la croyance commune au libre arbitre (i.e. cette croyance doit être réputée être une illusion).

Voilà pourquoi Peacocke considère une certaine forme d'introspectionnisme comme condition nécessaire si l'on entend relever d'une manière ou d'une autre le Défi de l'Intégration.

Il est important de remarquer ce que cette conclusion *ne dit pas*. Cette conclusion ne dit pas qu'un échec de l'introspectionnisme amène nécessairement à nier l'existence du libre arbitre (ni même à en douter). Dire que l'expérience commune est illusoire, dans le sens qui nous intéresse, ce n'est pas dire que son contenu est *faux*. Prenons un exemple dans le célèbre film *Matrix*. Au début du film, Neo fait confiance à ses expériences sensibles, et croit sur la base de ces expériences qu'il a un corps humain avec deux bras et deux jambes. Mais en réalité, toutes ses expériences sensibles sont des illusions produites directement par la Matrice dans son cerveau. Lorsque Morpheus lui fait prendre la « pilule rouge », Neo sort de la matrice et découvre qu'il a *effectivement* deux bras et deux jambes. L'expérience qu'il avait auparavant (d'avoir deux bras et deux jambes) était donc une illusion véridique. Quelque temps plus tard, Neo revient volontairement dans la Matrice ; il reçoit alors le même genre d'illusions qu'au début, mais cette fois, il *sait* que ce sont des illusions. Il reçoit en particulier l'expérience sensible d'avoir deux bras et deux jambes, et il sait que cette expérience, comme toutes les autres, est un pur produit de la Matrice, bien qu'il se souvienne par ailleurs que son contenu est véridique.

Dans le cas de l'expérience du libre arbitre, on peut également imaginer un scénario d'illusion véridique. Supposons par exemple qu'on parvienne un jour à établir les choses suivantes : « l'expérience commune du libre arbitre est une pure illusion liée aux variations de pression artérielle ; elle n'apporte donc aucune justification à croire à son contenu C ; mais *par ailleurs* ce contenu a pu être décrit précisément (par analyse conceptuelle), et il se trouve que les progrès en neurosciences nous ont permis de découvrir que ce contenu était, en fait, véridique. » Si ce scénario était le cas, nous nous trouverions à l'égard de l'expérience du libre arbitre exactement dans le même genre de situation que Neo à l'égard de son expérience d'avoir deux bras et deux jambes (lorsqu'il revient dans la matrice) : il s'agirait pour nous d'une expérience dont nous savons qu'elle est parfaitement illusoire bien qu'elle soit véridique. Il est clair que dans un tel scénario, notre perspective de sens commun ne serait pas préservée, mais bel et bien détruite, ... même si la perspective éclairée qu'on lui substitue est accidentellement d'accord avec telle ou telle thèse. Si donc le Défi de l'Intégration consiste à

préserver notre croyance de sens commun (en harmonisant une théorie de son contenu avec une théorie de son mode de connaissance), alors un tel scénario d'illusion véridique met en échec le Défi de l'Intégration. Un tel scénario reste un scénario fondamentalement sceptique.

Pour donner raison au sens commun, il ne suffit pas de sauver ses affirmations de la fausseté, il faut encore sauver ses raisons de la simple illusion. Si donc le sens commun appuie sa croyance au libre arbitre sur l'expérience d'être libre, toute épistémologie anti-introspectionniste sera, dans un sens important, une philosophie sceptique. Telle est la motivation fondamentale en faveur de l'introspectionnisme, et cette motivation est indépendante de toute question de compatibilité (ou même de toute autre question d'analyse du contenu de l'expérience commune).

I.6) Arguments Contra

Face à cet argument *pro*, il me semble qu'on peut distinguer deux grandes familles d'arguments *contra* : d'une part les objections de « l'expérience introuvable », et d'autre part les objections de « l'expérience injustifiable ».

Le meilleur exemple d'une *objection de l'expérience introuvable* est sans doute l'objection de Donagan :

« Malheureusement, la liberté du choix telle qu'on l'entend habituellement est parfois défendue [...] comme quelque chose que l'on peut connaître par expérience. [...] Mais] lorsque j'exerce ma capacité à choisir, je n'expérimente pas mon pouvoir d'exercer cette capacité autrement, comme je pourrais expérimenter mon choix comme plaisant ou difficile. [...] Si l'expérience de pouvoir choisir autrement que je ne choisis est quoi que ce soit d'autre que la conscience de faire un choix et la croyance que j'aurais pu me retenir de le faire, je n'ai aucune idée de ce qu'elle peut être. » (Donagan 1987, 180–182)

Le problème de l'introspectionnisme, selon Donagan, c'est qu'une prise en compte réaliste de nos actions telles que nous les vivons ne permet pas de mettre au jour une « expérience agentive » telle que l'introspectionniste la voudrait. Ce n'est même pas que l'expérience soit incapable d'apporter une justification ; peut-être pourrait-elle justifier si elle était là, mais elle n'y est pas. L'introspectionniste se serait inventé une phénoménologie qu'il ne possède pas en réalité, par simple « réquisit théorique », c'est-à-dire en supposant qu'il *doit bien* l'avoir puisque cette phénoménologie est le meilleur candidat pour fonder l'épistémologie du libre arbitre.

Ce type d'objection, qui a sans doute eu beaucoup d'influence contre toutes sortes de projets introspectionnistes, n'est pas pertinent compte tenu de la démarche que j'adopte dans ce travail. En effet, une telle objection repose fondamentalement sur une certaine théorie précise de l'expérience du libre arbitre (une certaine théorie du type et du contenu de cette expérience), pour défendre que l'expérience telle qu'elle est décrite dans *cette* théorie (une expérience de ce type, avec ce contenu) n'existe pas dans notre phénoménologie réelle. Par exemple, il est clair chez Donagan que l'expérience qu'il cherche (et ne trouve pas) est une expérience dont le contenu est le « pouvoir de choisir autrement » et dont le type doit être plus que simplement épistémique (« autre que [...] la croyance que j'aurais pu me retenir de le faire »). Autrement dit, il suit clairement une démarche par modèle et test. Ses motivations pour cela sont évidentes : il se situe dans une discussion « entre libertariens » et recherche pour ce contenu libertarien la meilleure théorie de la justification ; aussi n'aborde-t-il la question de l'expérience du libre arbitre qu'avec l'espoir (déçu) d'y trouver le contenu

libertarien. Mais en adoptant cette démarche, il passe totalement à côté de la question qui, dans ma démarche, est centrale, à savoir la question de l'expérience *effective* du libre arbitre et de la véridicité de son contenu. Il se pourrait très bien qu'une expérience à contenu libertarien soit introuvable, et que pourtant notre expérience agentive, telle qu'elle est réellement, nous apporte une excellente justification en faveur de son contenu (en l'occurrence non libertarien). Autrement dit, tout ce que dit Donagan est parfaitement compatible avec la thèse que « la croyance au libre arbitre est justifiée par introspection », selon ma terminologie.

Je ferais la même réponse pour toute objection de l'expérience introuvable partant d'une théorie précise du contenu (ou du type) de l'expérience agentive : si on *ne trouve pas* une expérience ayant tel contenu C, alors quel que soit par ailleurs l'intérêt théorique du contenu C, ce contenu n'est pas le concept de libre arbitre qui nous intéresse ici puisque ce qui nous intéresse est le caractère illusoire ou non de l'expérience agentive *réelle*. Et donc l'absence d'une telle expérience n'aura aucun impact sur la plausibilité d'une épistémologie introspectionniste *du libre arbitre*.

Ne pourrait-on pas imaginer une objection de l'expérience introuvable plus radicale ? C'est-à-dire une objection de l'expérience introuvable qui ne reposerait pas sur une certaine théorie préalable de l'expérience recherchée ? Cela ne pourrait vouloir dire qu'une seule chose : qu'il n'y a pas d'expérience agentive *du tout* (ou en tout cas, pas d'expérience agentive à l'origine du conflit intuitif avec le déterminisme). J'ai rejeté cette idée au chapitre 1. Et à partir du moment où *il y a* une expérience agentive à l'origine du conflit intuitif du libre arbitre, il est toujours possible de s'interroger sur le caractère justifiant ou non de *cette* expérience (en faveur de son contenu), et de rejeter comme non pertinente l'inexistence de toutes autres sortes d'expériences.

La famille d'arguments qui nous intéressera vraiment est donc la seconde, à savoir les *objections de l'expérience injustifiable*, ou « injustifiante ». Toutes ces objections consistent à dire que, même s'il y a une expérience du libre arbitre, elle n'apporte pas de justification pour croire à son contenu.

Ces objections s'appuient parfois sur une théorie particulière du contenu de l'expérience. Par exemple, si Kane et Balaguer rejettent l'épistémologie introspectionniste, c'est clairement à cause des conditions incompatibilistes exigeantes, qui ne peuvent (d'après eux) être confirmées ou infirmées qu'empiriquement. Chez William James, l'argument anti-introspectionniste repose non seulement sur une certaine théorie du contenu (la « possibilité de l'alternative »), mais surtout sur des observations indépendantes qui infirment ce contenu dans de nombreux cas (et entraînent par conséquent un soupçon *général* sur la fiabilité de ce type d'expérience). Dans la mesure où elles s'appuient sur une théorie du contenu, les objections de Kane, Balaguer, ou James, ne peuvent pas être pertinentes à ce stade de ma démarche. Nous les rencontrerons de nouveau dans la troisième partie de cette recherche, sous une forme conceptuellement plus détaillée.¹²⁰

Ce qui m'intéresse à ce stade, ce sont les objections à l'introspectionnisme qui ne reposent *pas* sur une certaine théorie du contenu, c'est-à-dire les objections à l'introspectionnisme a-compatibiliste proprement dit. On pourrait s'étonner qu'il existe de telles objections : après tout, si on n'a aucune idée du contenu C, comment peut-on être sûr que notre expérience E ne pourrait jamais apporter aucune justification à croire C ? C'est pourtant quelque chose de ce genre que soutient l'objection traditionnelle, celle de Spinoza et Leibniz : quel que soit le degré de liberté que requiert notre expérience agentive, cette expérience, en tant que simple expérience, peut être projetée (par l'imagination) sur un objet

¹²⁰ Nous retrouverons également, en troisième partie, l'objection de Mill qui à sa manière repose aussi sur une théorie détaillée du contenu, bien qu'en l'occurrence il s'agisse d'un contenu compatibiliste.

qui ne possède même pas ce degré de liberté parce qu'il n'en possède absolument aucun. Plus précisément, il me semble qu'on peut distinguer ici trois formes d'objections.

La première forme est *l'objection de la pure impression*. La meilleure instantiation est l'argument de Vivian cité plus haut. L'idée est la suivante : une expérience que *p* est un événement (ou un état) logiquement distinct du fait que *p* ; une fois qu'on a fait clairement cette distinction entre le fait et l'expérience (ou l'apparence), on se rend compte que l'introspection ne peut nous donner accès qu'à l'expérience du libre arbitre et jamais au *fait* du libre arbitre lui-même.

Les deuxième et troisième formes admettent bien sûr la distinction entre expérience que *p* et fait que *p* ; cependant, elles considèrent que l'origine du problème ne réside pas dans la pure distinction conceptuelle entre expérience et fait, mais dans le fait qu'il est *possible* d'avoir l'une sans l'autre. Selon le sens de « possible » qu'on retiendra, cette intuition peut s'entendre de deux manières.

La deuxième forme entend le possible au sens du possible métaphysique. On peut l'appeler *l'objection de l'impression faillible*. Il me semble que c'est la lecture la plus plausible des arguments de Spinoza et Leibniz : en imaginant une aiguille aimantée, ou une pierre, qui aurait le même genre d'expérience d'être libre que *nous* avons, ces auteurs semblent bien mettre en évidence une certaine *possibilité* de notre expérience sans le fait de la liberté, ... mais ils ne peuvent pas avoir voulu dire que, pour autant que nous puissions le savoir, nous sommes peut-être des aiguilles aimantées ou des pierres. Autrement dit, lorsque Spinoza et Leibniz considèrent leurs scénarios comme *possibles*, ils ne peuvent avoir à l'esprit que la possibilité logique ou métaphysique, pas la possibilité épistémique.

La troisième et dernière forme de l'objection de l'expérience injustifiable part de l'idée que certains scénarios de fausseté sont non seulement possibles métaphysiquement (comme la pierre de Spinoza) mais surtout épistémiquement possibles, dans le sens où nous n'avons pas de justification à les croire faux. C'est sans doute ainsi qu'il convient de lire le scénario, proposé par van Inwagen, des dispositifs martiens indétectables (qui contrôlent nos actions tout en produisant en nous une expérience du libre arbitre). Dire que ces processus sont indétectables sert précisément à dire que ce scénario n'est pas seulement une possibilité *logique* de fausseté de notre expérience, mais une possibilité *épistémique* puisque nous n'avons aucun moyen de détecter si nous sommes ou non dans un tel scénario. Cette objection repose donc sur l'idée que l'impression du libre arbitre contient en elle « trop d'informations », qu'elle est « lourde » d'implications pour lesquelles nous n'avons aucune justification ; c'est pourquoi je l'appellerai *l'objection de l'impression lourde*.

Dans la deuxième partie de ce chapitre, je vais étudier en détail chacune de ces trois versions, en m'efforçant de montrer que chacune correspond à un argument sceptique radical qui renverse la justification des croyances perceptives (et/ou des croyances mathématiques). En tant que tel, cela ne suffit pas à « sauver » l'expérience du libre arbitre ; après tout, il y a des philosophes sceptiques qui accepteront pleinement ce rapprochement, et condamneront donc *à la fois* notre croyance au libre arbitre et nos croyances perceptives (voire nos croyances mathématiques). Mais dans toute cette thèse, je prendrai comme *présupposé* que nos croyances perceptives courantes (ainsi que nos croyances mathématiques) sont *justifiées* et constituent des *connaissances*.¹²¹ C'est-à-dire que j'exclurai par présupposé toute position philosophique sceptique consistant à dire que nos croyances perceptives (ou nos croyances mathématiques) sont injustifiées, ou ne sont pas des connaissances, ou consistant à *douter* que ces croyances soient justifiées ou constituent des connaissances. Cette démarche n'est pas du tout une forme de mépris ou de manque d'intérêt pour les positions sceptiques. Ces positions,

¹²¹ Et que nous sommes également *justifiés* à les tenir pour des connaissances (et à les tenir pour justifiées).

dans leur domaine, sont philosophiquement passionnantes et souvent défendues avec beaucoup d'intelligence. La raison de ma démarche et la suivante : *si* ce genre de position sceptique était la bonne position à adopter en épistémologie, *alors* il est fort probable que la croyance au libre arbitre devrait également être réputée injustifiée et hasardeuse... mais cette conclusion ne pourrait plus être tenue pour une condamnation gravissime de la croyance au libre arbitre en tant que telle car elle se retrouverait dans la même situation qu'à peu près toutes nos croyances. Si un immeuble s'écroule dans un tremblement terre qui a ravagé en même temps la ville entière, la raison n'est pas à chercher dans la fragilité de l'édifice mais seulement dans la gravité du tremblement de terre. Par conséquent, un défenseur du libre arbitre n'a pas à être particulièrement inquiet des arguments anti-libre arbitre qui sont en même temps des arguments sceptiques : non pas que ces arguments soient *faibles* (au contraire, les arguments sceptiques sont sans doute parmi les arguments philosophiques les plus forts), mais tout ce qu'ils peuvent montrer lorsqu'on les fait porter contre le libre arbitre, c'est que la croyance au libre arbitre est « aussi peu » justifiée que nos croyances perceptives ou mathématiques... et cette conclusion est très faible et très peu intéressante dans la discussion sur le libre arbitre lui-même.

Ma thèse repose sur la conviction que tous les principaux arguments anti-libre arbitre ont une portée sceptique. Et il est important, dialectiquement, de mettre au jour cette portée sceptique, faute de quoi on peut facilement manquer la faiblesse de leur conclusion, et être seulement frappé par leur force dialectique (qui est véritable, mais qui est la force du scepticisme lui-même). Les seuls arguments qui seraient vraiment problématiques pour la croyance au libre arbitre en tant que telle, ce sont les arguments qui mettraient en cause cette croyance *tout en évitant le scepticisme*. Pour voir s'il y a ou non de tels arguments, il faut bien partir de l'hypothèse que le scepticisme est faux et voir ce qui s'ensuit. Ce qui s'ensuit, à mon sens, c'est que lorsqu'on prend toutes les mesures « anti-sismiques » nécessaires pour préserver l'édifice perceptif (ou mathématique) du tremblement de terre sceptique, on préserve en même temps et par la même occasion la croyance au libre arbitre (fût-ce dans sa variante libertarienne).

II. Trois interprétations de l'objection anti-introspection

II.1) L'objection de la pure impression

L'objection de la pure impression consiste à refuser à l'expérience commune du libre arbitre tout rôle dans le débat *du simple fait* qu'il s'agit d'une expérience ou d'une impression. C'est donc l'objection la plus radicale puisque les deux objections suivantes (et celles que nous verrons dans les chapitres suivants) reposeront uniquement sur la critique d'une *certaine catégorie* d'expériences, des expériences ayant une certaine caractéristique problématique. Dans sa radicalité, cette critique n'est pas vraiment défendable, et repose sur un état peu avancé de la discussion, mais le fait d'y répondre méthodiquement me permettra de poser des jalons conceptuels fondamentaux pour aborder les objections suivantes.

II.1.a. L'objection et son principe inférentialiste

Il existe deux façons de présenter cette objection ; la présentation la plus fréquente manque totalement sa cible et repose simplement sur une mécompréhension de la position

introspectionniste. Il nous faut pourtant écarter cette première version avant d'en venir à la seconde présentation qui, quant à elle, s'oppose véritablement à l'introspectionnisme (quoique de manière fort peu convaincante).

L'objection de la pure impression se présente le plus souvent comme la dénonciation d'un sophisme ou d'une confusion triviale de la part de l'introspectionniste : ce dernier aurait simplement oublié de distinguer entre l'*expérience* du libre arbitre et le *fait* du libre arbitre. Rappelons ici la formulation de Vivian :

« Il est important de se rendre compte que l'indéterministe affirme bien plus que ne peut justifier un acte d'introspection. Il ne rapporte pas seulement une expérience, un sentiment de liberté, qui peut exister ou ne pas exister, mais il affirme une proposition à propos de cette expérience, à savoir que la volonté n'est pas déterminée causalement, une proposition qui peut être vraie ou fausse. [...] Dire qu'un certain acte de volonté n'est pas déterminé causalement, ce n'est pas faire état d'un donné d'expérience, mais réfléchir sur l'expérience et soutenir une proposition à son sujet. » (Vivian 1964, 95)

L'acte d'introspection, nous rappelle Vivian, est un acte par lequel on accède à un donné d'expérience, ou à quelque chose qui arrive à notre esprit. Par exemple, c'est par introspection que je sais que j'ai une expérience visuelle devant moi. Mais lorsqu'il s'agit du *fait* ainsi représenté par mon expérience, ce n'est évidemment pas par introspection que je pourrai en avoir connaissance. L'idée d'oasis est dans mon esprit et je peux donc la trouver en inspectant à l'intérieur de mon esprit ; l'oasis elle-même, s'il y en a une, n'est pas le genre de chose qui peut tenir dans un esprit.

Notons tout de suite que ce n'est pas seulement l'*indéterministe* mais n'importe quel introspectionniste (fût-il compatibiliste ou agnostique sur la question de compatibilité) qui « affirme bien plus que ne peut justifier un acte d'introspection ». En effet, si on pense au scénario des dispositifs martiens de van Inwagen, il est clair pour tout le monde que ce scénario de manipulation exclut le libre arbitre, quelle que soit l'analyse précise qu'on en donne. Pourtant la négation de ce scénario a des conditions de vérité en dehors de l'esprit, et donc la simple introspection n'a pas accès au *fait* du libre arbitre, quelle que soit l'analyse qu'on en donne.

Une telle confusion entre l'expérience (ou l'impression) que *p* et le fait que *p* paraît si grave qu'on peut se demander si les « introspectionnistes » l'ont jamais réellement faite. Mais n'est-il pas évident qu'ils la font puisqu'ils défendent, justement, l'*introspection* du libre arbitre ? En réalité, la majeure partie du problème vient ici d'une ambiguïté sur le terme « d'introspection ». Lorsque cette ambiguïté est levée, on s'aperçoit que la critique d'une « simple confusion » manque totalement sa cible (bien que certaines formulations de l'introspectionnisme y aient prêté le flanc, comme nous le verrons).

La première chose à faire est donc de lever l'ambiguïté problématique sur le terme « introspectionnisme ». L'introspectionnisme est la thèse selon laquelle notre croyance (au libre arbitre) est justifiée par notre expérience (d'être libres). Mais il y a différentes manières dont une expérience peut fonder une croyance. Par exemple, lorsque j'ai une expérience visuelle d'une table en face de moi, on peut (au moins à première vue) considérer qu'elle justifie deux croyances distinctes :

- d'une part la croyance que « j'ai une expérience visuelle de table en face de moi »
- d'autre part la croyance que « il y a une table en face de moi ».

Si ces justifications apportent en outre une *connaissance*, l'une et l'autre connaissance reposeront alors, ultimement sur mon expérience ou mon impression de table en face de moi,

et on pourra donc, dans un sens large, les appeler toutes les deux des expériences « introspectives ». Dans un sens retreint du mot « introspection », cependant, il est clair que seule la première connaissance sera une « connaissance par introspection », c'est-à-dire une « connaissance *de* mon expérience ». La seconde, qui est une « connaissance *par* expérience » plutôt qu'une connaissance *de* mon expérience, sera alors appelée une connaissance « perceptive ».

La différence entre connaissance « introspective » et connaissance « perceptive » dans les sens précis que je viens de donner, porte d'abord sur le rapport formel entre contenu de l'expérience et contenu de la croyance, et non pas sur la nature du contenu lui-même. Je m'explique : dans le cas de la croyance introspective (au sens étroit), le contenu de la croyance (« j'ai une expérience visuelle de table ») est *différent* du contenu de l'expérience (« il y a une table »). Il est non seulement différent, mais surtout d'un niveau conceptuel ou représentationnel plus complexe. Dans le cas de la croyance perceptive, en revanche, le contenu de la croyance est *le même* que le contenu de l'expérience elle-même (ou en tout cas, il est de même niveau représentationnel).¹²² Cette distinction formelle (même contenu / contenu différent) pourrait être opérée quelle que soit la nature du contenu de l'expérience, que celui-ci ait des conditions de vérité (ou des « vérifacteurs ») situées dans l'esprit, dans un bureau, ou dans le désert du Sahara. Donc pour savoir si une théorie est introspectionniste au sens étroit ou introspectionniste au sens large (et plus proche de la théorie de la perception), la seule question pertinente à poser n'est pas de savoir où se situent les vérifacteurs, mais plutôt la question suivante : « lorsque vous dites que votre expérience que '*p*' vous apporte une justification, voulez-vous dire qu'elle apporte une justification à croire que '*p*', ou qu'elle apporte une justification à croire que '*j'ai une expérience que p*' ».

Revenons à l'introspectionnisme du libre arbitre. Il est évident que l'épistémologie du libre arbitre qu'on appelle classiquement « introspectionnisme » est en fait introspectionniste dans le sens large, et non pas dans le sens étroit : elle fonde la croyance au libre arbitre sur une expérience ayant précisément ce contenu. De fait, il serait peut-être moins égarant de parler de théorie « perceptive » du libre arbitre. O'Connor, par exemple, parle constamment de « perception » de la causalité agentive.¹²³ Si beaucoup d'auteurs ont parlé d'introspection, c'est non seulement à cause de l'importance particulière donnée à l'expérience, mais aussi sans doute parce que les conditions de vérité de cette expérience semblent intuitivement se situer sinon dans notre esprit, du moins dans notre corps – comme par exemple notre pression sanguine ou notre satiété gastrique. Il y aurait donc là une forme de perception (une connaissance *par* expérience et non *de* l'expérience) qui serait fondamentalement tournée vers « l'intérieur » : une « intéro-ception », pourrait-on dire. Je ne pense pas qu'il soit indispensable d'introduire ici une innovation terminologique compliquée, du moment que le point de clarification a été compris : la tradition introspectionniste est en fait une théorie de l'introspection au sens large, c'est-à-dire de la connaissance *par* expérience.

Qu'en est-il de l'objection de Vivian ? Il est évident qu'elle repose sur le sens étroit d'introspection : l'introspection, nous dit Vivian, ne peut que « faire état d'un donné d'expérience », c'est-à-dire que l'introspection (au sens étroit) ne peut délivrer que des croyances de la forme « j'ai l'expérience que *p* ». Dans ce sens étroit, il est incontestable que pour pouvoir introspecter « j'ai le libre arbitre », il faudrait que cette proposition soit une

¹²² Cette précision sert à éviter le problème des croyances perceptives erronées dont l'erreur tient à une mauvaise récupération du contenu de l'expérience. Par exemple, j'ai une expérience de table rouge en face de moi, mais à cause d'une certaine forme de confusion, cette expérience me fait produire la croyance qu'il y a une chaise rouge en face de moi. Dans ce cas, les deux expériences n'ont pas le même contenu, mais elles ont certainement des contenus de même niveau de complexité représentationnelle.

¹²³ "It is natural for the agency theorist to maintain that they involve the perception of the agent-causal relation" (O'Connor 1995, 199). Voir également (O'Connor 2000, 124).

description d'expérience, au même titre que « j'ai un mal de tête terrible » ou « j'ai un air qui me trotte dans la tête ». Autrement dit, il faudrait que le libre arbitre soit lui-même conçu comme *étant* une expérience et non comme *faisant l'objet* d'une expérience. Mais qui irait défendre que le libre arbitre *est* l'expérience d'être libre, c'est-à-dire *se réduit* à cette expérience ? Cela voudrait dire que quiconque a cette expérience est *ipso facto* (et par définition) libre. Ce n'est évidemment pas ce que veulent dire les introspectionnistes. Il est d'ailleurs difficile de donner un sens à une telle théorie : si « être libre » se définit comme « avoir telle ou telle impression », et qu'on ajoute que le *contenu* de cette impression est « d'être libre », on a alors une régression vicieuse dans la définition. Car « être libre » voudra dire « avoir l'impression d'être libre », c'est-à-dire « avoir l'impression d'avoir l'impression d'être libre » etc. Il faut bien, d'une manière ou d'une autre, que l'impression du libre arbitre soit l'impression de quelque *fait*.

Le reproche de « confusion » entre expérience et fait repose donc fondamentalement sur une profonde incompréhension des théories introspectionnistes. Il faut néanmoins admettre que certaines formulations malheureuses de l'introspectionnisme ont prêté le flanc à cette incompréhension. Malebranche, en particulier, dans sa théorie du « sentiment intérieur » de la liberté, au lieu de rapprocher la connaissance du libre arbitre de la connaissance perceptive (connaissance *par* expérience), la rapproche au contraire des connaissances introspectives au sens étroit, c'est-à-dire de la connaissance *de* nos expériences :

Lorsqu'un bien particulier nous est présenté, nous avons sentiment intérieur de liberté à son égard, comme nous en avons de notre plaisir et de notre douleur, lorsque nous sentons. Nous sommes même convaincus de notre liberté par la même raison qui nous convainc de notre existence. [...] Il n'en est pas de même du sentiment intérieur, comme de nos sens extérieurs. (Malebranche 1979, 807)

A mes yeux, ce développement malebranchiste est un accident historique fort malheureux, et s'oppose à la conception la plus naturelle de l'introspectionnisme.¹²⁴

Mais s'agit-il vraiment d'une idiosyncrasie historique ? L'idée malebranchiste d'un accès transparent et infaillible à notre liberté n'est-elle pas, au moins, suggérée par notre expérience elle-même ? C'est ce qu'une analyse phénoménologique de Galen Strawson pourrait nous amener à penser. Strawson – qui, on l'a vu, ne défend pas en fin de compte l'introspectionnisme – donne une description de notre conviction du libre arbitre en des termes tels qu'on pourrait croire à une véritable *identité* entre l'expérience du libre arbitre et le fait du libre arbitre :

Considérez [...] *a*, un agent ordinaire doué d'intelligence, de conscience de soi et de finalité. [...] Il est (à ce qu'il semble) tout à fait clair pour lui qu'il est capable de choisir librement. Mais alors, n'est-il pas certain qu'il *est ipso facto* capable de choisir librement ? Peut-être *a* est-il familier des théories déterministes, et les croit vraies. Mais, se dit-il, quand bien même il serait prêt, dans vingt minutes, à dire (à juste titre) que ce qu'il a fait était déterminé, cela ne change rien au fait que maintenant il sait qu'il est capable, maintenant, de choisir tout à fait librement. C'est (à ce qu'il semble) un fait inéliminable, pour lui, qu'il en est ainsi. [...] Tel est le cœur de la liberté comme 'fait expérientiel'. C'est le fait central de la phénoménologie de la liberté. (Strawson 1986, 60)

¹²⁴ Nous verrons dans la prochaine section, p. 127, que cette idiosyncrasie de Malebranche trouve son origine dans une certaine interprétation de l'infailibilisme cartésien.

Dans le contexte de la réflexion de Strawson, il faut prendre le « *ipso facto* » dans son sens fort : ce qui est suggéré, c'est que l'expérience faite par *a* pourrait être une condition constitutive et même *suffisante* du fait que *a* est libre.¹²⁵ Être libre, ça voudrait dire avoir cette expérience. Nous avons vu que cette conception du libre arbitre comme *étant* une expérience ('sa propre' expérience, pour ainsi dire) posait une difficulté d'intelligibilité considérable... et Galen Strawson (malgré des efforts compliqués pour construire une notion adéquate de « fait expérientiel »)¹²⁶ conclut lui-même que cette théorie est tout bonnement inintelligible.

S'il lui accorde autant d'attention, c'est parce qu'elle représente pour lui « le fait central de la phénoménologie du libre arbitre », c'est-à-dire la seule épistémologie introspectionniste à laquelle notre phénoménologie puisse nous amener. Comme cette théorie est inintelligible, c'est alors l'introspectionnisme lui-même qui se trouve ainsi réfuté. Mais ce « fait central de notre phénoménologie », tel que le décrit Strawson, est tout à fait contestable. Sur quoi repose l'analyse de Strawson ? Principalement sur le caractère *irrésistible* de notre conviction d'être libre, et plus précisément sur le fait que nous garderions cette ferme conviction (et nous nous sentirions justifiés à la garder) même si nous devenions convaincus que son contenu est, quant aux faits objectifs, faux. On pourrait dire que la conviction conférée par l'expérience du libre arbitre est « résiliente » à toute information objective. Voilà ce qui, d'après Strawson, nous amène naturellement à penser que notre expérience est *nécessairement* véridique comme « fait expérientiel », quoi qu'on puisse nous révéler à propos du fonctionnement objectif de notre action.

Une bonne manière de comprendre le raisonnement de Galen Strawson est de le comparer à celui de son père, Peter Strawson, que nous discuterons en détail au chapitre 4 (sec. II.4.). Peter Strawson considère que notre acceptation du libre arbitre (manifestée dans les « attitudes réactives » par lesquelles nous tenons les autres agents pour libres) est « résiliente » à une éventuelle preuve du *déterminisme* : si l'on nous prouvait le déterminisme, nous continuerions à tenir les agents pour libres (en adoptant les « attitudes réactives »). Il en conclut que notre concept de libre arbitre est *compatible* avec le déterminisme, puisqu'il ne serait pas remis en question par une preuve du déterminisme. L'argument de Galen Strawson peut être vu comme une radicalisation de celui-ci : pour Galen Strawson, la conviction que donne notre expérience du libre arbitre est résiliente à *n'importe quelle* information sur les

¹²⁵ Le contexte est celui d'une réflexion sur les théories « Subjectivistes » de la liberté, c'est-à-dire, dans le vocabulaire de Strawson, les théories qui considèrent que la *croyance* au libre arbitre est une condition nécessaire et constitutive de la liberté elle-même. Plus précisément, Strawson cherche ici à comprendre l'idée kantienne selon laquelle « tout être qui ne peut agir autrement que sous l'idée de la liberté est *par cela même*, au point de vue pratique, réellement libre » (Kant 1907, 183). A partir de là Strawson interprète la théorie de Kant comme un cas très particulier de théorie Subjectiviste : la croyance au libre arbitre y serait non seulement une condition nécessaire et constitutive du libre arbitre mais, à cause de son caractère inévitable, elle serait même une condition *suffisante* du libre arbitre. C'est ce que Strawson appelle une « commitment theory ».

¹²⁶ Qu'est-ce que cette notion de « fait expérientiel » ? Il s'agit en fait d'un effort théorique pour faire disparaître la distinction entre expérience que *p* et fait que *p*. Dans certains cas typiques, comme l'illusion d'une oasis, il est impossible de ne pas distinguer l'expérience et le fait : le contenu de notre expérience visuelle est une oasis, mais au sens objectif, ce que nous voyons (et voyons mal), c'est juste le désert devant nous, parce qu'il n'y a pas en fait d'oasis. Mais Strawson prend l'exemple d'une illusion d'optique, celle de la spirale de Frazer (des cercles concentriques qui donnent l'illusion d'une spirale unique), pour montrer que dans ce cas-là nous pourrions être portés à décrire la situation différemment : il n'y a pas seulement une expérience « comme » d'une spirale, c'est vraiment une spirale que nous voyons, et un peintre pourrait même utiliser ce schéma dans le but de nous *faire voir* une spirale. Il y a donc bien là un *fait* de spirale, qui va au-delà de l'expérience *comme de* spirale (... et compatible avec le fait que, sur le papier, ce sont des cercles qui sont tracés). Strawson dresse alors l'analogie avec le cas de l'expérience de la liberté : « Consider the parallel. What is depicted? A spiral. This really is a picture of a spiral. What is experienced? Freedom. This really is experience of freedom. It is not just experience as of freedom. For freedom is what is experienced. » (Strawson 1986, 68). Malheureusement, toute cette discussion ne parvient pas, à mes yeux, à faire oublier un instant la confusion qui la sous-tend... et qu'on ne peut pas, sans injustice, attribuer aux introspectionnistes, sauf exceptions regrettables.

faits extérieurs ; il en conclut que la véridicité de cette expérience est compatible avec n'importe quel fait extérieur, c'est-à-dire que les seules conditions de vérité de cette expérience sont purement expérientielles (pour être libre, il *suffit* d'avoir l'expérience d'être libre).

Le problème de ces deux arguments, c'est que la résilience d'une expérience à la démonstration d'un certain fait ne permet pas du tout de montrer une thèse de compatibilité (entre le contenu de l'expérience et le fait en question). Au chapitre 4, je développerai cette réponse à l'argument de Peter Strawson. En ce qui concerne l'argument de Galen Strawson, si nous avons une démonstration scientifique de l'absence objective de libre arbitre¹²⁷, alors le caractère résilient de notre conviction aurait simplement pour conséquence que nous serions fortement portés à croire deux choses contradictoires, l'une par une expérience irrésistible, l'autre par une preuve objective. Serions-nous alors *justifiés* à croire à notre expérience irrésistible ? On peut imaginer les deux cas de figure. Soit (1) la *justification* donnée par l'expérience du libre arbitre est elle aussi « résiliente » à n'importe quel argument, même scientifique, et dans ce cas nous serions justifiés à rejeter les données des sciences sur la base de notre expérience. Soit, plus probablement, (2) les données scientifiques apporteraient une *meilleure* justification que notre expérience du libre arbitre, et dans ce cas la persistance de la conviction au libre arbitre serait une tare de notre système cognitif, c'est-à-dire la persistance inéliminable d'une croyance injustifiée. Dans les deux cas, c'est bien le même fait objectif que l'expérience *prétend* justifier et que les réfutations indépendantes *prétendent* réfuter. Comme on ne peut pas nier que leurs prétentions portent sur le même fait objectif, cela veut simplement dire qu'une des deux prétentions est fausse. Ou pour le dire autrement, comme il est *indéniable* que l'expérience du libre arbitre a un contenu à prétention objective, *soit* ce contenu peut être réfuté et alors en cas de résilience elle devient une simple illusion (sans justification), *soit* la justification qu'elle apporte est elle-même résiliente, et alors cela veut simplement dire qu'il n'est pas possible d'apporter de justification plus forte qu'elle. Dans les deux cas, ce que l'expérience prétend apporter, c'est une justification à croire un fait objectif.

On peut donc retenir que l'introspectionnisme, dans sa version la plus naturelle et la plus forte, reconnaît parfaitement la différence entre l'expérience du libre arbitre et le fait du libre arbitre. Simplement, il affirme qu'il y a entre les deux le rapport suivant : l'*expérience* du libre arbitre nous apporte une raison de croire que le *fait* du libre arbitre est le cas.

Après ces clarifications, nous pouvons enfin aborder la véritable *objection de la pure impression*. L'objecteur, en effet, ne sera pas satisfait par ces clarifications. Sa réaction sera la suivante : une fois qu'on accepte de distinguer l'expérience et le fait, ne s'ensuit-il pas trivialement qu'établir l'existence de l'expérience ne permet pas d'établir l'occurrence du fait, puisqu'il est maintenant admis que ce sont deux questions différentes ? Ou pour reprendre les termes de G. Strawson : « le fait que nous ayons [...] une telle expérience ne nous permet pas d'éviter la question de savoir si elle est effectivement véridique. » (Strawson 1986, 48). Pour être exact, le fait que deux questions soient distinctes n'implique pas qu'elles soient *indépendantes*, ni d'un point de vue logique (une réponse à l'une pourrait impliquer logiquement une réponse à l'autre), ni *a fortiori* du point de vue de la justification (une réponse à l'une pourrait justifier d'accepter une réponse à l'autre). Mais c'est précisément cette prémisse d'indépendance que l'objecteur suppose : pour lui, une simple impression que *p* ne peut jamais apporter une justification à croire au fait que *p*. On pourrait donc donner, en première approche, la formulation suivante à l'*objection de la pure impression* :

¹²⁷ Non seulement du déterminisme, mais de n'importe quelle autre condition objective qui pourrait être tenue pour incompatible avec le contenu de notre expérience du libre arbitre.

- (PI-1) Une simple impression que p ne peut jamais apporter une justification à croire que p .
 (PI-2) L'introspectionnisme soutient qu'une simple impression d'être libre nous apporte une justification à croire que nous sommes libres.
 (C) Donc l'introspectionnisme est faux.

Cette première formulation n'est pas encore parfaitement satisfaisante, car elle ne nous dit pas très clairement à quoi s'oppose une « pure impression », et par conséquent, elle ne permet pas de comprendre exactement en quoi consiste le principe épistémique sous-jacent (PI-1).

Le principe sous-jacent apparaîtra en réponse à la question suivante : si l'objecteur ne considère pas les « simples impressions » comme permettant à une croyance d'être justifiée, qu'est-ce qui, d'après lui, pourrait jouer ce rôle ? L'objecteur pourrait être tenté de dire : « il nous faut, à côté de la simple expérience que p , des *raisons* de croire que p », mais cette réponse ne sera pas très efficace dialectiquement car l'introspectionniste considèrera évidemment l'expérience du libre arbitre comme une « raison » d'y croire, si l'on entend par là quelque chose qui rend la croyance rationnelle ou fondée en raison. L'objecteur doit donc s'appuyer sur une certaine conception de ce qui peut compter comme une « raison ». Quelle conception ? La réponse est assez évidente lorsqu'on connaît les débats traditionnels en philosophie de la connaissance : l'objecteur s'appuie sur une conception *inférentialiste* des raisons.

Dans la conception inférentialiste des raisons (ou de la justification), avoir une raison de croire que p , c'est être capable de produire un *raisonnement* en faveur de p , c'est-à-dire un ensemble de propositions dont les premières (les prémisses) doivent permettre d'arriver par *inférence* à la conclusion que p (en passant éventuellement par plusieurs étapes). Deux précisions s'imposent ici. Tout d'abord, il faut remarquer que ce ne sont pas les propositions elles-mêmes (disons, q , r , s) qui justifient la *croyance* que p : les propositions q , r , s sont dans un certain rapport logique avec la *proposition* p (elles impliquent p , ou elles rendent p probables, par exemple), mais pour ce qui est de justifier la *croyance* que p , il faut bien que l'agent épistémique appuie cette croyance sur les *croyances* que q , que r et que s . Autrement dit, dans la conception inférentialiste, une raison de croire que p (pour S) est nécessairement une croyance (ou un ensemble de croyances) que S possède. Deuxièmement, pour que les croyances que q , que r , que s justifient la croyance que p , il faut évidemment qu'elles soient elles-mêmes des croyances justifiées : une croyance validement déduite d'une préjugé sera elle aussi un préjugé et non une croyance justifiée.

Que *certaines* raisons soient de nature inférentielle, personne ne le nie : il est incontestable que, dans un grand nombre de cas, nous savons répondre à la question « pourquoi crois-tu que p ? » en avançant d'autres propositions crues par nous. « Pourquoi dis-tu que Roosevelt a été un bon président ? – Parce qu'il a su faire face avec intelligence à une crise économique sans précédent. » Il est même probable que, la plupart du temps dans la vie quotidienne, lorsqu'on est amené à rendre raison d'une croyance, la réponse à apporter est une raison de nature inférentielle. Aussi n'est-il pas surprenant que nous ayons tendance à assimiler tout bonnement le concept de raison avec celui de raison inférentielle. Comme nous allons le voir, cette assimilation pose un grave problème philosophique, mais pour l'instant, ce qui nous intéresse est la manière dont l'objecteur de la pure impression applique l'inférentialisme au cas du libre arbitre :

Argument anti-introspectionniste de la pure impression

(Infer.) Une croyance ne peut être justifiée que par inférence à partir d'autres croyances elles-mêmes justifiées.

(PI-2') L'introspectionnisme soutient que la croyance au libre arbitre est justifiée autrement que par inférence à partir d'autres croyances (elle est justifiée par une « simple impression »).

(C) Donc l'introspectionnisme n'offre pas une bonne théorie de la justification.

L'origine inférentialiste de l'objection de la pure impression est particulièrement visible dans une incompréhension à laquelle elle donne parfois lieu. Certains auteurs, ne considérant même pas la possibilité qu'une justification puisse être autre chose qu'une croyance en une proposition, reconstruisent la position introspectionniste sous la forme d'une mauvaise inférence. C'est le cas, par exemple, de Ledger Wood :

L'argument introspectif ou psychologique. [...] « J'ai l'impression d'être libre, donc, je suis libre », tel est l'argument le plus simple, et peut-être le plus puissant en faveur de la liberté. (Wood 1941, 388)

ou encore de Gregg Caruso dans « Consciousness and Free Will: A Critique of the Argument from Introspection »:

Du fait que j'ai *l'impression* d'être libre, il ne s'ensuit pas nécessairement que je *suis* libre. L'impression pourrait être une illusion. C'est un sophisme de croire que l'on pourrait établir une conclusion métaphysique simplement à partir de la phénoménologie. (Caruso 2010, 220–221)

Cette critique fait évidemment fausse route en présentant l'introspectionnisme comme une théorie qui proposerait un *argument* en faveur du libre arbitre, c'est-à-dire un raisonnement où le libre arbitre s'ensuivrait à partir de certaines croyances de départ. Pour l'introspectionniste, la croyance de *S* au libre arbitre est fondée sur l'expérience *elle-même*, et non pas sur un premier acte par lequel il produirait la croyance « j'ai l'expérience d'être libre », puis une inférence à partir de cette croyance « donc je suis libre ».

Bref, l'introspectionnisme n'est pas *du tout* un inférentialisme, c'est une épistémologie dans laquelle une croyance que *p* peut parfois être justifiée par la « simple impression » que *p* (par opposition à : « un ensemble de *croyances* qui permettent d'inférer que *p* »). Mais c'est précisément cela qui gêne l'objecteur de la pure impression (dans sa version la plus pertinente) : pour un tel objecteur, une bonne théorie de la justification *doit* être inférentialiste.

II.1.b. Portée sceptique de l'inférentialisme

Le problème de l'objection de la pure impression, c'est que le principe inférentialiste sur lequel elle repose est une source bien connue du scepticisme radical, c'est-à-dire du scepticisme qui considère comme injustifiées nos croyances perceptives les plus communes. Elle ne satisfait donc pas au critère anti-sceptique établi au point I.6. de ce chapitre : quand bien même cette objection marcherait contre l'introspectionnisme, elle établirait simplement la conclusion (très faible) que l'introspection du libre arbitre est dans une « aussi mauvaise » posture que toute perception du monde extérieur.

Pourquoi le principe inférentialiste mène-t-il au scepticisme de la perception ? On pourrait donner une première réponse sous la forme d'un argument exactement parallèle à l'Argument de la pure impression, et dirigé cette fois contre une théorie « fondationnaliste » de la justification perceptive. La théorie fondationnaliste de la perception que j'ai à l'esprit soutient simplement ceci : les croyances perceptives (comme « il y a une table en face de moi ») sont justifiées par les impressions perceptives de même contenu (l'impression visuelle que « il y a une table en face de moi »). L'argument serait alors le suivant :

Argument anti-perceptionniste de la pure impression

(Infer.) Une croyance ne peut être justifiée que par inférence à partir d'autres croyances elles-mêmes justifiées.

(PI-3) La théorie fondationnaliste de la perception soutient que les croyances perceptives sont justifiées autrement que par inférence à partir d'autres croyances.

(C) Donc la théorie fondationnaliste de la perception n'offre pas une bonne théorie de la justification perceptive.

Cet argument montre que le principe (Infer.) interdit *une certaine* théorie de la justification perceptive, mais il ne met pas directement en évidence la portée sceptique du principe, car il se pourrait que le fondationnalisme de la perception soit simplement une très mauvaise hypothèse épistémologique et qu'il y ait d'autres théories très simples à disposition.

La meilleure façon d'exhiber la portée sceptique de (Infer.), c'est de recourir à un argument sceptique aussi ancien que le scepticisme même et qui a toujours gardé, au cours des siècles, sa position centrale en épistémologie, à savoir l'*Argument de la régression épistémique*,¹²⁸ également appelé *Trilemme d'Agrippa*.¹²⁹ La « régression » dont il s'agit est une conséquence directe du principe (Infer.) :

- pour qu'une croyance 1 soit justifiée, elle doit être fondée sur une autre croyance 2, mais il faut aussi que cette autre croyance 2 soit elle-même justifiée, autrement 1 ne sera pas justifiée, voyons donc si la croyance 2 peut être justifiée.
- pour que la croyance 2 soit justifiée, elle doit être fondée sur une autre croyance 3, mais la croyance 3 doit elle aussi être justifiée, autrement 2 ne le sera pas, et donc 1 non plus.
- pour que la croyance 3 soit justifiée... etc.

¹²⁸ Il faut noter ici une ambiguïté terminologique : par « argument de la régression épistémique », on entend parfois non pas l'argument sceptique lui-même s'appuyant sur le principe inférentialiste, mais le *modus tollens* de cet argument, qui part de l'anti-scepticisme et conclut qu'il faut rejeter l'inférentialisme. J'appelle le premier « argument *sceptique* de la régression épistémique » et le second « argument *fondationnaliste* de la régression épistémique ». En l'absence de précision, je ferai référence à l'argument sceptique.

¹²⁹ La première occurrence importante d'un argument de la régression se trouve chez Aristote, dans les *Analytiques Premiers*, livres I et II. L'argument est pris par Aristote dans son sens fondationnaliste et non dans son sens sceptique. La première version importante de l'argument dans son sens sceptique est rapportée par Sextus Empiricus dans les *Esquisses Pyrrhoniennes*, livre I, ch.15, n°164-169. On peut la reconstruire sous la forme d'un trilemme, et Diogène Laërce attribue à Agrippa la paternité de ce type d'argument, d'où le nom classique de *Trilemme d'Agrippa*. Le même type de raisonnement a été reconstruit, souvent de manière indépendante, à presque toutes les époques de la philosophie, notamment par Pascal (*De l'esprit géométrique*, 1^{ère} section). On le retrouve au cœur des discussions entre le cercle de Vienne et Popper, sous le nom de « Trilemme de Fries » (voir Barberousse, Kistler, and Ludwig 2000, 10 sqq). La présentation contemporaine, que je suivrai ici, remonte principalement des années 1970, notamment aux travaux de Laurence Bonjour (1978; 1985) et William Alston (1976). Mais la littérature contemporaine sur la question est devenue considérable. Je renvoie le lecteur aux bibliographies données par John Post (pour les débuts de la discussion) et Michael Huemer (pour les développements récents) dans (Dancy, Sosa, and Steup 2010, 32, 450).

Il n'est d'ailleurs pas besoin d'entrer dans une discussion philosophique complexe pour faire apparaître le problème de la régression épistémique : il suffit qu'un interlocuteur malintentionné (par exemple un enfant facétieux) décide de nous reposer indéfiniment la question « pourquoi ? » à chacune des raisons que nous lui apportons.¹³⁰ Mais il ne suffit pas de faire apparaître une régression pour qu'il y ait *argument* sceptique ; après tout, toute régression n'est pas viciieuse. Ce que fait apparaître la régression des raisons, c'est que chaque croyance justifiée doit apparaître au sein d'une *chaîne de raisons* (1, 2, 3, etc.) ; pour qu'il y ait argument sceptique, il faut alors mettre en évidence que les chaînes de raisons ne *peuvent pas* se développer d'une manière qui assurerait la justification de la croyance de départ. Or il y a trois façons, et trois seulement, dont une chaîne peut se développer :

- la chaîne peut s'achever sur une croyance qui ne repose pas sur une autre croyance (exemple: 1 est justifiée par 2, qui est justifiée par 3, qui n'est pas justifiée par une autre croyance).
- la chaîne peut revenir circulairement sur une croyance qui figurait déjà à un autre moment de la chaîne (exemple : 1 est justifiée par 2, qui est justifiée par 3, qui est justifiée par 1)
- la chaîne peut faire appel, à l'infini, à des croyances toujours différentes (exemple : supposons que tout entier naturel n soit le nom d'une croyance, et qu'on ait la relation suivante, pour tout n , la croyance n est justifiée par la croyance $n+1$)

L'argument sceptique est un « trilemme » dans le sens où il repose sur l'affirmation qu'aucune de ces trois options ne permet de délivrer une croyance justifiée.

Argument sceptique de la Régression Epistémique

(Infer.) Une croyance ne peut être justifiée que par inférence à partir d'autres croyances elles-mêmes justifiées (i.e. au sein d'une chaîne de raisons inférentielles).

(2) il n'existe que trois sortes de chaînes de raisons : chaîne terminée, chaîne circulaire, et chaîne infinie.

(3) une chaîne terminée ne peut pas justifier une croyance car sa terminaison n'est pas elle-même justifiée de manière inférentielle (d'après Infer.).

(4) une chaîne circulaire ne peut pas justifier une croyance car la circularité épistémique est viciieuse.

(5) une chaîne infinie ne peut pas justifier nos croyances humaines car nous n'avons pas les moyens d'avoir effectivement de telles chaînes.¹³¹

(C) Aucune croyance (au moins humaine) ne peut être justifiée.

Il n'est pas difficile de voir comment la théorie fondationnaliste de la perception répond à cet argument sceptique : le fondationnaliste ne conteste pas que notre système de croyances contienne des chaînes de raisons (puisqu'il admet normalement que *certaines* croyances sont justifiées de manière inférentielle). Mais il refusera l'idée (exprimée par Infer.) que *chaque*

¹³⁰ Ce scénario élémentaire fait apparaître le problème de la régression épistémique pour la simple raison que notre pratique commune de la question « pourquoi ? » est (au moins typiquement) une demande de justification inférentielle.

¹³¹ On peut avoir d'autres raisons de refuser la justification par une chaîne infinie : on peut penser par exemple que la régression infinie est, dans le cas de la justification, intrinsèquement viciieuse. L'argument pourrait être le suivant : une inférence n'est pas en elle-même un producteur de justification mais seulement un *transmetteur* de justification. En mettant de simples transmetteurs bout à bout à l'infini, on n'arrivera jamais à produire la moindre justification s'il n'y a, par principe, aucun producteur de justification. Imaginons par analogie un système hydraulique constitué de tubes mis bout à bout à l'infini, sans qu'intervienne jamais une fontaine ou une pompe à eau : un tel système ne saurait produire la moindre goutte d'eau.

maillon de la chaîne doive être rattaché à une autre croyance elle-même justifiée inférentiellement. Par conséquent, il pourra refuser la prémisse (3) et accepter la légitimité épistémique des « chaînes terminées », en soutenant que ces chaînes se terminent par des croyances qui ont des raisons, certes, mais des raisons non inférentielles. Autrement dit, la prémisse à incriminer dans l'argument, celle qui contient en germe le scepticisme, c'est la prémisse inférentialiste. Telle est en tout cas la leçon classique de l'argument de la régression. Ne pourrait-on pas refuser les prémisses (4) et (5) ? Certaines solutions controversées au problème de la régression reviennent effectivement à incriminer ces prémisses. J'aborderai ces solutions un peu plus loin. Mais pour des raisons de clarté, il est préférable d'accepter pour l'instant la leçon classique et majoritaire qui incrimine (Infer.), et de considérer la conclusion à en tirer pour l'introspectionnisme. Nous verrons ensuite que les solutions incriminant (4) ou (5) ne changent pas, en définitive, cette conclusion.

La conclusion à tirer pour notre discussion, si le principe (Infer.) est un principe qui mène au scepticisme de la perception, c'est que l'argument de la pure impression, *qu'il marche ou non*, n'est pas un argument très dangereux pour l'introspectionnisme du libre arbitre. En effet, *même si* l'argument de la pure impression marchait, *alors* il montrerait seulement que la croyance introspective au libre arbitre est épistémiquement « aussi faible » que toute croyance perceptive (et à vrai dire que toute croyance humaine absolument) ; et inversement si l'argument ne marche pas pour les croyances en général, c'est que son principe est faux, et on peut alors être certain que ce principe faux ne constitue pas un problème pour l'introspectionnisme. Compte tenu de la « démarche anti-sceptique » que j'adopte, cette conclusion conditionnelle *en elle-même* est dialectiquement suffisante pour montrer que l'argument de la pure impression ne pose pas de problème sérieux à l'introspectionnisme du libre arbitre. C'est-à-dire que d'un point de vue strictement dialectique il n'est même pas besoin d'entrer dans les détails de la *réponse* aux arguments (à l'argument sceptique et à l'argument anti-introspection qui repose sur le même principe) pour répondre à l'objection.

Ce type de conclusion conditionnelle sera donc à mes yeux la principale réponse à chacun des arguments anti-introspectionnistes que je vais aborder. Cependant, je ne m'y restreindrai pas : pour chaque argument anti-introspection, après avoir montré qu'il *doit* y avoir une solution ou une autre (si on veut éviter le scepticisme de la perception), j'envisagerai la nature de la solution possible, et ce pour deux raisons. D'abord pour satisfaire un intérêt intellectuel qui semble légitime en philosophie : la question de savoir *comment* répondre à tel ou tel argument sceptique – comment nos croyances sont effectivement justifiées – dépasse largement l'enjeu de savoir si le scepticisme est vrai ou faux.¹³² Ce qui est vrai dans le cas général est vrai aussi dans le cas particulier de la croyance au libre arbitre : un philosophe qui croirait à l'introspectionnisme avec une certitude absolument tranquille et parfaitement indéboulonnable pourra néanmoins avoir un intérêt théorique pour la question de savoir comment l'introspection en vient à apporter une justification. Deuxièmement, un certain nombre d'objections que j'aborderai au fur et à mesure *supposent* d'avoir adopté une certaine solution aux problèmes précédents, ou au moins une certaine terminologie que la résolution des problèmes précédents permet d'introduire. Entrer dans les solutions particulières permettra donc de poser des jalons théoriques aux discussions suivantes.

Pour l'objection qui nous occupe, comme pour les suivantes, la réponse conditionnelle devrait donc être dialectiquement *suffisante* comme défense de l'introspectionnisme, mais une

¹³² Barry Stroud a souligné de manière particulièrement convaincante que l'empressement à considérer le problème sceptique comme tout bonnement résolu était souvent le meilleur moyen de passer à côté des leçons passionnantes que ce problème peut révéler : « adopting something we take to be an acceptable answer to a philosophical problem might be just what prevents us from learning the lesson that a deeper understanding of the source of the problem could reveal. » (Stroud 1984, xi).

solution du problème circonstanciée devrait permettre d'y voir plus clair. Evidemment, une solution circonstanciée devra aussi prendre certains partis parmi des débats importants en épistémologie ; si un lecteur refuse certains de mes partis pris, je l'invite à s'en tenir à ma défense conditionnelle.

II.1.c. Solution : évidentialisme faible et justification basique

Il existe quatre grandes familles de solutions à l'argument sceptique de la régression : fondationnalisme, externalisme¹³³, cohérentisme, et infinitisme. Dans cette section, je vais montrer que la solution fondationnaliste est à la fois très plausible en elle-même, et surtout qu'elle est la plus adaptée pour pouvoir formuler les problèmes épistémologiques suivants. Les solutions cohérentistes et infinitistes, qu'on peut faire correspondre au refus de (4) et de (5), seront abordées dans un second temps. Mais je me concentrerai tout d'abord sur la discussion entre fondationnalistes, externalistes et sceptiques, qui tourne essentiellement autour du principe inférentialiste lui-même et de ses motivations intuitives.

Plus précisément, je vais montrer que la solution fondationnaliste permet de faire droit à certaines intuitions qui sont au cœur de la motivation à accepter (Infer.) – et sans lesquelles il n'y a même pas le début d'un problème épistémologique – mais qu'elle permet de ressaisir ces intuitions d'une manière suffisamment faible pour éviter la conséquence sceptique. L'intuition épistémologique qui motive l'argument de la pure impression est une intuition internaliste, ou évidentialiste ; mais tandis que l'objecteur formule cette intuition internaliste sous la forme d'un principe (Infer.) qui ruine l'introspectionnisme (... et la perception), je proposerai un principe censé satisfaire le cœur de cette intuition, tout en laissant intacts l'introspectionnisme et la perception. En ce sens, le fondationnalisme sera considéré comme une voie moyenne entre l'internalisme trop austère du sceptique et le déni d'intuition internaliste de l'externaliste. Autrement dit, ma démarche (ici comme pour les objections suivantes) consistera non pas à *éliminer*, mais plutôt à *canaliser* les intuitions épistémologiques de l'objecteur anti-introspection.

Le cœur du débat entre sceptiques, fondationnalistes et externalistes repose sur la thèse internaliste, telle que je l'ai définie en introduction, c'est-à-dire sur le réquisit d'un justifieur interne positif pour qu'une croyance soit impeccable du point de vue internaliste. Comme je l'ai déjà évoqué,¹³⁴ cette thèse est équivalente à la thèse *évidentialiste*, qui exige que toute croyance soit fondée sur des données (*evidence*) ou des raisons, si du moins on entend ces raisons et données en un sens internaliste.¹³⁵ Pour bien voir le problème, il faut repartir de la définition de la « justification » et voir comment elle s'applique en détail à l'argument de la régression.

¹³³ L'externalisme peut aussi, à bon droit, être décrit comme un fondationnalisme ; dans ma classification, je suis l'usage de BonJour (BonJour and Sosa 2003, 13–14) qui utilise le terme « fondationnalisme » pour le fondationnalisme internaliste, et le terme « externalisme » pour le fondationnalisme externaliste.

¹³⁴ Voir l'introduction note 49, et les références qui y sont données.

¹³⁵ Certains auteurs soutiennent qu'on peut avoir une conception *externaliste* de l'*evidence*, et que l'externalisme et l'évidentialisme sont donc compatibles. L'exemple le plus important est Timothy Williamson qui va jusqu'à soutenir que la conception internaliste de l'*evidence* est une pure déformation du concept commun : « alternative theories of evidence distort the concept in the attempt to make evidence something that we can infallibly identify. Characteristically, they interiorize evidence: it becomes one's present experience, one's present degrees of belief, or the like. » (Williamson 2000, 193). Quoi qu'il en soit de l'usage courant du mot *evidence* en anglais, l'usage technique reste internaliste chez la plupart des externalistes eux-mêmes, par exemple chez Bergmann qui choisit d'abandonner l'évidentialisme et son réquisit universel d'*evidence* (Bergmann 2006, 131–132). C'est ce sens de *evidence*, ou cette conception des « raisons », que j'utiliserai dans ma discussion.

La notion de « justification », comme impeccabilité épistémique interne, est essentiellement relative à des normes épistémiques, et plus précisément à des normes internalistes qu'une croyance doit satisfaire : une croyance est justifiée si et seulement si elle n'enfreint aucune des normes internalistes proprement épistémiques (c'est-à-dire non déontologiques). Un désaccord à propos de la justification (ainsi définie) sera donc typiquement un désaccord sur la nature ou la liste des *normes internalistes proprement épistémiques* qui doivent être satisfaites pour qu'une croyance puisse être dite impeccable. C'est bien ce qui se passe dans le cas de l'argument de la régression – si du moins on le lit comme un argument de la régression *de la justification*.¹³⁶ Le cœur du débat est de savoir s'il y a ou non des normes *positives* parmi les normes épistémiques internalistes. Tout le monde admet qu'une croyance, pour être impeccable du point de vue interne, doit satisfaire certaines normes négatives (la non contradiction, la non circularité, l'absence de réfutation etc.) mais ce qui fait débat, c'est de savoir si l'impeccabilité internaliste requiert l'existence dans la sphère internaliste (c'est-à-dire mentale) d'un événement, d'un fait, ou d'un quelque chose qui joue le rôle de justifieur positif. L'intuition *internaliste* ou *évidentialiste* répond qu'il faut un tel justifieur, que pour être justifié il faut nécessairement *avoir* une justification, ou encore que pour croire que *p* de manière justifiée, il faut bien avoir des *raisons*, des données (*evidence*) qui nous fondent à croire que *p*. Il semble en effet que les normes strictement négatives soient largement insuffisantes pour donner une évaluation internaliste de nos croyances car elles peuvent être satisfaites de manière « vide » ou triviale : imaginons par exemple un homme emprisonné, et ayant les yeux bandés, ne recevant aucune information sur ce qu'il se passe à l'extérieur de la prison. Cet homme pourrait produire un ensemble de scénarios sur ce qu'il se passe à l'extérieur ; peut-être certains scénarios seront-ils contradictoires ; peut-être d'autres scénarios seront-ils réfutés d'une autre manière, par exemple parce que la croyance à de tels scénarios engendrerait immanquablement une circularité épistémique. Mais assurément il restera encore plusieurs scénarios, disons {S1, S2, S3}, qui passent tous ces tests négatifs, et qui sont tous aussi plausibles les uns que les autres. Supposons à présent que le prisonnier décide de croire à l'un de ces scénarios, S1. Bien sûr, la croyance en S1, même s'il se trouvait par hasard qu'elle est vraie, ne constituerait pas une connaissance – cela personne ne le nie, car d'un point de vue externaliste elle ne satisfait pas à la norme anti-hasard. Mais il semble que la croyance en S1 est critiquable *même du point de vue strictement internaliste*, c'est-à-dire qu'en regardant simplement ce qui se passe dans l'esprit du prisonnier, on a les moyens de critiquer sa croyance, bien que celle-ci n'ait violé aucune norme négative. Le problème de cette croyance, c'est que le prisonnier n'avait rien qui favorise positivement S1 par rapport à S2 et S3. Il semble donc que pour être impeccable d'un point de vue strictement internaliste, une croyance doit reposer sur quelque chose d'interne qui la favorise positivement ; il ne suffit pas qu'elle satisfasse aux normes négatives comme la non-contradiction ou la non-circularité. Telle est en tout cas l'intuition de base de l'*Evidentialisme* ou de l'*Internalisme*.

Comment cette intuition évidentialiste entre-t-elle en jeu dans le débat sur l'argument de la régression ? Le sceptique, tout d'abord, prend son point de départ dans l'intuition évidentialiste, qu'il formule sous la forme du principe inférentialiste (Infer.), qui requiert, pour toute croyance, une autre croyance servant de base inférentielle. Pour plus de clarté, on pourrait diviser ce principe en deux étapes :

¹³⁶ La même structure d'argument est très souvent appliquée à la notion de *connaissance*, ou à des notions de *justification* différentes de celle que j'emploie ici. Qu'il soit entendu que l'argument que je discute ici s'applique strictement à la notion d'impeccabilité épistémique interne.

(Evid.) Une croyance ne peut être justifiée (ne peut satisfaire à toutes les normes épistémiques internes) que si elle possède un justifieur positif (une raison, une *evidence*).
 (Infer.-E) Le justifieur positif (raison, *evidence*) d'une croyance ne peut être qu'une autre croyance lui servant de base inférentielle.
 donc (Infer.-J) Une croyance ne peut être justifiée que par inférence à partir d'autres croyances elles-mêmes justifiées.

Le sceptique présentera rarement les choses sous la forme d'un tel raisonnement : il considérera peut-être que le principe d'*inférentialisme de l'evidence* est l'expression même de son intuition évidentialiste, et n'en est pas séparable. Il n'en reste pas moins qu'on peut les distinguer logiquement et les discuter séparément. L'externaliste, de son côté, refuse absolument toute forme d'évidentialisme : il accepte des normes négatives internalistes (une croyance doit être non réfutée, *undefeated*), mais il n'y a pas de normes positives internalistes. C'est-à-dire, notamment, qu'une croyance qui serait vraie, qui satisferait aux normes externalistes (anti-hasard, peut-être d'autres normes), et qui satisferait aux normes internalistes négatives, pourrait être réputée *absolument impeccable*, parfaite, optimale, quand bien même elle ne serait positivement étayée par rien au point de vue internaliste. S'il y a une intuition évidentialiste, on peut donc dire que l'externaliste adopte (au moins à première vue) une attitude de déni d'intuition : il n'y a rien à garder dans cette intuition de départ. Enfin le fondationnaliste accepte le principe (Evid.), mais il conteste l'*inférentialisme de l'evidence* : pour lui, une *evidence* ou un justifieur positif peut être autre chose qu'une croyance dont le contenu impliquerait ou rendrait probable le contenu de la conclusion.

Qu'est-ce qui pourrait alors jouer le rôle de justifieur positif ? Il n'y a pas une infinité de solutions compte tenu des données même du problème : comme nous parlons de normes internalistes, le justifieur doit être quelque chose qui appartient à la vie mentale du sujet ; comme il doit justifier une croyance, c'est-à-dire qu'il doit justifier d'adopter p de préférence à $\sim p$, ou même de préférence à une proposition r n'ayant rien à voir avec p ,¹³⁷ il doit donc avoir un contenu propositionnel, ou au moins un certain « potentiel » propositionnel (c'est-à-dire être suffisamment « riche » pour pouvoir être légitimement décrit comme ayant une affinité particulière avec tel contenu propositionnel plutôt que tel autre) ; enfin, il ne doit pas être une croyance.

Le meilleur candidat répondant à tous ces critères est assurément la notion d'expérience (doxastique) ou d'apparence que nous avons définie au chapitre 1, c'est-à-dire une attitude propositionnelle rapportée par les expressions « il semble que p » ou « il apparaît que p », une attitude qui n'est ni la croyance que p , ni une simple disposition à former la croyance que p , et dont les expériences perceptives, les expériences mémorielles, les intuitions ou encore les apparences de conscience introspective sont des cas particuliers. Evidemment, les expériences perceptives jouent un rôle d'archétype pour l'ensemble de la catégorie. Avec leur riche phénoménalité, leur « imagerie sensible » comme dit Plantinga, les expériences perceptives ont le grand avantage d'être des données manifestes de notre phénoménologie, des éléments internalistes dont l'existence ne peut échapper à aucun analyste, et qui sont donc immédiatement disponibles lorsqu'on recherche un justifieur internaliste potentiel pour les croyances perceptives. Mais la catégorie doit être définie de manière plus large si l'on exige un justifieur pour *toutes* les croyances non inférentielles, qui semblent inclure les croyances

¹³⁷ Cette condition semble indispensable pour exclure des solutions totalement *ad hoc*, dont la forme caricaturale sera un « justifieur universel ». Supposons par exemple qu'un « internaliste » dise ceci : chaque sujet épistémique S a une conscience de soi, et cette conscience de soi (cet unique phénomène mental qu'est la conscience de soi) rend justifiée n'importe quelle proposition que S pourrait être amené à croire. On aurait alors une théorie avec un « justifieur » positif pour toute croyance... mais d'une manière qui viole manifestement toute notion intuitive de justification.

mémorielles, les croyances aux axiomes mathématiques tels que $1+1=2$, ou autres vérités *a priori* telles que « aucun chien n'est un ensemble ». On peut reprendre ici le texte déjà cité au chapitre 1 (p.72) et qui montre comment on peut généraliser la notion d'*evidence* expérientielle, en partant des expériences phénoménales pour l'étendre aux expériences doxastiques en général :

On peut commencer par noter qu'il y a plusieurs sortes d'expériences. Premièrement, il y a l'*imagerie sensible*, le genre d'expérience que vous avez tout particulièrement dans la vision, mais aussi dans l'audition, l'odorat, le goût, et aussi le toucher. Pour employer la terminologie de Chisholm, dans ce genre d'expérience nous avons une *apparence* de telle ou telle sorte. L'imagerie sensible joue un rôle considérable dans la perception ; les croyances perceptives sont formées *en réponse* à l'imagerie sensible et *sur la base* de cette imagerie.

Pourtant, ce n'est pas le seul type d'expérience qui accompagne la formation de croyance. [...] la formation de souvenirs *mémoriels* n'est généralement pas accompagnée d'expérience phénoménale, ou alors seulement d'une imagerie sensible fragmentaire, fugitive, indistincte, difficile à observer. Vous vous souvenez que vous êtes allés à une soirée à Novosibirsk ; il y a un peu d'imagerie, certes, cependant, elle est fugitive, indistincte, et telle que lorsque vous essayez de l'observer en y appliquant votre attention, elle disparaît. Mais il y a un autre type d'expérience qui est présente: la croyance que c'était à Novosibirsk (et pas, par exemple, à Cleveland) semble *vraie, acceptable, naturelle* ; elle s'impose à vous ; elle semble d'une certaine manière inévitable (il est difficile de trouver les bons termes). La croyance *a l'air* d'être correcte, acceptable, naturelle. Elle ne vous fait pas le même effet qu'une croyance que vous pensez être fausse. La même chose s'applique à la croyance *a priori*.

(Plantinga 2000, 110–111)

Il est important de noter ici que Plantinga ne cherche pas tant à caractériser de manière purement descriptive, au cas par cas, les phénoménologies distinctes de la croyance mémorielle, *a priori* ou perceptive. Une telle démarche au cas par cas insisterait plutôt sur les phénoménologies propres à chacune. Ce qui intéresse Plantinga c'est d'identifier *un* type de phénoménologie qui est *commune* aux différentes sources de croyances. Pourquoi ? Tout simplement parce que l'enjeu est de trouver le type requis par l'évidentialisme, et que l'évidentialisme est une contrainte épistémique *universelle*. Quelle que soit la formulation exacte de l'évidentialisme, il faudra qu'elle exige de toute croyance « la même chose », le même type d'*evidence*, et donc cette contrainte *doit* être suffisamment générale pour pouvoir s'adapter à différents types de croyances. Il y a donc une motivation forte, pour l'évidentialiste, à avoir une notion très générale d'expérience, ... mais cette généralité est aussi la source des réticences anti-évidentialistes les plus fortes. Je terminerai donc cette discussion du débat « sceptique vs externaliste vs fondationnaliste » par une discussion des difficultés que pose la « généralité » du concept d'expérience.

La grande généralité du concept d'expérience comme « apparence » ou « semblance » pose deux types de problèmes lorsqu'on décide de faire jouer un rôle justificatif à ce concept. Tout d'abord, le concept est si général qu'il peut sembler aveuglément généreux d'accorder la justification à toute croyance accompagnée d'une telle « semblance ». Deuxièmement, le concept est si général qu'il semble non seulement *applicable* à tout type de croyance (cela, c'est ce que nous cherchions), mais en outre trivialement *satisfait* par toute croyance absolument, et par conséquent, il ne peut en aucune manière constituer une « contrainte » significative sur la justification. Autrement dit, un concept si général ne peut jouer de rôle intéressant ni comme condition suffisante ni comme condition nécessaire de la

justification. Le premier problème ne nous concerne pas pour l'instant : pour répondre à l'argument de la régression, le fondationnaliste internaliste (l'évidentialiste) n'a pas besoin d'affirmer que *toutes* les expériences confèrent une justification ; il a seulement besoin de *certaines* justification par expérience pour bloquer la régression des raisons inférentielles. C'est seulement dans les parties suivantes que nous verrons si on peut dire, et en quel sens, que *toute* expérience doxastique confère une justification. Le deuxième problème est plus important car il vise le cœur de l'évidentialisme, l'idée qu'il y a une condition nécessaire de fondement pour l'impeccabilité interne des croyances. L'objection a été notamment formulée par Plantinga, contre l'internalisme de Conee et Feldman. L'argument de Plantinga est le suivant :

Considérez de nouveau la croyance que $2 + 1 = 3$; en plus de l'imagerie sensible il y a aussi quelque chose comme un *sentiment d'attraction* envers cette croyance ; elle a l'air correcte [...] Peut-être faut-il dire qu'il y a une sorte d'inclination, ou d'impulsion, à accepter cette proposition plutôt que d'autres ; à vrai dire, peut-être que l'*evidence* impulsive n'est rien d'autre que la manifestation phénoménale du fait que vous croyez effectivement la proposition en question. [...]

[L'évidentialiste] peut définir l'*evidence* d'une manière très large, en incluant sous cette catégorie ce sentiment d'impulsion à croire, ce sentiment que la proposition est vraie ou correcte. S'il entend l'*evidence* dans ce sens, alors, à mon sens, il a raison d'affirmer que l'*evidence* est une condition nécessaire de [la connaissance].¹³⁸ Mais alors, il a un *autre* problème. [...] Prise en ce sens, la condition évidentialiste n'ajoute rien à la condition de croyance. Vous avez une *evidence* impulsive pour p par le simple fait de croire que p ; p est proportionné à ce genre d'*evidence* seulement si vous croyez que p . Il n'est même pas possible que vous croyiez que p sans avoir d'*evidence* impulsive pour p ; comment pourriez-vous croire que p si p ne vous semble pas être vraie ? (Plantinga 1996, 360)

Autrement dit, il n'y a plus aucun intérêt à exiger d'une croyance qu'elle soit proportionnée à l'*evidence* si, par définition même de ce qui compte comme *evidence*, toute croyance satisfait nécessairement cette condition. Le langage de « l'impulsion à croire » semble assimiler ce type d'*evidence* à quelque chose de fondamentalement dispositionnel : or il est évident qu'on ne peut avoir la croyance que p sans être d'une manière ou d'une autre disposé à croire que p . La condition évidentialiste est donc absolument triviale.

La meilleure réponse à l'argument anti-évidentialiste (et donc externaliste) de Plantinga a été apportée par... Plantinga lui-même, qui a changé d'avis sur l'*evidence* impulsive à partir de *Warranted Christian Belief* (Plantinga 2000), où il l'appelle désormais d'un nom plus noble : « l'expérience doxastique ». Le cœur de l'argument de Plantinga était de dire que toute croyance est nécessairement associée à une *evidence* impulsive. Mais, premièrement, même si cela était vrai, il ne s'ensuivrait pas que la condition de proportionnalité aux *evidences* soit une condition triviale. Il y a en effet deux manières d'avoir une attitude épistémique inadaptée aux *evidences*, l'une (impossible d'après (Plantinga 1996)) serait d'avoir une croyance que p sans *evidence* que p , mais l'autre est d'avoir une *evidence* qui *devrait* m'amener à croire que p , et de ne pas croire que p pour autant. Par exemple :

¹³⁸ Pour Plantinga, c'est évidemment en étant une condition nécessaire *pour la justification* que ce sens d'*evidence* serait une condition nécessaire de la connaissance, car dans le contexte de son argument, tout l'enjeu est de savoir si l'on peut trouver un concept intéressant de justification, *c'est-à-dire* une condition nécessaire à la connaissance et qui ressemble au genre de condition recherché par l'évidentialiste.

Un sceptique pathologique pourrait avoir le même genre d'expériences doxastiques que nous tous, et pourtant être incapable de former les croyances appropriées. [...] Ce genre de réaction est exclu par la rationalité interne. (Plantinga 2000, 111)¹³⁹

La notion de rationalité interne, à laquelle Plantinga reconnaît ici un intérêt épistémologique, n'est rien d'autre que notre notion de justification ; et il reconnaît explicitement, dans le même paragraphe, que cette notion de rationalité interne implique comme condition nécessaire de « former les croyances appropriées en réponse aux expériences doxastiques ». ¹⁴⁰ Ce que Plantinga reconnaît ainsi, c'est que l'évidentialisme internaliste n'est pas une contrainte triviale, puisqu'on peut l'enfreindre au moins par omission (en ne croyant pas ce pour quoi on a de bonnes expériences doxastiques)¹⁴¹.

Mais, deuxièmement, on peut aller plus loin et montrer que l'on peut aussi l'enfreindre par excès (en croyant ce pour quoi nous n'avons pas d'expérience doxastique). L'idée que toute croyance impliquerait nécessairement une *evidence* impulsive correspondante est, me semble-t-il, très liée à la compréhension *dispositionnelle* d'une telle expérience. Mais nous avons vu, au chapitre 1, que les expériences ou apparences telles que les définit Huemer ne sont *pas* des simples dispositions à croire. Son argument est le suivant :

Un individu peut être disposé à croire une proposition pour des raisons qui n'ont rien à voir avec le fait que la proposition semble être vraie. Par exemple, je pourrais être disposé à croire que je vais aller au ciel après ma mort, non pas parce que cela me semble vrai, mais parce que je veux que ce soit vrai. Le fait qu'il apparaisse vrai à quelqu'un que *p* est [...] un état mental distinct qui parfois amène à croire que *p*, et non pas la croyance que *p* elle-même ou l'inclination à le croire. (Huemer 2007, 31)

Cet argument répond à la question finale de Plantinga : « Comment pourriez-vous croire que *p* si *p* ne vous semble pas être vraie ? ». Une façon (peut-être y en a-t-il d'autres ?) serait de croire que *p* par vœu pieux plus ou moins conscient, ou par une décision volontaire de croire, si cela est possible. Cette possibilité révèle deux choses : d'une part la non-trivialité de la contrainte évidentialiste, et d'autre part sa motivation profonde. Si la contrainte évidentialiste n'est pas triviale, même dans sa version « pas de croyance sans expérience doxastique », c'est principalement parce qu'il existe des philosophes (notamment William James) qui défendent une conception *volontariste*. Et partant cela fait apparaître une motivation fondamentale de l'évidentialisme, à savoir s'opposer à toute sorte de volontarisme ; cette opposition est particulièrement sensible dans le débat classique entre James et Clifford. La faiblesse

¹³⁹ Dès (Plantinga 1996, n. 58), Plantinga avait aperçu la possibilité d'avoir une expérience doxastique sans avoir la croyance correspondante : « It is of course possible that in some sense *p* seems true to you, when the fact is you don't believe *p* ». Mais il n'avait pas perçu la pertinence de cette possibilité pour la discussion.

¹⁴⁰ Ceci permet au passage de montrer que l'expérience doxastique ne peut pas être réduite à « la simple manifestation phénoménale du fait que je crois que *p* » : chez le « sceptique pathologique », les expériences doxastiques communes ne *peuvent pas* être la manifestation de croyances communes, puisque celles-ci sont absentes. On pourrait essayer de soutenir que le sceptique pathologique a *la même impression* que nous avons lorsque se manifestent à nous phénoménalement nos croyances communes... mais si tel était le cas, on devrait dire qu'il a 'l'impression de croire les croyances communes', qu'il lui semble qu'il les croit. Mais ce n'est pas du tout la situation décrite au départ : dans la situation décrite au départ, il semble au sceptique qu'il a deux mains, mais il ne lui semble pas qu'il le croit.

¹⁴¹ Bien sûr, il ne s'agit pas de dire que nous devons, pour être rationnels, croire *toutes* les propositions pour lesquelles nous avons une *evidence*. Cela semblerait exiger un nombre de croyances infinies si du moins une croyance donne une raison de croire toutes ses conséquences logiques (connues). L'idée est seulement que *dans certaines circonstances*, celles notamment où se trouve le sceptique, ne pas croire que *p* alors qu'on a l'expérience doxastique que *p* est une réponse inappropriée aux données dont on dispose.

dirimante du volontarisme pour un évidentialiste, ce n'est pas tellement qu'il n'ait pas *d'argument* ou de *prémisses* pour asseoir sa croyance ; le problème est que sa croyance est déterminée par une attitude doxastique logiquement indépendante de la notion de vérité, et fondamentalement « productive » plutôt que « réceptive ». La notion d'expérience doxastique, en revanche, est définie comme « apparence *de vérité* de la proposition considérée », et elle est une expérience fondamentalement réceptive. On peut dire qu'elle est orientée à la réception du vrai.

Pour conclure cette discussion « sceptique vs externaliste vs fondationnaliste », retenons ceci : l'évidentialisme inférentialiste qu'accepte le sceptique est manifestement trop austère, puisqu'il mène à la régression des raisons ; mais le refus complet de tout fondement évidentiel interne (le fondationnalisme externaliste) n'est pas attrayant non plus puisqu'en tant que tel il autorise que nos croyances basiques ne reposent sur aucun élément positif interne qui soit orienté à la vérité. Le fondationnalisme internaliste satisfait, semble-t-il, les *desiderata* les plus importants de cette discussion, grâce à la notion d'expérience doxastique : les expériences doxastiques offrent à la fois un fondement interne orienté à la vérité, pour satisfaire l'intuition évidentialiste, et d'autre part un point d'arrêt à la régression des raisons, qui mène au scepticisme. Face au problème sceptique de la régression épistémique, la meilleure solution semble donc être d'accepter le fondationnalisme :

(Fond.) Toutes nos croyances justifiées

- soit reposent basiquement sur une expérience doxastique
- soit sont inférées à partir de croyances elles-mêmes justifiées, au sein de chaînes de raison remontant ultimement à des expériences doxastiques.

Appliquons à présent ce résultat fondationnaliste au problème qui nous intéresse, la justification de la croyance au libre arbitre, et la réponse à *l'argument de la pure impression*. Nous avons vu que cet argument reposait, en réalité, sur le principe inférentialiste :

(Infer.) Une croyance ne peut être justifiée que par inférence à partir d'autres croyances elles-mêmes justifiées.

Or le fondationnalisme refuse ce principe et lui substitue un principe évidentialiste plus faible :

(Evid.) Une croyance ne peut être justifiée que sur la base de données (*evidence*), ou de raisons.

mais

(Base) Certaines croyances que *p* sont justifiées par une raison non inférentielle, à savoir par une expérience doxastique que *p*.

Cette deuxième proposition fait apparaître directement l'erreur qu'il y avait dans la première formulation de l'argument, celle qui est sans doute la plus proche des intuitions de l'objecteur :

(PI-1) Une simple impression que *p* ne peut jamais apporter une justification à croire que *p*.

(PI-2) L'introspectionnisme soutient qu'une simple impression d'être libre nous apporte une justification à croire que nous sommes libres.

(C) Donc l'introspectionnisme est faux.

La proposition (Base) est la négation directe de la proposition (PI-1) : si l'on veut éviter le scepticisme (sans refuser tout évidentialisme), il faut admettre que certaines au moins de nos « simples impressions », certaines de nos expériences doxastiques, confèrent à leur contenu une justification. S'il y a quelque chose à reprocher à l'introspectionnisme, ça ne peut donc pas être le simple fait de faire reposer la justification de la croyance au libre arbitre sur notre expérience commune d'être libres.

II.1.d. Non pertinence des solutions cohérentiste et infinitiste

Il nous faut à présent envisager les deux dernières solutions possibles à l'argument sceptique de la régression, la solution cohérentiste et la solution infinitiste. Une interprétation simple et traditionnelle de la discussion consiste à dire que ces solutions refusent, respectivement, les prémisses (4) et (5) de l'argument, c'est-à-dire que le cohérentiste accepte les chaînes épistémiques circulaires, et l'infinitiste les chaînes épistémiques infinies. Ces deux solutions semblent initialement remettre en question la ligne argumentative que j'ai suivie jusqu'ici. En effet, s'il est possible d'éviter le scepticisme en refusant les prémisses (4) ou (5), c'est que le principe (Infer.) en lui-même n'est pas clairement un principe sceptique. On pourrait donc utiliser ce principe pour ruiner l'introspectionnisme du libre arbitre sans pour autant ruiner toute croyance perceptive. Et dans ce cas, la démarche anti-sceptique que j'adopte ne fonctionnerait pas. Mais en réalité, je vais montrer que ces deux solutions ne sont pas pertinentes pour notre discussion. Non pas qu'elles ne soient pas des solutions intéressantes au problème de la régression, mais si l'une d'elle est la bonne réponse au problème, *alors* elle oblige à décrire la justification perceptive à l'aide d'une reconstruction théorique qu'on pourrait tout aussi bien utiliser pour reconstruire la position introspectionniste du libre arbitre. Et par ailleurs, les reconstructions théoriques qu'elles imposent sont conceptuellement inadaptées pour aborder (pour *formuler* même) les objections épistémologiques que nous verrons par la suite.

L'infinitisme a été défendu principalement par Peter Klein (1999). L'argument principal de Klein en faveur de l'infinitisme est une version de l'argument de la régression qui assume comme prémisses *l'échec* de toutes les autres possibilités (fondationalisme et cohérentisme) ainsi que la négation du scepticisme : comme les chaînes de raisons ne peuvent ni s'arrêter ni être circulaires, et qu'il faut pourtant qu'il y ait de la justification, c'est donc que nos chaînes de raisons sont infinies. L'infinitisme est donc d'abord et avant tout une thèse structurelle extrêmement générale : l'argument de Peter Klein ne permet pas de donner la moindre idée de la *nature* des raisons présentes dans la chaîne infinie, ni même d'exhiber les premiers maillons. Bien sûr, dans certains cas, exhiber les premiers maillons sera facile : c'est le cas des croyances qui sont, de manière incontestée, inférentielles. Mais pour les cas problématiques, comme par exemple ma croyance que j'ai en ce moment mal aux dents, l'infinitisme ne nous donne pas de recette bien claire pour produire la chaîne infinie qui justifie cette croyance. Ce caractère extrêmement général de la thèse infinitiste n'est pas problématique dans une discussion d'épistémologie générale, mais lorsqu'il s'agit comme ici d'évaluer la rationalité d'une croyance particulière (de faire de « l'épistémologie appliquée »), cela nous laisse sans aucune règle applicable pour évaluer ladite croyance. Bien sûr, l'infinitisme requiert une infinité de raisons en renfort de la croyance au libre arbitre qui nous vient de l'expérience d'être libre... mais comment allons-nous déterminer si une telle chaîne existe ou non si nous n'avons pas la moindre idée de ce à quoi ressemble une telle chaîne dans le cas parallèle de la croyance d'avoir mal aux dents qui nous vient de l'expérience d'avoir mal aux dents ?

Ce raisonnement est à l'origine d'une des objections traditionnelles formulées contre l'infinitisme, notamment par Robert Audi.¹⁴² On peut l'appeler cette objection « anything goes », ou « tout est permis » :

Pour toute proposition, une chaîne infinie peut être imaginée (au moins esquissée) dont on puisse soutenir qu'elle apporte une justification pour la proposition. Ainsi, même pour une proposition qu'on croit être évidemment fausse, il nous serait facile d'imaginer des croyances pour l'étayer.¹⁴³ [...] Par conséquent, aussi ridicule que puisse être une proposition que j'affirme croire, il n'y a aucun moyen de m'attraper avec une croyance que je ne pourrais étayer de la même manière. Dans de telles circonstances, tout est permis. Mais cela ne mène à rien.
(Audi 1998, 189–190)

Klein aborde cette objection mais il en manque le cœur. Klein (1999, 312) commence par reconnaître que, si l'infinitisme est vrai, alors « il y a une justification potentielle pour toute proposition contingente », c'est-à-dire une chaîne de propositions infinies à laquelle un interlocuteur pourrait recourir. Mais il ne s'ensuit pas que « toute proposition contingente soit justifiée », car « l'existence d'une telle chaîne est une condition nécessaire, mais n'est pas suffisante. Ainsi, ce ne sont pas n'importe quelles chaînes dotées des bonnes propriétés structurelles qui rendent les croyances justifiées ». L'infinitisme n'arrive donc pas à la conclusion que toutes les croyances (contingentes) sont justifiées. Soit, mais ce que montre l'argument d'Audi, c'est que l'infinitisme ne nous donne pas les moyens de trancher entre les chaînes justifiantes et celles qui ne le sont pas : mes croyances farfelues ne seront peut-être pas toutes justifiées, mais il n'y aura jamais moyen de « me coincer ». Et en ce sens, l'infinitisme « ne mène à rien » : il ne peut jouer aucun rôle normatif en épistémologie.

Pour pouvoir évaluer une croyance particulière (par exemple la croyance commune au libre arbitre), il faudrait d'abord donner des critères distinctifs entre chaînes infinies justifiantes et chaînes infinies injustifiantes. Et en particulier, il faudrait donner une théorie de la chaîne infinie qui justifie ma croyance d'avoir mal aux dents en ce moment, ou ma croyance qu'il y a une table en face de moi. A première vue, ces croyances sont basiques. L'infinitiste devra donc les reconstruire. Une reconstruction possible serait la suivante : ma croyance qu'il y a une table (B1) est fondée sur ma croyance que *j'ai l'impression visuelle d'une table* (B2intro) et sur ma croyance que *B2intro est fiable* (B2fiab). A son tour, ma croyance B2fiab est fondée sur ma croyance que *j'ai l'expérience doxastique que B2 est fiable* (B3intro) et sur ma croyance que *B3intro est fiable*. Etc. *ad infinitum*. Supposons qu'on admette une telle théorie pour rendre compte de la chaîne infinie qui justifie la croyance perceptive B1. Que s'ensuivrait-il pour l'introspectionnisme du libre arbitre ? Tout simplement que la position introspectionniste pourrait très bien être étayée par une reconstruction infinitiste exactement parallèle. Certes l'infinitisme rejette la construction fondationnaliste de l'introspectionnisme, mais rien n'empêche de reconstruire l'introspectionnisme selon sa reconstruction préférée du fondement perceptif.

Par ailleurs, il est évident que cette reconstruction (et à mon sens n'importe quelle autre reconstruction infinitiste de la perception si elle doit être minimalement plausible) fait

¹⁴² Le même type d'objection est formulée par (Post 1980). C'est la version de Post que Klein discute directement.

¹⁴³ Audi donne l'exemple suivant : « Take the obviously false proposition that I weigh at least 500 pounds. I could back up a belief of this by claiming that if I weigh at least 500.1 pounds, then I weigh at least 500 (which is self-evident), and that I *do* weigh at least 500.1 pounds. I could “defend” *this* by appeal to the propositions that I weigh at least 500.2 pounds, and that if I do, then I weigh at least 500.1. And so forth, until the challenger is exhausted. » (*ibid.*).

crucialement appel aux notions de fiabilité des sources épistémiques, qui est au cœur du fondationnalisme. Autrement dit, dès qu'on voudra préciser la théorie infinitiste de la perception de manière à fournir des critères applicables d'évaluation des croyances, on aura très vite recours à des outils conceptuels essentiellement fondationnalistes. La reconstruction infinitiste de ces outils conceptuels n'ajoutera rien de particulier à leur usage pour évaluer telle ou telle croyance, et il est donc préférable d'utiliser directement ces outils dans leur contexte le plus naturel, c'est-à-dire en acceptant le fondationnalisme.

Pour résumer : si l'on est infinitiste, alors soit tout est permis, et il n'y a plus de problème pour la rationalité de la croyance libertarienne, soit tout n'est *pas* permis, mais les raisons pour lesquelles tout n'est pas permis ont plus à voir avec le fondationnalisme et sa terminologie propre qu'avec l'infinitisme.

La même leçon générale s'applique au cohérentisme, mais pour des raisons légèrement différentes. Pour simplifier, le cohérentisme est confronté au dilemme suivant :

Le cohérentisme traditionnel est inacceptable parce que ce qu'il défend est une forme suffisamment déguisée de pétition de principe ; et les formes apparemment plus plausibles de cohérentisme ne sont que du fondationnalisme déguisé.
(Klein 1999, 297)

Commençons par la première branche du dilemme : ce que Klein appelle ici le « cohérentisme traditionnel », c'est l'interprétation du cohérentisme comme étant, de plein droit, une réponse au problème de la régression, à savoir la réponse qui accepte les chaînes de raisons circulaires. On pourrait dire que c'est l'interprétation du cohérentisme comme « circularisme ». ¹⁴⁴ Le cohérentisme comme simple circularisme pose deux types de problèmes : le problème de la circularité en tant que tel, et le problème de l'isolation.

Le problème de la circularité revient à dire ceci : une croyance ne peut pas se justifier elle-même (antiréflexivité), ni être justifiée par une deuxième croyance qui la justifierait en retour (antisymétrie), et le fait d'allonger la « boucle de justification » de 2 croyances à 3 croyances ou plus ne change rien au problème épistémologique de la circularité ; cela peut éventuellement rendre la circularité *moins visible*, mais fondamentalement la structure justificationnelle circulaire reste toujours une forme de pétition de principe. ¹⁴⁵ Le circularisme revient généralement à refuser la dernière étape et à accepter les justifications circulaires « si les cercles sont suffisamment 'grands' » (Fumerton 1985, 43). Cette position (qui fait une différence entre les formes immédiates et les formes éloignées de circularité) est souvent jugée trivialement inacceptable. Je ne serais pas aussi catégorique : nous verrons en effet au chapitre 3 que certaines formes de circularité épistémique sont particulièrement difficiles à éliminer et amènent de plus en plus d'épistémologues à établir des discriminations entre différentes formes de circularité. Pour anticiper un peu, disons que la forme de circularité qu'il est le plus difficile d'éliminer concerne notre justification à croire à la fiabilité de nos sources perceptives, par exemple la fiabilité de la vue. A première vue, il est très plausible de soutenir que nous avons une justification à croire que notre vue est fiable, par exemple parce que nous connaissons le fonctionnement de la rétine... mais nous ne le connaissons que sur la base d'informations obtenues (au moins en partie) par la vue elle-même. Ce type de circularité (qui est au cœur du « problème du critère » comme nous le verrons) peut amener à douter que

¹⁴⁴ On peut douter que cette interprétation du cohérentisme soit si « traditionnelle » que le dit Klein. Les formes historiques les plus anciennes du cohérentisme se conçoivent assurément comme des alternatives au fondationnalisme, mais font reposer leur solution sur la notion de *cohérence* bien plus que sur l'acceptation d'une quelconque forme de circularité. Voir (Olsson 2012, sec. 1).

¹⁴⁵ Voir par exemple (Plantinga 1993, 74–76).

toute circularité est inacceptable... mais elle repose crucialement sur la question de la fiabilité des sources épistémiques ultimes, autrement dit sur une question essentiellement fondationnaliste. L'objection de la circularité place donc le cohérentisme dans une position comparable à celle de l'infinitiste un peu plus haut : elle ne réfute peut-être pas trivialement toute forme de cohérentisme, mais dès qu'on voudra évaluer si la circularité de *telle* croyance particulière est vicieuse ou acceptable, il faudra recourir à des critères et à une terminologie fondationnaliste et non pas à l'outillage conceptuel du cohérentisme comme tel.

Le deuxième problème auquel est confronté le cohérentisme dans sa version la plus fruste, c'est le problème de l'isolation.¹⁴⁶ Pour le cohérentisme simple, notre système de croyances est constitué, non pas d'un ensemble de « boucles de raisons » indépendantes certes, mais plutôt d'une trame, ou d'un réseau, holiste de liens de justification où toutes les croyances dépendent de toutes les autres (d'une manière ultimement circulaire). Et la notion de justification se réduit alors à une certaine propriété interne de la trame : la cohérence. Le problème, c'est que cette condition peut être satisfaite de manière totalement indépendante des expériences que nous avons du monde extérieur. Ceci est extrêmement contre-intuitif parce qu'il semble qu'une théorie de la justification doive garder des conditions qui visent d'une manière ou d'une autre la vérité :

Puisque la théorie n'assigne aucun rôle essentiel à l'expérience, il y a peu de raison de penser qu'un système de croyances cohérent reflètera adéquatement le monde extérieur.
(Olsson 2012, sec. 1)

Mais c'est aussi contre-intuitif de manière plus directe, parce que les expériences que nous faisons (comme par exemple les expériences sensibles) semblent *évidemment* pertinentes pour déterminer si notre système de croyances est justifié ou non :

La théorie est engagée à refuser à tout état mental, autre que nos croyances, la moindre pertinence pour la justification. C'est contre-intuitif – intuitivement, un individu qui aurait exactement les mêmes croyances que moi, mais des expériences sensibles totalement différentes, serait irrationnel.
(Huemer 2010, 24)

On retrouve en fait l'argument du « prisonnier » que nous avançons p.111 contre l'externalisme : dans les deux cas, le problème est qu'on peut satisfaire les conditions de justification de manière totalement « isolée » de l'expérience du monde extérieur.¹⁴⁷ Dans les faits, la plupart des grands systèmes cohérentistes ont en fait proposé des aménagements au

¹⁴⁶ Comme le rappelle (Olsson 2012, sec. 1), le même type d'objection est parfois désigné sous le nom de « alternative systems objection ». Michael Huemer l'appelle « input objection » dans (Dancy, Sosa, and Steup 2010, 23). L'argument anti-cohérentisme de Plantinga (1993, 80–84) appartient à la même famille. Mais il est important de noter que la version proposée par Plantinga est strictement dirigée contre le cohérentisme de la *connaissance* ; ses contre-exemples ne marcheraient pas contre une théorie cohérentiste de la *justification*.

¹⁴⁷ La proximité entre externalisme et cohérentisme n'est pas surprenante si l'on remarque que la cohérence peut être conçue comme une norme strictement négative sur le système de croyances : dans la mesure où l'externalisme est la thèse qui n'accepte que des normes négatives sur la justification, et que le cohérentisme est la thèse qui n'accepte que la cohérence comme norme de justification, il est clair que le cohérentisme simple peut être conçu comme un type particulier d'externalisme, et se trouve donc confronté au problème général de l'externalisme. Du point de vue du fondationnalisme internaliste, la différence entre externalisme et cohérentisme est principalement une différence d'accent : là où le fondationnaliste est inquiet à l'idée d'abandonner toute norme positive interne (tout ancrage dans l'expérience), l'externaliste essaie de calmer cette inquiétude en insistant « mais n'oubliez pas qu'il y a des normes positives *externes* qui sont très importantes, à savoir la fiabilité », tandis que le (simple) cohérentiste essaie de calmer cette inquiétude en insistant « mais n'oubliez pas qu'il y a des normes *négatives* internes qui sont très contraignantes, à savoir la cohérence ».

cohérentisme simple, pour prendre en compte d'une manière ou d'une autre l'expérience. Mais comme le remarque Klein, ces « formes apparemment plus plausibles de cohérentisme ne sont que du fondationnalisme déguisé », et ma réponse appliquée au problème du libre arbitre sera donc légèrement différente.

L'observation que le cohérentisme subtil est, ultimement, une forme (très) particulière de fondationnalisme, a été faite par de nombreux auteurs.¹⁴⁸ Encore faut-il bien la comprendre. Le plus simple est de partir du cohérentisme de Quine, et de la manière dont il intègre l'expérience pour répondre au problème de l'isolation. Chez Quine (Quine and Ullian 1978), les croyances individuelles sont justifiées en fonction de leur place dans la « trame des croyances » (*web of belief*). C'est la partie proprement cohérentiste de son épistémologie. Mais, ainsi qu'il est bien connu, la théorie de Quine conserve l'idée selon laquelle nos croyances doivent affronter d'une manière ou d'une autre le « tribunal de l'expérience » pour être justifiées. Ce qu'il conteste dans le fondationnalisme classique, c'est seulement le « dogme » que chaque croyance puisse affronter le tribunal de l'expérience *individuellement* :

Nos énoncés sur le monde extérieur affrontent le tribunal de l'expérience sensible, non pas individuellement, mais seulement collectivement. (Quine 1951, 38)

Mis à part ce caractère holiste, on peut dire que dans la théorie de Quine la justification de la trame des croyances (comme totalité) repose sur l'expérience, à peu près dans le même sens où, selon le fondationnaliste classique, la justification de telle croyance repose sur telle expérience. Dans les deux cas, il s'agit d'un apport *positif* et *externe* aux croyances qui permet de répondre au problème de l'isolation. En ce sens, on peut dire que la théorie de Quine reste profondément fondationnaliste, même s'il s'agit d'un fondationnalisme *holiste*. En effet, si on essaye de reformuler la position Quinienne sous la forme d'une réponse à l'argument de la régression, on obtiendra à peu près les étapes suivantes :

- Par quoi est justifiée telle croyance que *p* ?
- Par la trame totale des croyances.
- Mais cette trame de croyances ne serait pas justifiante si elle n'était elle-même justifiée (première étape de régression) ; par quoi donc cette trame de croyances est-elle justifiée ?
- Par l'expérience. (Fin de la régression.)

On voit alors que ce qui permet ici d'éviter la régression sceptique, ce n'est pas une l'acceptation d'une quelconque « boucle de justification », mais bien le refus du principe inférentialiste lui-même : une croyance *peut* être justifiée par autre chose qu'une croyance.¹⁴⁹ Elle peut (et doit) tout d'abord être justifiée par une trame holiste de croyances. Peut-être n'est-ce pas là une entorse très grave à l'inférentialisme qui peut évidemment accepter qu'une croyance soit justifiée par *plusieurs* croyances. Mais surtout ladite trame de croyances peut être justifiée par quelque chose qui n'a rien de doxastique, l'expérience.

¹⁴⁸ Outre la citation de Klein, et (Olsson 2012, sec. 1), voir notamment (Plantinga 1993, 78–79), et les références qu'il donne en note 20 : « [The coherentist] should not be seen as endorsing circular reasoning [...] *He is instead pointing to a condition under which a belief is properly basic.* [...] Seen from the present perspective, therefore, the coherentist reveals her true colors as a nonstandard foundationalist with unusual views about what is properly basic. »

¹⁴⁹ N'en déplaise à son disciple Davidson qui faisait du principe inférentialiste la *caractéristique* même du cohérentisme : « [R]ien ne peut compter comme raison d'une croyance si ce n'est une autre croyance. » (Davidson 1986, 310)

Si on admet que les théories cohérentistes les plus défendables se ramènent, d'une manière ou d'une autre, à une forme complexe de fondationnalisme,¹⁵⁰ quelles conséquences doit-on en tirer pour le problème de la croyance introspective au libre arbitre, et l'évaluation de sa justification ? D'un point de vue général, le fait de redonner à l'expérience un rôle positif de justification rend *possible* que notre expérience du libre arbitre justifie à croire au libre arbitre. Mais évidemment, un Quinien n'est pas tenu à affirmer que *toutes* les expériences justifient à accepter leur contenu. Pour le Quinien, c'est la trame des expériences (TE) qui justifie une certaine trame de croyances (TC). Ce qui se passera donc, normalement, est la chose suivante : certaines expériences de (TE) seront telles que leur contenu sera aussi le contenu d'une des croyances de (TC), mais d'*autres* expériences de (TE) ne seront pas ainsi « validées » par (TC). Comment savoir quelles expériences seront validées et lesquelles ne le seront pas ? La réponse sera fondamentalement holiste : c'est seulement « collectivement » (en fonction de la trame entière des expériences) qu'on pourra déterminer (TC) et donc qu'on pourra déterminer si la croyance correspondante à telle expérience est présente dans (TC). Dès lors, si on veut évaluer la position introspectionniste « traduite » dans le langage Quinien, c'est-à-dire une position selon laquelle l'expérience du libre arbitre est « validée » dans la trame de croyance que permet de justifier notre expérience holiste, on sera reporté à la question suivante : notre trame holiste d'expériences (en tant qu'elle contient l'expérience E) justifie-t-elle une trame holiste de croyances dans laquelle se trouve la croyance que nous avons le libre arbitre ? Il me semble évident que la question, à ce niveau holiste, est absolument inévaluable : dès qu'on voudra l'évaluer, il faudra d'une manière ou d'une autre « atomiser » la question, voir comment *localement* telle information donnée par les sens (la vue, le toucher) pourrait poser problème à la justification de *cette* croyance particulière (que nous sommes libres). Autrement dit : quand bien même le cohérentisme subtil de Quine serait la bonne théorie de la justification en général, dès qu'on voudra évaluer la justification d'une croyance individuelle telle que la croyance au libre arbitre, on sera obligé de recourir à des types de questions et à des outils conceptuels qui sont ceux du fondationnalisme proprement dit, c'est-à-dire du fondationnalisme « atomiste ».

Pour toutes les raisons que je viens d'exposer, le cohérentisme et l'infinitisme n'offrent pas un cadre théorique pertinent pour aborder le problème épistémologique du libre arbitre.¹⁵¹ A l'issue de cette longue discussion sur le fondationnalisme (et ses divers compétiteurs), peut-être est-il profitable d'en résumer la motivation : ma thèse générale est que toutes les objections auxquelles fait face l'introspectionnisme ont pour parallèle exact un argument sceptique radical. Par conséquent *tout* épistémologue non sceptique (qu'il soit fondationnaliste ou non) devra rejeter ces objections. Ce raisonnement général me permet donc de défendre l'introspectionnisme sans avoir à m'engager pour telle ou telle thèse épistémologique, si forte soit-elle. Mais par ailleurs, les meilleures objections à

¹⁵⁰ Voir (Olsson 2012) pour d'autres réponses cohérentistes au problème de l'isolation ; l'essentiel pour mon propos, c'est que chacune d'entre elles revient, d'une manière ou d'une autre, à redonner un rôle positif aux expériences sensibles.

¹⁵¹ Une autre raison repose sur la distinction entre fondationnalisme *justificatif* et fondationnalisme *doxastique*. (Audi 1998, 196) fait remarquer qu'une version intéressante du cohérentisme consiste à soutenir que nos croyances perceptives sont doxastiquement (ou psychologiquement) basiques, mais néanmoins dépendantes d'autres croyances quant à leur justification, ce qui permet d'éviter au moins la circularité doxastique particulièrement désastreuse. De même, un infinitiste n'ira jamais soutenir que les croyances en nombre infini qui justifient à croire que *p* ont été d'une manière ou d'une autre engagées psychologiquement dans la dérivation de la croyance que *p* : elles fondent donc *justificativement* cette croyance sans la fonder doxastiquement. Si le cohérentisme et l'infinitisme reviennent à séparer le fondement justificatif et le fondement doxastique des croyances, alors ils reviennent à refuser la contrainte de « Bon Fondement » étudié p.94 et la « motivation doxastique » qui est au principe de toute cette enquête sur la justification introspectionniste.

l'introspectionnisme seront désormais formulées dans un contexte fondationaliste. La raison en est que, si on accepte une autre épistémologie (non-sceptique), soit ces objections disparaissent totalement¹⁵², soit elles devront être reformulées en utilisant une traduction (par exemple une traduction cohérentiste) à l'aide d'un outillage conceptuel qui, ultimement, est un outillage fondationaliste. Dans le premier cas, l'introspectionnisme se trouvera trivialement défendu contre l'objection, puisqu'elle disparaît ; dans le second, le lecteur n'aura qu'à reformuler ma réponse, en même temps que l'objection, selon sa traduction (cohérentiste ou infinitiste) préférée.

C'est donc à titre méthodologique ou conditionnel que j'accepterai désormais la thèse fondationaliste proprement dite, à savoir :

(Base) Certaines croyances que *p* sont justifiées par une raison non inférentielle, à savoir par une expérience doxastique que *p*.

Cette thèse permet de dire qu'il est *possible* que l'expérience doxastique d'être libre confère une justification positive à croire au libre arbitre. Elle suffit donc à contrecarrer l'objection reprochant à l'introspectionnisme de reposer sur une « simple expérience ». Mais les objections que nous allons voir à présent reprochent à l'introspectionnisme de reposer sur un *certain type*, ou certaine catégorie problématique, d'expériences. Pour répondre à ces questions le seul principe fondationaliste sera insuffisant.

II.2) L'objection de l'impression faillible

Les deux objections que je vais considérer partent toutes les deux de l'intuition suivante : l'expérience du libre arbitre échoue à fournir une justification pour son contenu non pas simplement parce qu'elle n'est qu'une expérience, mais parce qu'elle *pourrait* être fausse. Et cette *possibilité* d'erreur est traditionnellement mise en évidence par des scénarios (possibles) dans lesquels l'expérience est présente sans que son contenu soit vrai : scénario de la pierre, de l'aiguille aimantée, des dispositifs martiens implantés dans le cerveau.

La distinction fondamentale entre les deux objections tient à l'interprétation qu'on donne de la « possibilité » dans l'argument. Cette notion peut renvoyer soit à la modalité épistémique, soit à la modalité métaphysique (ou logique dans un sens large). Au sens épistémique, l'objection consiste à dire que les scénarios d'erreur pourraient être le cas « pour autant que nous puissions le savoir », c'est-à-dire qu'ils sont compatibles avec tous les faits auxquels nous avons directement accès, ou compatibles avec toutes nos expériences. Au sens métaphysique, l'objection consiste à dire que les scénarios d'erreur n'impliquent aucune contradiction ou impossibilité métaphysique. Ou encore, que les scénarios d'erreur ne sont pas logiquement exclus... ce qui n'empêche pas qu'ils puissent être exclus par ce que nous savons par ailleurs, ou par nos expériences. Il est clair que les scénarios de la pierre ou de l'aiguille aimantée visent à établir la simple possibilité métaphysique,¹⁵³ tandis que le scénario des dispositifs martiens est construit pour être un scénario *épistémiquement* possible.

¹⁵² Par exemple, un grand nombre d'objections envisagées dans cette étude disparaissent totalement si on accepte l'externalisme.

¹⁵³ Spinoza ne *peut* pas avoir voulu dire que, pour autant que nous puissions le savoir, ou pour autant que nos expériences permettent de le dire, nous sommes peut-être des pierres ayant l'expérience de librement tomber vers le bas... tout simplement parce que nous ne faisons même pas cette expérience ! Spinoza veut évidemment dire ceci : nous pourrions être *comme* des pierres qui font l'expérience illusoire de tomber librement vers le bas, lorsque nous faisons l'expérience de faire librement *telle autre chose*. (Certes, nous faisons parfois l'expérience de tomber comme des pierres, mais dans ces cas-là, notre expérience n'est pas l'expérience de le faire librement).

Dans cette section, je mettrai en évidence le caractère sceptique, et donc non problématique pour cette étude, de l'objection infailibiliste (celle reposant sur la possibilité métaphysique d'erreur), et j'appliquerai ensuite au cas du libre arbitre la solution anti-sceptique la plus reçue pour pouvoir passer aux objections suivantes.

II.2.a. L'objection et son principe infailibiliste

Les objections classiques de Spinoza et Leibniz contre l'introspectionnisme cartésien peuvent être lues de deux manières différentes, l'une qui repose essentiellement sur les scénarios de la pierre ou de l'aiguille, l'autre où les scénarios ne jouent qu'un rôle purement illustratif et non argumentatif. Reprenons la formulation de Leibniz :

« La raison que M. Descartes a alléguée, pour prouver l'indépendance de nos actions libres par un prétendu sentiment vif interne n'a point de force. [...] C'est comme si l'aiguille aimantée prenait plaisir de se tourner vers le nord ; car elle croirait tourner indépendamment de quelque autre cause, ne s'apercevant pas des mouvements insensibles de la matière magnétique. » (Leibniz 1969, 131–132)

Selon une première lecture, l'essentiel de l'argument est d'établir la distinction conceptuelle entre l'expérience positive d'être libre et l'absence d'expérience d'être causé : l'être humain serait dans le second cas, non dans le premier, et n'aurait donc rien pour justifier positivement sa croyance d'être libre. Deux remarques sur cette lecture de l'argument. Premièrement, l'argument ainsi compris est un argument de l'expérience inexistante qui repose sur une certaine théorie phénoménologique du type d'expériences que nous avons effectivement ; c'est donc dans la deuxième partie que j'étudierai la légitimité de sa prémisse phénoménologique. En amont de l'étude phénoménologique, je m'en tiens à ma réponse générale aux arguments de l'expérience inexistante : ces arguments peuvent peut-être défendre que nous n'avons pas une expérience d'être libre *d'un certain type*, mais il est incontestable que nous avons au minimum une expérience *doxastique* commune d'être *positivement* libre. Et les scénarios de Leibniz et Spinoza ne remettent pas *cela* en question : il nous *semble vrai* que nous sommes positivement libres, quand bien même cette expérience doxastique reposerait sur une *absence* d'expérience *sensible* d'un certain type. Deuxièmement, l'argument ainsi compris ne repose aucunement sur la possibilité qu'il y ait d'*autres* êtres (la pierre, l'aiguille) ayant une expérience négative sans expérience positive : il suffit que ce soit *notre* expérience. Les scénarios de l'aiguille ou de la pierre n'ont alors aucun autre rôle que d'illustrer ce que signifie la distinction conceptuelle entre expérience positive et expérience négative. Ils jouent plus un rôle « psychologique » qu'un rôle argumentatif.

Mais selon une autre lecture, les scénarios de l'aiguille ou de la pierre servent à exhiber la *possibilité* que l'expérience du libre arbitre soit illusoire. Dans cette autre lecture, il ne s'agit plus de dire que nous n'avons pas l'expérience du libre arbitre : nous *avons* une certaine forme d'expérience d'être libres (fût-ce une expérience doxastique) mais c'est une expérience qui est telle qu'elle pourrait très bien être possédée par la pierre ou l'aiguille, des êtres fondamentalement non libres. On pourrait formaliser l'argument de la manière suivante :

Argument de la pierre

- (1) Il est métaphysiquement possible qu'une pierre, bien qu'absolument déterminée dans sa chute, ait l'expérience doxastique de tomber librement.
- (2) Donc, il est métaphysiquement possible que l'expérience doxastique d'être libre soit non véridique.¹⁵⁴
- (3) S'il est métaphysiquement possible qu'une expérience doxastique soit non véridique, alors elle ne confère pas de justification épistémique à son contenu.
- (C) L'expérience E d'avoir le libre arbitre n'apporte aucune justification à croire au libre arbitre.

Dans cet argument, le scénario de la pierre est en somme une expérience de pensée permettant d'établir une prémisse de possibilité métaphysique (la proposition 2). C'est la prémisse 3 qui correspond ensuite au principe épistémologique central de l'argument.

Cet argument est-il convaincant ? La réponse est clairement non, pour la simple raison que le principe épistémologique invoqué en (3) est un principe extrêmement radical que fort peu de philosophes accepteraient aujourd'hui. Il s'agit en effet d'une certaine version (fondationnaliste) de l'*Infailibilisme de la Justification*.¹⁵⁵ Si une expérience est faillible, elle ne peut pas conférer de la justification. Elle ne peut pas faire partie de ces expériences qui, pour le fondationnaliste, se trouvent au fondement de l'architecture des raisons. Inversement, cela veut dire que cette architecture des raisons devra reposer uniquement sur des expériences infailibles. La portée sceptique de l'infailibilisme est un *topos* philosophique, et suffit à rendre l'argument de la pierre particulièrement fragile.

Quant à la proposition (2), la possibilité métaphysique que l'expérience E soit non-véridique, un défenseur de l'introspectionnisme a donc fort peu de raisons de la contester : l'expérience du libre arbitre, comme l'expérience visuelle d'une table en face de moi, est l'expérience d'un certain *fait*. Comme je l'ai montré plus haut, l'introspectionniste (l'intéroceptionniste) conçoit clairement la distinction logique entre ce fait et l'expérience elle-même, et dans la mesure où les deux sont logiquement distincts, l'introspectionniste n'a donc pas de raisons, *a priori*, de refuser la possibilité métaphysique que l'un existe sans l'autre. On *pourrait* envisager que l'analyse conceptuelle du contenu de l'expérience E rejette *in fine* la possibilité métaphysique que ce contenu soit faux lorsque E est instanciée. Mais en amont de toute analyse conceptuelle, c'est-à-dire du point de vue de l'introspectionnisme indexical, la position naturelle et par défaut semble être d'accepter cette possibilité métaphysique.¹⁵⁶

Dans ces conditions, on peut s'étonner que l'argument de Leibniz et Spinoza, au lieu de soutenir le principe (3), s'attache particulièrement à renforcer la proposition (2) à l'aide d'une expérience de pensée. Ce qui est sans doute le plus étonnant dans cet argument, c'est le fait d'apporter en renfort d'une thèse de compossibilité qui ne pose pas grand problème, une

¹⁵⁴ J'utilise ici le terme « non-véridique » de préférence à « fausse » ou à « illusoire » pour les raisons suivantes. Dire qu'une expérience « pourrait être fausse » est ambigu : en un sens, l'expérience que j'existe pourrait aussi être fausse (son contenu pourrait être faux)... dans un monde possible où je ne la possède pas. Mais ce qui nous intéresse ici est la possibilité de l'erreur *dans le monde même* où l'expérience est faite par le sujet. Dire qu'une expérience « pourrait être illusoire » évite de problème, mais dans ma terminologie introduite en p.95, le terme d'illusion désigne le caractère hallucinatoire de l'expérience, et il est compatible avec la vérité (accidentelle) du contenu halluciné. Donc ce que montre l'expérience de pensée de la pierre (en tout cas de manière immédiate et valide), c'est que l'expérience du libre arbitre peut être *non-véridique*.

¹⁵⁵ Une version plus générale de l'infailibilisme de la justification s'appliquerait à la justification inférentielle en même temps qu'à la justification basique (ou par expérience) à laquelle je me restreins ici.

¹⁵⁶ Le principe que j'utilise ici implicitement est le principe de « présomption de compossibilité » : il est justifié de présumer que deux faits sont métaphysiquement compossibles en l'absence de raisons de penser qu'il ne le sont pas. Cette présomption n'exclut pas que l'introspectionniste puisse

expérience de pensée dont la possibilité métaphysique est en revanche tout à fait discutable : la possibilité d'une pierre douée d'expérience. Mais, pour reprendre une célèbre question de Wittgenstein, « Pourrait-on se représenter une pierre pourvue de conscience ? » (Wittgenstein 2005, pt. I, §390). Comme Wittgenstein l'a fait observer, il est loin d'être évident que nous puissions nous représenter une pierre qui a des expériences. Pour le même genre de raisons, il est loin d'être évident que nous puissions nous représenter une rose qui a des dents : « Car où la rose pourrait-elle avoir des dents ? » (Wittgenstein 2005, pt. II, xi, p.311). De même qu'on ne saurait pas bien comment insérer des dents sur une rose, on ne saurait pas bien comment intégrer des expériences conscientes dans une pierre. L'expérience de pensée de la pierre heureuse de tomber pose donc un problème conceptuel important, et il semble étrange de vouloir l'utiliser pour établir la faillibilité de l'expérience E qui, quant à elle, ne pose pas vraiment problème. Pour caricaturer cette stratégie, on pourrait aussi essayer de défendre la prémisse (2) sur la base de l'expérience de pensée suivante : « C'est comme si le nombre 2 prenait plaisir d'être pair ; car il croirait l'être librement, ne s'apercevant pas des raisons logiques qui le déterminent à être pair ». ¹⁵⁷ En fin de compte, il me semble que le recours à l'image de la pierre ou de l'aiguille joyeuses est vraiment faible dialectiquement et que son succès est dû plutôt à une forme d'efficacité rhétorique, car elle permet à Spinoza une ultime pointe contre l'orgueil humain : « Telle est cette liberté humaine que tous se vantent de posséder » (Spinoza 1986, 304). A ce titre, je suivrai l'exemple de (Dennett 1984) et rangerai ces métaphores terrifiantes et déprimantes (mais dialectiquement très faibles) dans la catégorie des « grands méchants loups » ¹⁵⁸ du débat sur le libre arbitre. Mieux vaut donc libérer cet argument infaillibiliste de la très faible prémisse d'expérience de pensée et s'en tenir à ceci :

Argument infaillibiliste anti-introspection

(2) Il est métaphysiquement possible que l'expérience E d'être libre soit non véridique.

(Infal.) S'il est métaphysiquement possible qu'une expérience doxastique soit non véridique, alors elle ne confère pas de justification épistémique à son contenu.

(C) L'expérience E d'avoir le libre arbitre n'apporte aucune justification à croire au libre arbitre.

Mais même en l'état, la réponse semble évidente : certes, il est à peu près certain que l'expérience E est faillible, mais le principe infaillibiliste est faux ! Insister gravement sur la faillibilité de notre expérience E ne représente donc pas un grand risque pour l'introspectionnisme, et on peut s'étonner que cet argument ait eu autant d'influence.

En réalité, il faut reconnaître que les premières versions historiquement importantes de l'introspectionnisme (celles de Descartes et Malebranche) ont largement prêté le flanc à l'argument infaillibiliste. En effet, Descartes et Malebranche défendaient à la fois une thèse infaillibiliste en épistémologie, et une conception infaillibiliste de l'expérience du libre arbitre, de sorte qu'en exhibant la simple faillibilité de l'expérience du libre arbitre, on atteignait effectivement le cœur de leur position.

Nous avons déjà évoqué (p. 102) l'élément fondamental de l'infaillibilisme Malebranchiste, à savoir sa théorie du « sentiment intérieur ». Pour Malebranche, notre expérience du libre arbitre est infaillible exactement au même titre que notre expérience du

¹⁵⁷ Cette expérience de pensée serait évidemment contrefactuelle car, comme le sait tout bon latiniste, en réalité « le numéro deux se réjouit d'être impair » *Numero deus impare gaudet* (Virgile, *Eglogues*, VIII, v.75) cf. (Gide, *Paludes*).

¹⁵⁸ Dennett (1984, 4) parle de « bugbears », c'est-à-dire ces monstres terrifiants des histoires enfantines, qui se cachent sous les lits et dans les placards. Je n'ai pas trouvé de meilleur équivalent en français que le « grand méchant loup ».

plaisir ou de la douleur, ou encore que notre expérience d'avoir une expérience visuelle comme d'une pomme rouge. Notre expérience visuelle comme d'une pomme rouge, en elle-même, est faillible : il peut ne pas y avoir de pomme rouge en face de moi. Cette expérience est faillible parce qu'elle relève des sens extérieurs. Mais mon expérience d'*avoir une expérience* comme d'une pomme rouge n'est pas faillible, car elle relève du sentiment intérieur.

Lorsqu'un bien particulier nous est présenté, nous avons sentiment intérieur de liberté à son égard, comme nous en avons de notre plaisir et de notre douleur, lorsque nous sentons. [...] Il n'en est pas de même du sentiment intérieur, comme de nos sens extérieurs ; ceux-ci nous trompent toujours en quelque chose, lorsque nous suivons leur rapport, mais notre sentiment intérieur ne nous trompe jamais.
(Malebranche 1979, 807)

Nous avons déjà vu que cette théorie était tout à fait problématique parce qu'elle revient à faire du libre arbitre une « donnée immédiate de la conscience », c'est-à-dire lui-même une *expérience* d'un certain type. Or qu'il y ait une expérience du libre arbitre est indubitable, mais que le libre arbitre *soit* lui-même réductible à une expérience subjective, c'est particulièrement problématique.

Il me semble que cet infailibilisme Malebranchiste est une déformation d'un infailibilisme beaucoup plus subtil chez Descartes. Chez Descartes, l'infailibilité du libre arbitre est d'abord une indubitabilité qui provient de l'étape du doute hyperbolique :

Au même temps que nous doutions de tout, et que nous supposions même que celui qui nous a créés employait son pouvoir à nous tromper en toutes façons, nous apercevions en nous une liberté si grande, que nous pouvions nous empêcher de croire ce que nous ne connaissions pas encore parfaitement bien. Or, ce que nous apercevions distinctement, et dont nous ne pouvions douter pendant une suspension si générale, est aussi certain qu'aucune autre chose que nous puissions jamais connaître.
(Descartes 1953, 588)

Malgré le verbe « apercevoir », il n'est pas clair que Descartes conçoive ici le libre arbitre comme une sorte de donnée immédiate de la conscience, une donnée qui garderait son statut de donnée même dans l'hypothèse sceptique la plus radicale. Ce qui est indiscutable, c'est seulement ceci : le libre arbitre est une *condition* qui est préservée dans la description du scénario de doute hyperbolique. C'est une proposition qui n'est pas mise en doute par le scénario sceptique. Et on peut même ajouter : c'est une proposition qui est essentielle pour pouvoir mettre en doute les autres propositions (l'existence du monde extérieur etc.). En effet, c'est bien la *liberté* de suspendre mon assentiment qui me permet d'entrer dans le doute hyperbolique. Autrement dit, l'indubitabilité du libre arbitre chez Descartes ressemble beaucoup plus à ce que Wittgenstein a appelé des « propositions charnières » du doute, qu'à une donnée immédiate de la conscience : si nous ne pouvons pas en douter, ce n'est pas parce qu'il est toujours immédiatement donné, mais parce qu'il est une condition de possibilité du doute lui-même. Certains diront une condition « logique », d'autres une condition « grammaticale », d'autres encore une condition « transcendantale ». Par là, l'argument de Descartes ressemble beaucoup plus aux arguments transcendants à la Kant ou à la Donagan, qu'à une position introspectionniste proprement dite. Faut-il alors cesser de considérer Descartes comme un introspectionniste ? Je laisserai cette question aux historiens, mais il me

semble qu'on peut dire ceci : la position de Descartes a des formulations introspectionnistes¹⁵⁹ qui sont clairement l'origine de la tradition introspectionniste moderne et la référence des discussions anti-introspectionnistes ; mais son argument infaillibiliste (ou indubitabiliste) peut également amener à considérer Descartes comme étant à l'origine de la tradition transcendantale moderne.

En tout, ce qu'il importe de remarquer pour mon propos, c'est que l'argument infaillibiliste de Descartes *l'éloigne* logiquement de son propre introspectionnisme : l'introspectionnisme et l'infaillibilisme ne font pas bon ménage, et Malebranche, qui a tenté de concilier les deux dans sa théorie du « sentiment intérieur », a produit ainsi une version particulièrement faible et vulnérable de l'introspectionnisme. Ce n'est pas cette version, historiquement située, qui nous intéressera ici : les auteurs qui défendent aujourd'hui une position introspectionniste ont intelligemment laissé derrière eux ce malheureux accident de l'histoire de l'introspectionnisme.

Etant admis que l'expérience du libre arbitre est *faillible*, il ne nous reste plus qu'à montrer pourquoi cela n'empêche en rien de fonder sur elle une croyance justifiée.

II.2.b. Portée sceptique de l'infaillibilisme

Le problème de l'argument infaillibiliste anti-introspection est le même que celui de l'argument inférentialiste anti-introspection : il mène à un scepticisme radical, car il remet en cause la justification de nos croyances perceptives. Par conséquent, même s'il marchait, il ne permettrait pas de montrer que la croyance introspective au libre arbitre est *en elle-même* particulièrement problématique. Pour mettre en évidence la portée sceptique de l'infaillibilisme, je me situerai dans un cadre fondationaliste pour toutes les raisons évoquées en II.1.

La manière la plus simple et la plus directe de montrer la portée sceptique de l'infaillibilisme est l'argument suivant :

Argument infaillibiliste anti-perception

(P) Il est métaphysiquement possible que nos expériences perceptives soient non véridiques.

(Infal.) S'il est métaphysiquement possible qu'une expérience doxastique soit non véridique, alors elle ne confère pas de justification épistémique à son contenu.

(C) Nos expériences perceptives n'apportent aucune justification à croire à leur contenu.

Cet argument est évidemment le parallèle exact de l'argument anti-introspection ; il n'en diffère que par le fait qu'il est tourné contre une théorie fondationaliste de la perception. Le parallèle entre les deux arguments a l'avantage de montrer clairement pourquoi l'argument anti-introspection ne peut pas être un argument vraiment problématique : si l'on ne considère comme justifiées que les croyances qui reposent sur des expériences infaillibles, alors certes

¹⁵⁹ Notamment le titre de l'article 39 des *Principes*, cité plus haut p. 81... mais il faut remarquer que la rédaction du titre français n'est pas clairement attribuable à Descartes lui-même. Mais le vocabulaire de « l'expérience » se retrouve également dans les réponses aux Cinquièmes Objections : « [ce que vous niez touchant l'indifférence de la volonté] est tel que chacun le doit plutôt ressentir et expérimenter en soi-même que se le persuader par raison. [...] Ne soyez donc pas libre, si bon vous semble ; pour moi, je jouirai de ma liberté, puisque je la ressens en moi-même [*illam apud me experiar*]. [...] Et peut-être que je trouverai plus de créance en l'esprit des autres en assurant ce que j'ai expérimenté, et dont chacun peut aussi faire épreuve en soi-même [*apud se experiri*] ». (Descartes 1953, 499)

notre croyance au libre arbitre s'effondre, mais c'est tout l'édifice de la perception qui s'effondre en même temps. Cet argument anti-perception a cependant un inconvénient, c'est qu'il s'attaque à une *certaine* théorie fondationnaliste de la perception, à savoir la théorie selon laquelle nos expériences perceptives fournissent une justification *directe* à accepter leur contenu (le Fondationnalisme Direct de la perception). Or sans remettre en cause le fondationnalisme, on pourrait soutenir d'autres théories sur la manière dont nos expériences perceptives fournissent une justification. On pourrait soutenir des versions *indirectes* du fondationnalisme de la perception. Et parmi ces théories fondationnalistes indirectes, peut-être est-il possible d'en trouver une qui préserve le principe infaillibiliste ? En d'autres termes, on pourrait essayer de fonder l'édifice de nos croyances perceptives en le faisant reposer uniquement sur des bases infaillibles. C'est évidemment la tentative cartésienne que je suis en train de décrire. Si cette tentative réussit, alors l'infailibilisme ne mène pas au scepticisme, et pourrait donc, semble-t-il, constituer un principe plausible pour réfuter l'introspectionnisme du libre arbitre. En réalité, ce problème n'est qu'apparent, pour deux raisons. Tout d'abord il y a de très bonnes raisons de penser que la tentative cartésienne échoue. Et deuxièmement, quand bien même la tentative cartésienne réussirait, sa défense de la perception pourrait être immédiatement transposée à l'expérience du libre arbitre. Je vais développer ces deux points en repartant d'un bref aperçu de la solution cartésienne (ou de type cartésien).¹⁶⁰

Le fondationnalisme infaillibiliste consiste donc à n'accepter au fondement de l'édifice des raisons que des bases infaillibles. Quelles seront ces bases infaillibles ? Pas le témoignage, assurément. Pas la perception non plus : comme l'expérience perceptive d'une table rouge n'est pas infailliblement associée à une table rouge, il est clair qu'elle ne pourra pas délivrer comme croyance basiquement justifiée qu'il y a une table rouge. Mais l'expérience perceptive d'une table rouge est en revanche infailliblement associée à la vérité de la proposition suivante : « j'ai l'expérience perceptive d'une table rouge ». Elle pourra donc me justifier à croire cette proposition. Et d'après les définitions données plus haut, le mécanisme par lequel une expérience que *p* fonde la croyance que « j'ai l'expérience que *p* » correspond au sens étroit du mot *Introspection*. Descartes peut donc accepter, au fondement de notre édifice des raisons, les croyances introspectives au sens étroit. Par ailleurs, Descartes admet également au fondement de notre édifice de raisons les croyances claires et distinctes, telles que « pour penser il faut être », ou certaines vérités mathématiques comme « $1 + 1 = 2$ ». Toutes les croyances mathématiques ne sont pas connues de manière claire et distincte : certaines sont seulement déduites à partir de principes, et sont trop complexes pour que nous en ayons immédiatement une notion claire et distincte. Et pour ces croyances déduites, nous pourrions nous tromper dans nos déductions. Mais pour ce qui est des croyances mathématiques que nous voyons clairement et distinctement être vraies, on peut soutenir qu'elles sont infaillibles. Pour transposer dans notre vocabulaire, cela signifie que notre expérience doxastique pourra fonder la justification à croire que « $1 + 1 = 2$ », par exemple, parce que cette expérience doxastique est infaillible. Pour désigner cette faculté de fonder des croyances basiques sur des expériences doxastiques mathématiques ou notionnelles, j'emploierai le terme (malheureusement très polysémique) de *Raison*.¹⁶¹ L'infailibilité de l'Introspection (au sens

¹⁶⁰ Dans la présentation qui suit, je ne prétends pas donner une analyse historique de l'épistémologie de Descartes. Ce qui m'intéresse, c'est une certaine position dialectique, à savoir la position fondationnaliste infaillibiliste, dont Descartes est souvent considéré comme le représentant le plus important.

¹⁶¹ Je vois au moins cinq usages importants de ce terme (en laissant de côté l'usage kantien qui oppose « Raison » et « Entendement »). Dans le sens 1, que je retiens, la raison est la faculté basique délivrant les « vérités de raisons », les vérités mathématiques ou notionnelles – par opposition aux autres facultés basiques éventuelles, en particulier la perception ou l'introspection. Dans le sens 2, le mot raison est parfois employé pour désigner la faculté inférentielle (déductive ou inductive) – par opposition à toutes les facultés basiques, à la fois la perception, l'introspection et la raison-1. Dans un 3^e sens, on désigne parfois la combinaison des deux premières facultés, c'est-à-dire qu'on attribue à la raison-3 toutes les croyances attribuables ultimement (soit

étroit) et de la Raison, si du moins on accepte que ces deux facultés sont effectivement infaillibles,¹⁶² permet donc à l'infailibiliste d'éviter le scepticisme *universel* – une position qui a l'inconvénient d'être assez clairement auto-réfutante. Mais si l'on s'en tient strictement à ces deux sources, on est tout de même confiné au solipsisme. L'étape suivante consiste à ajouter, parmi les sources infaillibles, une source *inférentielle*, à savoir la *Déduction*.¹⁶³ Peut-on sortir du solipsisme en opérant des *déductions* à partir de nos croyances basiques *a priori* d'une part et *introspectives* d'autre part ? C'est là en tout cas le « projet de Descartes ». On connaît les grandes lignes de la solution cartésienne : l'introspection m'apprend que « j'ai des expériences sensibles comme d'un monde extérieur », et la raison¹⁶⁴ m'apprend que « l'être parfait (Dieu) existe et ne pourrait vouloir que mes expériences sensibles soient trompeuses ». La déduction conclut alors : « donc mes expériences sensibles comme d'un monde extérieur ne sont pas trompeuses ». Cette harmonieuse collaboration entre introspection, raison et déduction, pour « récupérer » en contexte infailibiliste la crédibilité de nos sens faillibles, n'a plus beaucoup d'adhérents aujourd'hui. Les épistémologues contemporains sont soit infailibilistes et sceptiques (c'est le cas de Peter Unger dans son livre *Ignorance* (1975)), soit faillibilistes par anti-scepticisme (c'est le cas de la très grande majorité, ainsi que le constate (Reed 2002, 143) : « Le faillibilisme est accepté par à peu près tous les épistémologues contemporains »).¹⁶⁵ L'origine de ce consensus est assez évidente : la solution cartésienne

directement soit inférentiellement) à des expériences doxastiques notionnelles – la raison-3 s'oppose alors l'expérience, c'est-à-dire l'ensemble des croyances attribuables ultimement (soit directement soit inférentiellement) à des expériences perceptives ou introspectives. Dans un 4^e sens, la raison s'oppose aux croyances testimoniales, c'est-à-dire aux croyances ultimement attribuables à la raison-3 ou à l'expérience d'autrui ; la raison-4 est donc la faculté de produire des croyances ultimement fondées uniquement sur mes propres facultés basiques. Dans un 5^e sens enfin, la raison s'oppose à la foi conçue comme faculté surnaturelle ou religieuse ; dans la mesure où la faculté testimoniale n'est pas surnaturelle et n'a pas forcément un contenu religieux, le sens 5 de la raison réunit donc à la fois la raison-4 et les facultés testimoniales naturelles non religieuses ; d'une manière générale, la raison-5 est donc la faculté naturelle de produire (basiquement ou inférentiellement) des croyances ne reposant pas ultimement sur un témoignage religieux. (NB : la deuxième partie de la définition sert à exclure du domaine de la raison-5 les croyances, en droit possible, qui seraient produites de manière naturelle sur la base d'un témoignage religieux).

¹⁶² Le point n'est pas absolument évident : la Raison et l'Introspection sont assurément les *meilleurs* candidats si on cherche des facultés infaillibles, mais en dehors d'un tel réquisit, on peut trouver des raisons de mettre en doute leur infaillibilité. Du côté de la Raison, il est évident que nous pouvons avoir des intuitions mathématiques à la fois basiques et erronées ; Descartes résout ce problème grâce à la distinction entre intuitions mathématiques claires et distinctes, et intuitions obscures et/ou confuses. Mais les discussions du début du XX^e siècle sur les premiers principes des mathématiques peuvent amener à douter que nous ayons les moyens de discerner ainsi entre nos intuitions mathématiques : l'intuition du principe de compréhension non restreinte, dont Russell a montré qu'elle menait à des contradictions, ne semble pas moins claire ni moins distincte que l'intuition des principes qui le remplacent dans le système axiomatique ZF. Il semble donc que les mathématiques doivent nécessairement reposer sur des intuitions ultimes qu'aucun critère interne ne permet de distinguer d'une intuition erronée. Voir notamment (Bealer 1999, 31) : « the existence of the paradoxes suggests that the infailibilist theory of intuition is mistaken ». Du côté de l'Introspection (au sens étroit), on peut également mettre en doute l'infailibilité du processus par lequel la croyance introspective tente de ressaisir descriptivement le contenu de l'expérience. Voir (BonJour 2001, 24–26). Cette possibilité sera abordée plus loin, car elle est au cœur des discussions de la deuxième partie de cette étude, où nous verrons justement des désaccords sur l'introspection du contenu de l'expérience du libre arbitre.

¹⁶³ Il est clair que l'induction et l'abduction sont des sources inférentielles faillibles.

¹⁶⁴ Peut-être aidée de l'introspection de l'existence en moi d'une Idée de Parfait, ou à partir d'une preuve ontologique de l'existence de Dieu.

¹⁶⁵ La discussion sur l'infailibilisme fait généralement entrer dans la discussion un troisième personnage, le « contextualiste » (ou le « variantiste »). Il se trouve, d'ailleurs, que la position de Peter Unger dans ses derniers ouvrages, bien qu'il s'agisse encore d'un type de scepticisme, est également et peut-être prioritairement une position contextualiste. Mais la discussion sur le contextualisme ne nous concerne pas vraiment ici : il consiste à dire que le sens du mot « savoir » peut varier d'un contexte à un autre. Dans certains contextes (le contexte de la vie quotidienne, notamment) nous utilisons un sens de « savoir » faible, faillibiliste, et qui rend vraie la proposition que « nous savons qu'il y a un monde extérieur ». Mais dans d'autres contextes (notamment le

elle-même est considérée comme difficilement acceptable (car peu d'épistémologues, théistes ou non, sont prêts à faire reposer l'épistémologie sur une preuve de l'existence de Dieu *a priori*, ou par l'idée de parfait), et il n'est pas clair du tout qu'il y ait des solutions alternatives pour récupérer la perception sensible en contexte infaillibiliste. Il ne semble donc pas y avoir de moyen d'éviter le scepticisme du monde extérieur si l'on accepte l'infaillibilisme.

Mais supposons un instant que la solution cartésienne (ou une autre du même genre) réussisse à sauver un infaillibilisme non sceptique. Que s'ensuivrait-il pour la discussion sur la faillibilité de l'expérience du libre arbitre ? Faudrait-il dire que l'argument infaillibiliste anti-introspection condamne efficacement l'expérience du libre arbitre *sans* condamner en même temps les expériences sensibles ? Absolument pas : si l'on accepte une stratégie cartésienne pour sauver de manière indirecte les expériences sensibles (qui sont faillibles), on pourra de la même manière indirecte sauver l'expérience du libre arbitre en dépit de sa faillibilité. On pourra adopter, par exemple, la défense suivante : certes notre expérience d'être libre est faillible, mais Dieu serait trompeur si cette expérience était systématiquement erronée, donc elle ne l'est pas. Si quelqu'un objecte à cela que le sauvetage indirect de la perception est plus plausible que le sauvetage indirect de l'expérience du libre arbitre, il devra nécessairement trouver une *autre* caractéristique problématique de l'expérience du libre arbitre, une caractéristique qui la distingue des expériences perceptives, par exemple le fait que son contenu est métaphysiquement lourd, ou que nous n'avons pas d'indices de sa fiabilité, alors que nous en avons peut-être pour la perception. Autrement dit, il devra faire reposer son argument anti-introspection sur une caractéristique qui n'a plus rien à voir avec le problème de la faillibilité, et son argument aura alors tout intérêt à être envisagé dans le cadre faillibiliste consensuel plutôt que dans le cadre, complexe et marginal, d'une reconstruction infaillibiliste de (certaines) expériences faillibles.

Pour simplifier, je laisserai donc de côté l'éventualité d'une reconstruction cartésienne ou quasi cartésienne défendant à la fois l'infaillibilisme et l'anti-scepticisme. Je m'en tiendrai au consensus actuel selon lequel le principe infaillibiliste mène au scepticisme de la perception. Tout fondationnaliste se doit donc d'accepter au fondement de notre structure justificationnelle des expériences faillibles. La faillibilité de l'expérience du libre arbitre ne constitue donc pas un problème pour l'épistémologie introspectionniste.

II.2.c. Solution : sécurité faillibiliste de la connaissance, et justification *prima facie*

Le paragraphe précédent était essentiellement négatif : il consistait à refuser un certain principe (l'infaillibilisme de la justification). Pour éclairer la discussion, je vais à présent envisager ce qui s'ensuit positivement pour l'épistémologie du libre arbitre. Plus précisément, je vais aborder les deux questions suivantes : Peut-on faire droit à une intuition profonde qui serait à l'origine du principe infaillibiliste ? A quoi ressemble positivement une épistémologie qui refuse le principe infaillibiliste ? Pour répondre à la première question, je vais montrer qu'il y a bien une intuition légitime sur laquelle repose le principe infaillibiliste, à savoir une intuition fiabiliste, mais que cette intuition concerne la connaissance (et pas la justification) et que par ailleurs elle ne peut pas être aussi forte que le voudrait l'infaillibiliste. Pour répondre à la seconde question, j'exposerai la notion de justification *prima facie*, qui constitue le

contexte d'une discussion sur le scepticisme), le mot « savoir » acquiert un sens plus fort, peut-être même un sens infaillibiliste, dans lequel la phrase « nous savons qu'il y a un monde extérieur » est fausse. Mais pour chaque sens de « savoir », pris individuellement, il reste admis que *soit* il est faillibiliste, *soit* il mène au scepticisme, *tertium non datur*. Dans la discussion présente, j'applique ce dilemme à une notion unique, la notion de justification telle que je l'ai définie en introduction.

concept central d'une théorie faillibiliste de la justification. Dans les deux cas, bien évidemment, ces réponses seront appliquées directement à l'introspection du libre arbitre.

Sur quelle intuition repose le principe infaillibiliste ? L'intuition fondamentale qui amène à l'infailibilisme semble être l'intuition d'une incompatibilité entre la *connaissance* et la *possibilité de l'erreur*. On pourrait exprimer cette intuition de la manière suivante : s'il est possible que S se trompe à propos de p , alors on ne peut pas dire que S sait que p . Ou pour reprendre une formulation plus simple et classique :

(~PE) Si S sait que p , alors S ne peut pas se tromper.

Cet énoncé, en l'état, est notoirement ambigu, et pour lever ces ambiguïtés, le plus simple sera de recourir, à partir de maintenant, à quelques symboles formels très courants.¹⁶⁶ La première ambiguïté que recèle (~PE) concerne la portée de l'opérateur d'impossibilité « ne peut pas ». Si on lui donne la portée large, on obtient le principe de Factivité de la connaissance :

(F) $\sim\Diamond (Kp \wedge \sim p)$, ou encore $\Box (Kp \supset p)$

Le principe de factivité, comme nous l'avons vu en introduction, rappelle simplement l'idée universellement admise que la condition de vérité fait partie des conditions qui définissent la notion de « savoir ». Si on donne à l'opérateur modal la portée étroite, on obtient un principe qu'on peut appeler principe de Nécessité :

(N) $Kp \supset \sim\Diamond\sim p$, ou encore $Kp \supset \Box p$

Le principe de Nécessité est encore assez loin de l'infailibilisme proprement dit : il revient à dire que toute connaissance est connaissance d'une vérité nécessaire (par exemple, d'une vérité mathématique ou logique). L'infailibiliste, quant à lui, acceptera normalement comme connaissances certaines croyances à des vérités contingentes, à condition que ces vérités contingentes soient crues de manière infaillible ; l'exemple par excellence d'une croyance contingente et pourtant infaillible est évidemment le *cogito*. L'idée de l'infailibiliste n'est donc pas que la *proposition* crue ne peut pas être fausse (s'il y a connaissance) mais bien que la *croyance* ne peut pas être fausse. Autrement dit :

(Infail.) $Kp \supset \sim\Diamond (Bp \wedge \sim p)$, ou encore $Kp \supset \Box (Bp \supset p)$

Certains auteurs (Audi 1998, 304; Engel 2007, 27) ont suggéré que l'attrait intuitif de l'infailibilisme reposait principalement sur cette ambiguïté de (~PE), et sur la confusion avec le principe de Factivité, lequel est évidemment vrai.

Pourtant, l'intuition d'une tension entre connaissance et possibilité de l'erreur résiste à cette clarification nécessaire. La meilleure façon de faire apparaître cette tension, c'est de remarquer notre résistance à accepter des « propositions de loterie ».¹⁶⁷ Supposez par exemple que Laurent mette un jeton sur la case 'pair' d'une roulette ; il croit que son jeton va perdre ;

¹⁶⁶ Voici, pour le lecteur non habitué, les définitions de ces symboles : K désigne la notion de connaissance (knowledge), B la notion de croyance (belief), et J la notion de justification. Les symboles $\Box$ et $\Diamond$ désignent, respectivement, la nécessité et la possibilité. Les symboles $\supset$, $\rightarrow$, et $\Rightarrow$ désignent respectivement les conditionnels matériel, contrefactuel et strict.

¹⁶⁷ La littérature sur les « propositions de loterie » est très abondante. Un traitement magistral de ce problème est développé dans (Hawthorne 2004).

et la roulette lui donne raison. Il est clair dans ce genre de cas que la conjonction suivante est très malheureuse :

Jean savait que son jeton déposé sur 'pair' allait perdre, mais il était parfaitement possible que cette croyance s'avère fausse.

Les conjonctions de ce genre sont malheureuses, semble-t-il, précisément parce que la notion de connaissance contient une notion de « sûreté » ou de « sécurité » face aux erreurs possibles. Pour reprendre la formulation prudente de Hawthorne,

Savoir si une croyance vraie que *p* constitue une connaissance dépend, semble-t-il, de la facilité plus ou moins grande avec laquelle le sujet aurait pu se tromper à propos de *p* et à propos de questions semblables. (Hawthorne 2004, 56)¹⁶⁸

On ne peut donc pas faire comme si la notion de savoir était totalement indépendante de l'idée de « sécurité contre l'erreur ». Mais la question est de savoir quel type ou quel niveau de sécurité est requis pour qu'il y ait connaissance. Si on regarde notre usage courant du mot « savoir », il semble clair qu'on attribue à certains sujets la connaissance alors même que certaines possibilités d'erreur persistent. Par exemple, pour pouvoir dire que « Vincent sait que sa voiture est au garage », il faut sans doute pouvoir exclure la possibilité que sa femme ou son fils vient de descendre pour prendre la voiture, mais il n'est pas nécessaire de pouvoir exclure la possibilité que les martiens aient détruit la terre, à l'exception de la chambre où se trouve Vincent. De la même façon, pour pouvoir dire qu'une voiture est sûre, il est certainement nécessaire de pouvoir exclure la possibilité qu'elle tombe subitement en panne sur autoroute, mais il n'est pas nécessaire d'exclure la possibilité qu'elle tombe en panne à cause d'un champ magnétique induit par des martiens. Dans notre usage courant, la sécurité des croyances, comme la sécurité des voitures, ne semble pas requérir une sécurité *contre toute possibilité logique* de défaillance. Il semble donc que la notion commune de sécurité contre l'erreur n'oblige pas à accepter la thèse radicale de l'inaffabilité, avec ses conséquences sceptiques : sécurité contre certaines erreurs pertinentes, oui ; sécurité contre toute erreur possible, sûrement pas.

L'inaffabilité contemporain a pourtant un argument assez fort pour nous inviter à faire le passage de la notion de sécurité à la notion d'inaffabilité. L'argument de l'inaffabiliste Peter Unger (1975) consiste à reconnaître que nos usages courants du terme « savoir » ne sont pas sensibles à certaines possibilités d'erreur (et notamment aux possibilités d'erreur typiques des scénarios sceptiques). Mais il considère que le terme de « savoir » ressemble en cela à toute une série de termes d'évaluation, qu'il appelle « termes absolus » (Unger 1975, chap. 2). Un terme absolu, par exemple le terme « plat »... ou le terme « sûr » pour la voiture, peut être utilisé de manière graduelle : la table peut être « plus plate » que le canapé, la F1 de Schumacher peut être « plus sûre » que la 4L de Bourvil. Et en ce sens, on peut dire que la table est « plate » même s'il y a encore à sa surface des aspérités, ou que la F1 est « sûre » même s'il y a encore des possibilités de défaillance technique. On peut donc utiliser ces termes dans un sens plus ou moins fort, plus ou moins rigoureux. Mais ce qui caractérise les termes « absolus » par rapport à d'autres catégories de termes graduels, c'est qu'il existe un usage « limite » : une surface serait « plate » (absolument, ou strictement) s'il n'y avait plus *aucune* aspérité ; une voiture « sûre » (absolument, ou strictement) s'il n'y avait plus *aucune* possibilité de défaillance... et de même une croyance sera « sûre » (absolument, ou strictement) s'il n'y a plus *aucune* possibilité d'erreur. Et donc, une croyance ne sera une

¹⁶⁸ Ainsi que le rappelle Hawthorne, en note 16, l'association conceptuelle entre connaissance et sécurité a été défendue tout particulièrement par (Sosa 2000) et (Williamson 2000).

connaissance (absolument, ou strictement) que si elle est infaillible. L'idée de Unger consiste donc à accepter l'existence de sens faillibilistes du mot « savoir » ou « connaissance »... mais en soulignant qu'il s'agit là de sens « lâches » qui sont liés à un certain contexte (un contexte dans lequel certaines possibilités d'erreur ne comptent pas) ; et quand on prend la notion de sécurité au sens strict (c'est-à-dire dans un contexte où toutes les possibilités d'erreur comptent), on arrive nécessairement à la notion d'infaillibilité. En gardant cette notion de variation contextualiste, on peut même défendre que la notion de « connaissance » est *toujours* (analytiquement) infaillibiliste, qu'elle exclut toujours « toutes » les possibilités d'erreur...et que c'est l'ensemble des possibilités d'erreur (pertinentes) qui varie avec le contexte.¹⁶⁹ Concrètement, la position infaillibiliste sceptique reviendra à dire ceci : certes, la croyance de Vincent (que son jeton posé sur 'pair' va perdre) est menacée par des possibilités d'erreur plus évidentes que telle croyance perceptive commune (qu'il y a une table en face de moi), mais la différence entre les deux croyances est une simple différence de degré, une différence sur le nombre de possibilités, car l'important pour être une connaissance au sens strict, ce serait de pouvoir éliminer vraiment *toutes* les possibilités d'erreur. Appliquée au problème du libre arbitre, cette conclusion signifie que notre expérience du libre arbitre, quand bien même elle ne pourrait éliminer toutes les possibilités d'erreur, se retrouve essentiellement dans le même sac que toutes nos croyances (faillibles), ce qui, encore une fois ne met pas l'introspectionniste dans une position particulièrement problématique.

Qu'en est-il des réponses au scepticisme infaillibiliste ? Certaines réponses risquent-elles de poser problème pour l'introspectionnisme du libre arbitre ? On peut ranger les réponses dans deux grandes catégories, que (Hawthorne 2004) appelle l'invariantisme modéré et le contextualisme. L'invariantisme modéré, qui est la position traditionnelle, consiste à dire que sur l'échelle des différents niveaux de « sécurité » (c'est-à-dire des sécurités à l'égard d'un nombre plus ou moins grand de possibilités d'erreur), le sens du mot « savoir » (ou du moins le sens *privilegié* du mot « savoir ») correspond à un niveau de sécurité modéré, pas à un degré de sécurité absolu : pour « savoir », dans le sens important, il n'est pas nécessaire d'éviter *toutes* les possibilités d'erreur. L'invariantiste modéré (à la différence de l'invariantiste *sceptique*) refuse donc de dire que le sens privilégié de « savoir » soit un sens absolu... mais il est d'accord avec l'infaillibiliste (c'est-à-dire l'invariantiste sceptique) pour dire qu'il y a un unique sens privilégié du mot « savoir ». C'est ce point que conteste précisément le contextualiste. Pour le contextualiste,¹⁷⁰ tous les niveaux de sécurité se valent ; on peut donc dire que les croyances perceptives faillibles ne sont pas des « connaissances » *dans un certain sens*, mais ce sens n'est pas spécialement privilégié, il n'est pas « le sens strict », et en particulier, il n'a aucune légitimité pour nos énoncés quotidiens de la forme « S sait que p ». Pour des raisons de clarté, le cadre théorique que j'adopterai pour la suite de la discussion sur le libre arbitre sera beaucoup plus proche de l'invariantisme modéré que du contextualisme. Mais cela suppose quelques mots d'explication, compte tenu de l'importance de la tradition contextualiste aujourd'hui. Le grand problème de la « solution » contextualiste pour notre discussion, c'est qu'elle tourne essentiellement autour de la question de la polysémie du mot « savoir » (ou du mot « connaissance ») dans le langage naturel. Or cette question de polysémie du mot « savoir », j'ai essayé de l'écarter en introduction en stipulant de quel *concept* déterminé j'entendais parler en utilisant ce mot (quel que soit son usage courant). Plus précisément, la solution contextualiste repose sur la distinction entre des standards de sécurité communs (qui sont suffisamment bas pour que les croyances perceptives

¹⁶⁹ On exploite alors un phénomène bien connu en philosophie et en sémantique contemporaine, à savoir le caractère contextuellement variable des quantificateurs (ou des domaines de quantifications). Pour une synthèse récente sur cette question, voir notamment (Stanley and Gendler Szabó 2003).

¹⁷⁰ Une des défenses les plus importantes historiquement du contextualisme a été l'article de David Lewis (Lewis 1996). Pour une synthèse plus récente, voir (Rysiew 2011) ou de nouveau (Hawthorne 2004).

puissent les satisfaire) et des standards de sécurité élevés (mais liés à des contextes très particuliers) ; or j'ai souligné en introduction que mon projet ne s'intéressait qu'aux standards épistémiques communs et pas à un éventuel projet de rehaussement de nos standards épistémiques. Par conséquent, le *seul* sens de « savoir » qui nous intéresse ici, le seul degré de sécurité pertinent pour nous, c'est le sens suffisamment bas pour que nos croyances perceptives puissent le satisfaire. C'est ce niveau de sécurité que l'introspectionnisme du libre arbitre devra satisfaire également, s'il doit être sauvé. Une fois que cette question est posée clairement, on voit que l'existence ou non *d'autres* sens possibles du mot « savoir » est totalement périphérique à notre problème d'épistémologie appliquée. Le mot « savoir » est peut-être variant et sans signification privilégiée, mais notre intérêt ici restera attaché, invariablement et de manière privilégiée, au degré de sécurité que les croyances perceptives peuvent satisfaire.

Je vais donc brièvement exposer la manière dont l'invariantiste modéré (par opposition à l'invariantiste *sceptique*) peut réagir face à l'argument infaillibiliste. Il me semble qu'un invariantiste modéré aura naturellement tendance à reconduire l'intuition de « sécurité aux possibilités d'erreur » à une autre notion plus fondamentale... à savoir la notion « anti-hasard » dont nous avons vu qu'elle était au cœur de tout concept de connaissance. C'est notamment la démarche de Duncan Pritchard dans (Pritchard 2005). Pourquoi réinterpréter ainsi la notion « anti-possibilités d'erreur » en une notion « anti-hasard » ? Parce que le fait même de s'exprimer en termes quantifiables (au sens strict de la quantification logique) d'ensembles de possibilités d'erreur donne une bonne part de sa force argumentative à l'argument infaillibiliste : si l'on admet que le cœur de la notion de savoir est l'idée de rejeter les possibilités d'erreur, on aura immédiatement à l'esprit la possibilité d'un savoir maximal, ou privilégié, qui consiste à les rejeter *toutes*. Et les efforts pour séparer les possibilités d'erreur « dangereuses » des possibilités d'erreur « bénignes » donnera toujours une impression de compromis. Mais si l'on repart de la formulation logique du principe infaillibiliste :

(Infail.) $Kp \supset \sim \Diamond (Bp \wedge \sim p)$, ou encore $Kp \supset \Box (Bp \supset p)$, ou encore $Kp \supset (Bp \Rightarrow p)$

on constate qu'il y a deux manières logiquement équivalentes de le formuler. L'une, avec l'opérateur de possibilité, dit ceci :

(Infail.1) Si S sait que p , alors il n'y a pas de possibilité d'erreur pour sa croyance que p .

Mais l'autre formulation, à l'aide de l'opérateur de nécessité, a un « écho intuitif » légèrement différent :

(Infail.2) Si S sait que p , alors il y a un lien de dépendance logique entre sa croyance que p et le fait qu'elle soit vraie.

Or si l'intuition (Infail.1) est le passage à la limite de l'intuition de « sécurité aux erreurs », l'intuition (Infail.2) en revanche correspond à un passage à la limite de l'intuition « anti-hasard », telle que je l'ai formulée en introduction :

(AH) Si S sait que p , alors ce n'est pas par pur hasard que sa croyance que p est vraie (il y a un lien de dépendance entre sa croyance que p et le fait qu'elle soit vraie).

De même que (Infail.1) et (Infail.2) sont logiquement équivalents, de même toute version faillibiliste de l'intuition anti-hasard (par exemple un réquisit de dépendance *causale* entre fait

et croyance) reviendra à exclure certaines « possibilités d'erreur » tout en permettant d'autres. Il y a donc une certaine interchangeabilité des deux types d'intuition, quel que soit leur degré, et il est significatif de remarquer que les exemples traditionnellement avancés par les infailibilistes pour motiver l'intuition anti-possibilité d'erreur peuvent tout aussi bien être « embrigadés » comme intuitions anti-hasard. C'est le cas notamment de notre exemple de proposition de loterie : si Laurent ne sait pas que son jeton va perdre, c'est parce qu'il ne peut y avoir aucune sorte de dépendance entre sa croyance et le résultat qui sera donné par la roulette.

Si les deux intuitions de base sont interchangeables, il est cependant très important de savoir laquelle des deux est la plus fondamentale ; car elles ne se prêtent pas toutes les deux si facilement à l'argument infailibiliste du « passage à la limite ». Comme on vient de le voir, l'intuition anti-possibilités d'erreur se prête très facilement à un tel argument. Mais lorsqu'on part de l'intuition anti-hasard, lorsqu'on part de l'idée que pour avoir une connaissance que *p*, il faut que notre croyance ait une certaine dépendance par rapport au fait que *p*, il n'est pas évident du tout qu'on soit insensiblement et naturellement porté à considérer la dépendance *logique* comme modèle ou comme idéal d'une telle dépendance fait-croyance. Si l'on cherche fondamentalement à éliminer les possibilités d'erreur, il semble naturel de vouloir les éliminer toutes ; si on cherche à éliminer le hasard, il n'est pas clair du tout que la possibilité logique d'un démon cartésien (ou d'une manipulation martienne) puisse vraiment être catégorisée comme un « hasard ». Je ne développerai pas ici l'analyse détaillée de cette notion de hasard ; ce qui m'intéresse pour l'instant, c'est seulement l'idée suivante : l'intuition fondamentale et légitime qui peut amener parfois à l'infailibilisme n'est en fait rien d'autre que la condition anti-hasard par laquelle j'ai défini (stipulativement) mon concept de « connaissance ». La notion d'une *certaine* dépendance (non logique) entre fait et croyance permet de faire droit au cœur de cette intuition, sans tomber dans les affres du scepticisme.

Qu'en est-il de la *justification* ? La notion de justification est absente de toute la discussion qui précède sur l'infailibilisme, alors que l'argument auquel nous essayons de répondre était un argument contre la *justification* de la croyance au libre arbitre par l'expérience du libre arbitre. Si la notion de justification était absente, c'est parce qu'elle est effectivement indépendante des intuitions qui amènent au principe (Infail.). Ce principe en effet est ultimement motivé par l'intuition anti-hasard, et donc il est clair qu'il concerne les mérites *externes* de la croyance, et non pas ses mérites *internes*. De manière plus immédiate, l'infailibilisme est motivé par la condition anti-possibilités d'erreur, mais là aussi, il est clair que cette condition est, au moins de prime abord, une condition externe : une possibilité d'erreur (qu'elle soit réaliste ou farfelue) reste une possibilité d'erreur et donc une menace à la connaissance, que j'en aie conscience ou non, qu'elle soit représentée dans mon esprit ou non. Comme la condition anti-possibilités d'erreur est une condition externe, elle ne *peut pas* logiquement être une condition nécessaire à « l'impeccabilité épistémique interne », c'est-à-dire, dans ma terminologie, à la « justification ». On peut, bien sûr, ne pas s'intéresser à l'impeccabilité épistémique interne, mais dans cette partie, c'est justement cette question qui est posée : la faillibilité (logique) de notre expérience du libre arbitre rend-elle la croyance introspective au libre arbitre problématique d'un point de vue épistémique interne ? La réponse est clairement « non » : si la faillibilité de notre expérience du libre arbitre posait un problème épistémique, ce serait fondamentalement un problème de connaissance et non pas (de prime abord) un problème de justification ; mais par ailleurs, je viens de montrer que nous n'avons pas de bonnes raisons de penser que la faillibilité pose un problème pour la connaissance.

Pourtant, la question de l'infailibilité est liée à la question de la justification ; l'idée de partir de la notion de faillibilité pour constituer un problème de justification n'est pas une

simple erreur reposant sur la confusion entre impeccabilité externe (connaissance) et impeccabilité interne (justification). Il y a deux types de connexions entre infaillibilité et justification. Le premier type de connexion, abordé dès l'introduction de cette étude, est le principe de méta-cohérence : nous avons vu que le caractère hasardeux de la croyance de Gwendolen (que son fiancé s'appelle Ernest) ne posait pas *en tant que tel* un problème de justification (du moment que Gwendolen n'a aucune bonne raison de suspecter ce défaut), mais qu'il *pouvait* en droit se répercuter en problème de justification s'il était représenté dans le système de croyances de Gwendolen. Dans ce cas, avons-nous vu, ce qui poserait un problème de justification, c'est bien un élément interne, à savoir la *croyance* de Gwendolen qu'il y a un problème externe... quand bien même il n'y en aurait pas en réalité. De même, si nous avons la *croyance* que l'expérience du libre arbitre est faillible, et la *croyance* que la faillibilité pose un problème externe (et empêche le statut de « connaissance »), alors ces deux croyances (qu'elles soient vraies ou fausses en réalité) pourraient constituer la source d'un problème purement *internaliste*, un problème donc de justification. On pourrait alors formuler l'argument suivant contre la justification du libre arbitre :

Argument infaillibiliste de la connaissance à la justification

- (2) Il est métaphysiquement possible que l'expérience E d'être libre soit non véridique.
- (Infail.K) S'il est métaphysiquement possible qu'une expérience doxastique soit non véridique, alors elle ne permet pas que son contenu soit une *connaissance*.
- (C1) Donc l'expérience E d'avoir le libre arbitre ne permet pas de *connaître* que nous sommes libres.
- (3) Le philosophe est conscient des prémisses (2) et (Infail.K) ainsi que de leur conclusion (C1).
- (Méta-coh.) Si S est conscient qu'il est impossible de connaître que *p* par la méthode *M*, alors S ne peut pas être justifié à croire que *p* par la méthode *M*.
- (C2) Donc le philosophe ne peut pas être justifié à croire au libre arbitre par introspection.

Certains lecteurs trouveront peut-être plus claire la version semi-formalisée suivante :

Argument infaillibiliste de la connaissance à la justification

- (2) $\Diamond (E LA \wedge \sim LA)$
- (Infail.K) Si $\Diamond (E p \wedge \sim p)$ alors $\sim K_E p$.
- (C1) $\sim K_E LA$.
- (3) Pour le philosophe ϕ conscient de (C1), $J_\phi \sim K_E LA$.
- (Méta-coh.) Si $J_x \sim K_E p$ alors $\sim J_x p$.
- (C2) Donc pour le philosophe, $\sim J_\phi LA$.¹⁷¹

L'argument procède donc en deux étapes : une étape qui s'intéresse à l'existence *effective* ou externe de possibilités d'erreurs (qu'elles soient ou non représentées dans l'esprit du sujet), et une deuxième étape qui internalise ce problème et s'intéresse à la *représentation mentale* des possibilités d'erreur (que ces possibilités d'erreur soient effectives ou non). C'est la deuxième étape qui est cruciale pour constituer un problème de justification.

Que penser de cet argument ? Ne permet-il pas de montrer que les auteurs infaillibilistes sceptiques tels que (Unger 1975) et (Stroud 1984) ont réellement soulevé un problème de

¹⁷¹ Il est évident que ce qui est nié ici, c'est la possibilité pour le philosophe d'avoir une justification à croire au libre arbitre *sur la base de l'expérience E* ; j'aurais pu répéter l'indice *E* à côté de l'indice ϕ , mais il m'a semblé que cela ajouterait en complexité de notation plutôt qu'en clarté.

justification, contrairement à ma première réponse ? Trois remarques s'imposent ici : oui, ces auteurs sceptiques ont opéré (plus ou moins clairement) l'internalisation nécessaire pour produire un problème de justification ; non, l'argument résultant n'est pas convaincant, dans la mesure où il repose toujours sur la prémisse infaillibiliste ; et pourtant oui, l'argument résultant fait apparaître un vrai problème, indépendant de l'infailibilisme. Je reprends ces trois points de manière plus détaillée. Premièrement, les grands infaillibilistes sceptiques que sont Unger et Stroud ne se sont pas contentés de poser un problème pour la notion externaliste de connaissance, et pour parvenir à un problème interne, leurs arguments ont effectivement recours à l'internalisation du problème externe. Mais une difficulté de la discussion vient du fait que l'étape d'internalisation n'est pas toujours clairement distinguée de la première étape. Regardons par exemple le principe sur lequel repose l'argument sceptique de Stroud :

(Principe Stroud) Nous ne pouvons savoir une chose particulière que si nous avons éliminé certaines possibilités d'erreur que nous reconnaissons comme incompatibles avec le fait de savoir cette chose. (Stroud 1984, 24)

Dans la mesure où tout scénario sceptique est un scénario incompatible avec la connaissance du monde extérieur, il est clair que ce principe est infaillibiliste : *tout* scénario sceptique doit être éliminé pour qu'il y ait connaissance. Mais ce principe contient également l'élément d'internalisation nécessaire pour poser un problème de justification : l'internalisation est assurée par la précision que *nous* devons *avoir éliminé* les possibilités d'erreur. Stroud ne s'intéresse donc pas ici simplement au fait que les possibilités d'erreur doivent être absentes *en soi* ou *effectivement* ou *en externe*. Il exige que le sujet lui-même se représente leur absence.¹⁷² Autrement dit, le principe de Stroud ne permet pas de distinguer clairement, et d'évaluer séparément le réquisit externe d'absence de possibilité d'erreur et le réquisit interne de capacité à exclure les possibilités d'erreur. Un auteur qui permet de faire cette distinction de façon particulièrement claire est Duncan Pritchard : pour Pritchard (2005, chap. 6), la question de la présence ou de l'absence de possibilités d'erreur est ce qu'il appelle le *hasard épistémique véritique*, et c'est un problème externaliste. Et la question de l'accès ou non (par le sujet) à ces possibilités d'erreur est ce qu'il appelle le *hasard épistémique réflexif*, et c'est un problème internaliste. Pritchard montre parfaitement l'indépendance des deux questions : il se pourrait que nos croyances soient parfaitement fiables d'un point de vue externe, mais réflexivement hasardeuses (si nous n'avons aucun moyen de *savoir* qu'elles sont fiables) ; il se pourrait inversement que nous n'ayons aucune bonne raison de douter de la fiabilité de nos croyances... mais que par manque de chance elles se trouvent être néanmoins véritiquement hasardeuses.

Deuxième point : une fois qu'on a clairement identifié ces deux étapes, il paraît évident qu'un argument qui utilise *à la fois* ces deux étapes (que ce soit de manière explicite ou implicite) échoue, à cause de ce que nous avons déjà dit de l'infailibilisme de la connaissance : en effet, la première étape de l'argument « de la connaissance à la justification » est l'argument infaillibiliste de la connaissance. Comme nous avons vu que cet argument était faible, il s'ensuit que tout argument qui l'inclut ou le présuppose est *a fortiori* au moins aussi faible. Pour le dire autrement : puisque la faillibilité ne pose *pas* de problème

¹⁷² Il y a un débat abondant sur ce que signifie exactement « éliminer une possibilité d'erreur » : une lecture souvent jugée comme trop forte pour rendre l'argument sceptique acceptable consiste à dire que S a éliminé (ou peut éliminer) une possibilité *P* si et seulement *S sait* que $\sim P$. Une autre lecture qui serait sans doute trop faible consiste à dire que S a éliminé *P* si et seulement *S croit* que $\sim P$. Il y a entre les deux toutes sortes de solutions intermédiaires. On peut se rapporter notamment à (Pryor 2001, 99–100). En ce qui nous concerne ici, l'important est de remarquer que, dans la mesure où l'expression de Stroud s'exprime à la première personne (*nous* avons éliminé), il est clair que cette condition devra d'une manière ou d'une autre consister dans un état mental du sujet.

externe, il n'y a tout simplement aucun problème externe à internaliser. Supposons qu'on accepte le principe de méta-cohérence, et qu'on dise qu'on ne peut être justifié à accepter nos expériences sensibles ou notre expérience au libre arbitre que si nous sommes justifiés à les tenir pour des sources de connaissance ; comme leur faillibilité ne nous empêche pas de les considérer comme des sources de connaissance, elle ne nous pose pas non plus de problème de justification. La faillibilité de l'expérience du libre arbitre ne pose aucun problème interne *parce que* elle ne pose aucun problème externe.

Troisième point : l'argument d'internalisation du problème de faillibilité, bien qu'il échoue, attire notre attention sur toute une famille d'arguments possibles, fondés sur le principe de méta-cohérence (ou *des* principes de méta-cohérence). En effet, le fait qu'il n'y ait pas de problème infailibiliste externe à internaliser (tout simplement parce qu'il n'y a pas de problème infailibiliste externe) ne veut pas dire qu'il n'y ait *aucun* problème externe qu'on puisse chercher à internaliser. Si nous avons rejeté la contrainte d'infailibilité pour qu'il y ait connaissance, nous avons gardé la contrainte anti-hasard, c'est-à-dire la contrainte d'une certaine *dépendance* fait-croyance. Cette contrainte-là, qui est une contrainte externe, peut être internalisée via un principe de méta-cohérence. Et son internalisation est même le cœur du problème épistémique du libre arbitre : sommes-nous justifiés (en interne) à croire que notre expérience du libre arbitre peut constituer (du point de vue externe) une connaissance, c'est-à-dire une croyance dépendant adéquatement du fait qui est cru ? Je n'ai évidemment pas l'intention d'éluder cet argument-là. Mais dans cette section sur l'infailibilisme, je soulignerai simplement qu'il n'a, en fait, rien à voir avec la question de l'infailibilité : son principe fondamental ne concerne pas l'éventail (plus ou moins large) des possibilités d'erreur mettant en péril la connaissance ; il concerne les principes « méta-épistémiques » (i.e. les principes sur la justification à croire qu'on est justifié, le fait de savoir qu'on sait, la justification à croire qu'on sait, etc.), or ces principes peuvent poser exactement le même problème que l'infailibilisme soit vrai ou non.

Par conséquent, la première tentative pour transformer le problème (externe) de la faillibilité en un problème de justification échoue ; au mieux, cette tentative permet d'apercevoir un autre problème qui n'a en fait rien à voir avec l'infailibilisme proprement dit, un problème de métajustification ou de méta-cohérence que nous aborderons au chapitre 3.

Mais il y a une deuxième manière plus complexe dont le problème de l'infailibilisme est connecté au problème de la justification : c'est le lien entre la *faillibilité de la connaissance* et la *révisabilité de la justification*. Il convient ici d'être précis car les deux notions sont si proches qu'elles sont souvent tout bonnement confondues. La façon la plus simple de présenter la question de la révisabilité de la justification est de considérer un principe *irrévisabiliste*. Soit un individu S ayant la croyance que p à t , le principe irrévisabiliste dit que :

(Irrév.) S'il est possible que S soit justifié à croire que $\sim p$ à $t+n$ (i.e. si sa croyance que p est révisable), alors S n'est pas justifié à croire que p à t .

ou inversement :

(Irrév.) Si S est justifié à croire que p à t , alors il n'est pas possible que S soit justifié à croire que $\sim p$ à $t+n$ (i.e. que sa croyance que p soit, plus tard, à tenir pour fausse).

$J_{tp} \supset \sim \Diamond (B_{tp} \wedge J_{t+n} \sim p)$ ¹⁷³

¹⁷³ Ou plus simplement : $J_{tp} \supset \sim \Diamond J_{t+n} \sim p$; la formulation donnée dans le corps du texte permet de mieux mettre en évidence l'analogie avec l'infailibilisme.

Il est clair que l'irrévisabilisme de la justification a un rapport d'analogie avec l'inaffidabilisme de la connaissance :

(Infail.) Si S sait que p , alors il n'est pas possible que sa croyance que p soit fausse.
 $Kp \supset \sim \Diamond (Bp \wedge \sim p)$

Mais malgré les ressemblances qu'on peut apercevoir entre les deux principes, il y a aussi des dissemblances fondamentales : les deux principes ne concernent pas le même statut (impeccabilité interne vs impeccabilité externe) ; là où l'irrévisabilisme est un principe diachronique, l'inaffidabilisme est un principe synchronique ; enfin, l'irrévisabilisme exprime un rapport (de stabilité) entre justification et justification ($J_t p / J_{t+n} \sim p$), tandis que l'inaffidabilisme exprime un rapport (de dépendance) entre connaissance et *fait* ($Kp / \sim p$).

La différence entre les deux principes sera peut-être plus sensible à travers des exemples. On peut très bien envisager des croyances qui seraient irrévisables mais non pas inaffidables : supposons par exemple que je forme, à t , la croyance introspective que j'ai le même mal de dents que la veille ; il se pourrait que cette croyance soit erronée ; mais on pourrait défendre que cette croyance est irrévisable, puisqu'on ne voit pas comment je pourrais *par la suite* devenir mieux renseigné sur ma douleur présente que je ne le suis maintenant. Inversement, on peut envisager des croyances qui seraient inaffidables, et pourtant révisables : supposons par exemple que je croie au théorème de Gödel, et que ce théorème soit effectivement vrai ; puisqu'il s'agit d'une vérité logique, si elle est vraie, elle est vraie dans tous les mondes possibles, et donc il n'y a pas de monde possible où ma croyance est fausse ; ma croyance est donc inaffidable,¹⁷⁴ mais cela n'empêche pas qu'elle puisse reposer pour moi sur une justification tout à fait révisable (notamment si je l'accepte par l'autorité de la communauté mathématique et qu'un jour la communauté mathématique me fait croire qu'on a trouvé une erreur dans la démonstration de Gödel). On pourrait donc tout à fait accepter l'inaffidabilisme de la connaissance sans l'irrévisabilisme de la justification,¹⁷⁵ ou l'irrévisabilisme de la justification sans l'inaffidabilisme de la connaissance.¹⁷⁶

L'inaffidabilisme et l'irrévisabilisme sont donc logiquement indépendants. Et on peut déplorer qu'ils soient parfois confondus. C'est souvent le cas, notamment, dans les discussions sur la philosophie des sciences Popperienne. Dans ce contexte, ce qu'on appelle « inaffidabilisme », c'est la possibilité, pour toute théorie scientifique, d'être un jour réfutée,

¹⁷⁴ On peut être étonné qu'une telle croyance mathématique soit jugée *inaffidable* alors qu'il semble clair qu'en un sens important « je pourrais me tromper ». Ce qu'il faut remarquer, c'est qu'au sens de la possibilité logique ou métaphysique, celui-là même qui est utilisé dans le cas des « possibilités d'erreur » sceptiques, il n'est pas vrai que « je pourrais me tromper », tout simplement parce qu'il n'est pas vrai que « le théorème pourrait être faux ». Ma croyance est donc à l'abri de toute *possibilité logique* d'erreur, elle est fiable dans absolument tous les mondes possibles où elle est crue. C'est ce que (Pryor 2006) a appelé une croyance « hyper-fiable » (*hyper-reliable*). Le problème de l'inaffidabilité comme hyper-fiabilité, c'est qu'elle ne semble pas devoir être connectée d'une quelconque manière avec la façon dont nous fondons notre croyance : si je crois le théorème de Gödel sur la base du raisonnement « Gödel est un chat ; or tous les chats ont démontré le théorème de Gödel ; donc le théorème de Gödel est vrai », cela n'empêche pas ma croyance d'être hyper-fiable, alors qu'intuitivement il n'y a aucune connexion (si ce n'est une implication stricte) entre la source de ma croyance et sa vérité. Tout ce que cela montre, c'est que l'inaffidabilisme est vraiment très loin du compte lorsqu'il transforme le réquisit de connexion fait-croyance en un réquisit d'implication stricte. Mais c'est bien cela que fait l'inaffidabilisme, puisqu'il s'intéresse aux possibilités *logiques* de distorsion fait-croyance.

¹⁷⁵ Pour l'exemple de la croyance au théorème de Gödel, cela reviendrait à dire ceci : « ma croyance au théorème est justifiée *en dépit* du fait qu'elle soit révisable, mais elle n'est une connaissance *que parce que* elle est inaffidable ».

¹⁷⁶ Pour l'exemple de l'introspection faillible, cela reviendrait à dire ceci : « ma croyance introspective constitue une connaissance *en dépit* de son caractère faillible, mais elle n'est justifiée *que parce que* elle n'est pas susceptible d'être révisée par la suite. »

ou du moins remplacée ; autrement dit, c'est la possibilité qu'une théorie qui était justifiée à t (qui était la meilleure théorie à t) cesse d'être justifiée à $t+n$. C'est clairement le révisabilisme de la justification qui est ici en jeu, et non pas le faillibilisme de la connaissance.¹⁷⁷ Pourtant, il est clair que la confusion entre les deux notions n'est pas arbitraire et repose sur une connexion qu'on pourrait, en première approche, exprimer ainsi : « si la croyance que p à t peut être *révisée* (quant à sa justification) à $t+1$, c'est bien parce qu'elle *pourrait être fausse* (déjà à t) ». On vient de voir que ce raisonnement n'était pas valide puisque de la faillibilité d'une croyance il ne s'ensuit pas *dans l'absolu* qu'elle soit révisable. C'est-à-dire que *dans l'absolu* une croyance pourrait être menacée par des possibilités d'erreur que nous ne pourrions avoir aucune justification ultérieure à tenir pour actuelles (par exemple, une possibilité d'erreur introspective, qui serait *ex post facto* indétectable). Mais *dans le cas des croyances perceptives* (ou empiriques), les possibilités d'erreur les plus menaçantes ne sont pas, en fait, les possibilités indétectables (telles que les scénarios sceptiques avec martiens et démons) ; les possibilités les plus menaçantes pour la connaissance sont les possibilités d'erreur « courantes », c'est-à-dire celles que nous sommes régulièrement justifiés à tenir pour actuelles (j'ai cru voir que cette tour était ronde ; je vois à présent qu'elle est octogonale). Et par conséquent, si un philosophe faillibiliste veut éviter le scepticisme de la perception, il ne lui suffit pas d'accepter le *faillibilisme faible* :

(faillib. faible) Ma croyance que p peut constituer une connaissance même si elle est vulnérable à des possibilités d'erreur de type sceptique et indétectable.

mais il doit en outre accepter le *faillibilisme fort* :

(faillib. fort) Ma croyance que p peut constituer une connaissance même si elle est vulnérable à des possibilités d'erreur de type courant, i.e. un type d'erreur que j'ai déjà eu l'occasion de détecter pour moi-même.

Comme nos croyances perceptives sont affligées de faillibilité « courante », il faut bien que ce type de faillibilité soit compatible avec la connaissance, si du moins on veut défendre que nos croyances perceptives sont des connaissances ; et de même, il faut bien que ce type de faillibilité soit compatible avec le statut de « croyance justifiée », si du moins on veut défendre que les croyances perceptives sont justifiées. Le cas particulier des croyances perceptives, et de leurs erreurs couramment *observées*, permet de comprendre pourquoi le projet faillibiliste et le projet révisabiliste, quoique logiquement distincts, sont étroitement liés. Lorsqu'on est faillibiliste à propos des connaissances perceptives, il est « naturel » d'être également révisabiliste au sujet de leur justification, car l'irrévisabilisme mène au scepticisme

¹⁷⁷ Il serait compliqué d'entrer ici dans les classifications et définitions de Popper lui-même, d'autant qu'elles admettent une certaine souplesse. Dans (Popper 1977 Add. I.5), Popper désigne par « faillibilisme » la simple prise en compte de la faillibilité de nos croyances empiriques... et non pas la *compatibilité* de cet état de fait avec le statut de connaissance ou de croyance justifiée : « By 'fallibilism' I mean here the view, or the acceptance of the fact, that we may err, and that the quest for certainty (or even the quest for high probability) is a mistaken quest. ». Dans (Popper 1963, 228), en revanche, le faillibilisme est assimilé au falsificationnisme et désigne alors la thèse selon laquelle la révisabilité d'une croyance est non seulement *compatible* avec sa justification mais même *nécessaire* à sa justification : « fallibilists say, roughly speaking, that what cannot (at present) in principle be overthrown by criticism is (at present) unworthy of being seriously considered ». Quelques lignes plus loin, cette thèse « faillibiliste » est ensuite identifiée au refus du « justificationnisme » (c'est-à-dire, dans mon vocabulaire, de l'internalisme) : « we can never give positive reasons which justify the belief that a theory is true ». De ces trois définitions, thèse de la faillibilité de fait des sens, thèse de la révisabilité comme réquisit, thèse de l'absence de justifieur positif, aucune ne correspond à ce que j'ai défini comme faillibilisme, ni même comme révisabilisme, mais indépendant de la terminologie retenue, il est clair que la théorie de Popper défend ce que j'ai appelé révisabilisme.

de la justification (« aucune croyance perceptive n'est justifiée ») tout aussi sûrement que l'inafaillibilisme mène au scepticisme de la connaissance (« aucune croyance perceptive n'est une connaissance »).¹⁷⁸

Ayant refusé l'irrévisabilisme de la justification, il reste à donner une description positive de ce qu'est une théorie révisabiliste de la justification ; et en particulier, il faut expliquer la notion centrale de toute théorie révisabiliste, à savoir la notion de justification *prima facie*. Le propre d'une justification *prima facie*, c'est qu'elle « peut être réfutée ou désamorcée par des données supplémentaires ». ¹⁷⁹ Par conséquent elle permet de rendre compte du processus de *révision* de la justification. Pour mieux comprendre cette notion, on peut l'introduire par un exemple. Imaginons que Rémi voie, à quelque distance, une tour qui lui apparaît assez distinctement comme ronde. Comme il n'a aucune raison de suspecter un problème quelconque, il produit à t0 la croyance que cette tour est ronde. Dans le cadre internaliste fondationnaliste où nous nous situons à présent, la situation sera décrite ainsi : la croyance de Rémi ne viole aucune des normes négatives de justification (non-contradiction, non-circularité, etc.) et en outre, elle est fondée sur un justifieur interne positif, à savoir l'expérience visuelle dont le contenu (ou une partie du contenu) est que « cette tour est ronde ». La croyance de Rémi est donc *justifiée*. Imaginons maintenant qu'après quelques instants, à t1, Rebecca dise à Rémi que, d'après le guide touristique, la tour qui doit se trouver devant eux est une tour octogonale. Comme il n'est pas absolument sûr de sa vue à cette distance, Rémi révisé sa croyance, bien que la tour continue de lui *sembler* ronde. Mais à t2 Rebecca corrige ce qu'elle vient de dire : elle s'est tout simplement trompée d'endroit, et n'a du coup aucune information sur la tour qu'ils sont en train de voir. A la suite de cette nouvelle information, Rémi se remet à croire que la tour est ronde, car son impression visuelle est tout de même assez distincte. Comment le fondationnaliste décrira-t-il ces revirements de croyance (et de justification) ? La première chose à noter, c'est que ces revirements l'obligent à être *révisabiliste*, si du moins il veut dire qu'à chacune des étapes l'attitude épistémique de Rémi soit justifiée ; c'est-à-dire qu'il est obligé d'admettre que le « statut justifié » (justifiedness) d'une croyance peut changer à travers le temps. La proposition « cette tour est ronde » est justifiée à t0, puis sans justification à t1, puis de nouveau justifiée à t2. Mais cela ne nous dit pas exactement comme cette révisabilité interagit avec le fondationnalisme : si le « statut justifié » évolue, qu'en est-il de l'évolution du *justifieur* positif ? Le fondationnaliste dira-t-il que la proposition a un justifieur à t1, n'a plus de justifieur à t2, et acquiert de nouveau un justifieur à t3 ? Il pourrait dire cela, mais ce n'est clairement pas la description la plus naturelle de ce qu'il se passe, dans la mesure où le « justifieur », en l'occurrence, est une expérience visuelle qui existe tout aussi bien à t1 qu'à t0 et à t2. La description la plus naturelle peut être donnée par analogie avec les forces physiques s'appliquant à un pendule : si à t0 j'applique à la boule d'un pendule une force vers la droite, et que rien ne retient la boule, je la déplacerai vers la droite ; si à t1 quelqu'un d'autre vient appliquer vers la gauche une force de même valeur, la boule s'immobilisera, du fait de l'annulation des deux forces. La description admise d'un tel phénomène ne consiste pas à dire que les deux forces ont été

¹⁷⁸ Notez cependant que *même* dans le cas de la perception, on pourrait sans contradiction défendre que la révisabilité commune de nos croyances perceptives ne les empêche pas d'être parfois des connaissances mais les empêche d'être pleinement justifiées, ou inversement qu'elle ne les empêche pas d'être justifiées, mais qu'elle les empêche d'être au sens fort des connaissances. La connexion entre les deux projets tient donc simplement au fait que, en général, un auteur qui voudra défendre les croyances perceptives voudra les défendre *à la fois* comme étant des connaissances et comme étant justifiées.

¹⁷⁹ « *Prima facie* justification can be defeated or undermined by additional evidence. » (Pryor 2000, 534) Je reviendrai un peu plus loin sur mes traductions non habituelles pour *defeated* et *undermined*. Pour d'autres références sur la justification *prima facie*, on peut consulter notamment (Pollock and Cruz 1999, 32), qui introduisent clairement cette notion comme une alternative faible aux justifications irrévisables (*incorrigible justification*). Voir également (Audi 1998, 27–28) et (Engel 2007, chap. V), spécialement p.157 et 162.

supprimées à t_1 : elles continuent de s'appliquer ; mais comme elles s'appliquent conjointement et en sens contraires, elles *s'annulent*. Si par ailleurs la force appliquée vers la gauche est *supérieure* à celle que j'applique vers la droite, cette dernière non seulement sera annulée, mais elle sera contrecarrée. Par analogie, la force justifiante d'un justifieur positif peut être comprise comme une force qui tend à rendre une proposition justifiée ; cette force peut être annulée ou contrecarrée par un justifieur allant en sens contraire, mais lorsqu'elle est annulée ou contrecarrée, cela ne veut pas dire qu'elle n'existe plus. Ainsi, à t_0 , Rémi acquiert un justifieur pour croire que la tour est ronde, et ce justifieur est ensuite contrecarré par un justifieur plus fort en faveur du contenu « la tour n'est pas ronde » ; mais cela ne veut pas dire que le justifieur de départ a disparu, et la meilleure façon de le voir, c'est de constater que cette force justifiante se remet à opérer à t_2 aussitôt qu'on enlève la force justifiante allant en sens contraire. C'est cette « force justifiante » (parfois opérante, parfois annulée, parfois contrecarrée) que les épistémologues ont pris l'habitude d'appeler la « justification *prima facie* ».

La notion de « justification *prima facie* » est parfois considérée comme un outil épistémologique douteux parce qu'il serait trop permissif, trop tolérant, qu'il permettrait de justifier à peu près n'importe quoi. Si l'on reste sur cette impression, alors le fait de défendre l'introspectionnisme du libre arbitre en ayant recours à la notion de justification *prima facie*, ce *deus ex machina* des causes épistémologiques désespérées, pourrait être vu comme un cruel révélateur de la faiblesse de cette position. Mais cette réaction à l'égard de la notion de justification *prima facie* repose sur une mauvaise compréhension de la notion, qui est, lorsqu'on la comprend bien, extrêmement modeste. Pour dissiper cela, il convient donc de faire quelques remarques sur cette notion, en rappelant le peu qu'elle dit, et surtout en soulignant ce qu'elle ne *dit pas*.

Premièrement, la notion de justification *prima facie* est une conséquence directe des deux thèses dont nous venons de voir la très forte plausibilité anti-sceptique : le fondationnalisme internaliste et le révisabilisme. Si on considère que l'impeccabilité interne d'une croyance que p ne peut pas être une simple affaire de contraintes négatives mais suppose un élément positif en faveur de p (argument de l'isolation ou du prisonnier), alors il faut bien admettre quelque chose comme un justifieur positif. Et si par ailleurs, on admet des justifications *révisables* (comme *doit* le faire tout anti-sceptique pour la perception), alors il faut bien admettre que les justifieurs positifs ne rendent une croyance justifiée que de manière révisable. La notion de justification *prima facie* ne présuppose rien de plus radical que cela. On peut en effet la définir ainsi :

S est justifié *prima facie* à croire que p =_{df} S a un justifieur positif à croire que p pouvant procurer de manière révisable une justification à croire que p .

Deuxièmement, soutenir que notre structure justificationnelle (notamment pour nos croyances perceptives) doit contenir des justifications *prima facie* (des justifieurs positifs susceptibles d'être réfutés), cela n'engage pas directement à une certaine théorie sur la nature et le nombre de ces justifications *prima facie*. Et il peut y avoir débat sur la question. On pourrait, en droit, défendre que ces justifieurs positifs ne sont pas des expériences au sens où on l'entend ici, ou que seules *certaines* expériences peuvent jouer le rôle de justification *prima facie*. Compte tenu de la discussion précédente sur le fondationnalisme, je ne remettrai pas en question le fait que les justifieurs soient des expériences. J'aborderai en revanche dans la section suivante la question de savoir *quelles* expériences fournissent une justification *prima facie*. Pour l'instant, je n'ai donc pas montré que l'expérience du libre arbitre fournissait effectivement une justification *prima facie*, j'ai simplement montré que sa faillibilité (ou sa révisabilité) ne l'empêchait absolument pas d'en fournir.

Troisièmement, même si on soutient que toutes les expériences doxastiques fournissent une justification *prima facie*, la notion de justification *prima facie* ne permet pas de « rendre justifié tout et n'importe quoi » pour la simple raison que la justification *prima facie*, en tant que telle, ne nous dit rien du « statut justifié » des croyances. Le statut justifié d'une croyance, dans ce contexte, est généralement appelé « justification *ultima facie* ». Il correspond, si l'on veut, à la résultante de toutes les « forces justifiantes » qui s'appliquent sur le système cognitif du sujet. Et de même qu'un pendule immobile peut être soumis à une multitude de forces sans qu'aucune n'apparaisse finalement dans la résultante, de même il est naturel de penser que nos systèmes cognitifs sont soumis à une multitude de justifications *prima facie* dont la majeure partie s'entre-annulent et disparaissent au niveau de la justification proprement dite. Ce processus d'équilibrage ou d'annulation des justifications *prima facie* est ce qu'on appelle dans le vocabulaire technique le mécanisme des *defeaters*¹⁸⁰ – ce terme est souvent traduit en français par « défaiseurs », je préfère parler de « réfutations » ou de « réfuteurs ».¹⁸¹ Dans la mesure où on prend au sérieux ce processus des réfutations de justifications *prima facie*, on n'a aucune raison de craindre que la théorie soit trop permissive. Supposons, par exemple, qu'un opposant de la justification *prima facie* exprime l'inquiétude suivante : « avec la notion de justification *prima facie*, ne pourrait-on pas défendre que je serais justifié à croire qu'il existe une planète habitée à 200 années-lumières de la terre si subitement cette proposition se mettait à me sembler vraie (à cause d'un dysfonctionnement neuronal) ? Ou que le théorème de Fermat est faux, si cela se mettait à me sembler vrai ? ». On voit bien que ces inquiétudes viennent de la présence d'*objections* à ces croyances, des objections qui empêchent assurément ces croyances d'être justifiées *ultima facie*. Mais une théorie de la justification *prima facie* n'aura évidemment aucun problème à dire que ces deux croyances, quand bien même elles recevraient une justification *prima facie*, sont en présence de réfutations si importantes qu'elles n'ont aucune chance d'être *ultima facie* justifiées. Autrement dit, le souci légitime de ne pas « justifier tout et n'importe quoi », le théoricien de la justification *prima facie* le partage totalement, mais il l'exprime simplement en termes de justification *ultima facie* et de réfutations de justification. En somme, la justification *prima facie* n'est pas une théorie plus permissive qu'une autre, tout simplement parce que la justification *prima facie* n'est qu'une condition *nécessaire* mais pas *suffisante* pour être justifié (*ultima facie*) et que toutes les restrictions et contraintes souhaitables peuvent ensuite s'exprimer dans les *autres* conditions nécessaires, i.e. dans les conditions de non-réfutation. Certes, il fait partie de la notion même de justification *prima facie* que :

en l'absence de données réfutantes [*defeating evidence*], la justification *prima facie* à croire que *p* constituera une justification toutes choses considérées à croire que *p*.
(Pryor 2000, 534)

Mais cela revient simplement à dire que la justification *prima facie* est une condition nécessaire qui est « suffisante si les autres conditions nécessaires sont réunies » (ces autres conditions nécessaires pouvant être très contraignantes).¹⁸² Et on voit mal quelle autre

¹⁸⁰ Sur la notion de *defeater*, voir (Pollock and Cruz 1999, 196 sqq), (Plantinga 2000, chap. 11) et (Bergmann 2006, chap. 6). Plantinga introduit la distinction entre *defeaters* de *warrant* (de la connaissance) et *defeaters* de *rationality* (de la justification) ; il est clair que je ne considère ici que le second type de *defeaters*.

¹⁸¹ En anglais, *defeater* appartient au vocabulaire du combat ; dire qu'une justification est *defeated*, c'est dire qu'elle est vaincue ou battue par une considération plus forte. En français, ce sens ancien de « défaire » ne paraît plus, me semble-t-il, que dans le substantif « une défaite », et l'expression de « défaire une justification » évoque plutôt les travaux de Pénélope que ceux d'Hercule.

¹⁸² Il se trouve que Jim Pryor lui-même pose des restrictions à ce qui peut compter comme réfutation, et les pose comme faisant apparemment partie de sa *définition* de la justification *prima facie* : « I want us to understand

condition nécessaire on pourrait encore ajouter en plus des conditions positives (le justifieur) et des conditions négatives (non réfutation). Dans le cas du libre arbitre, dire que l'expérience du libre arbitre fournit une justification *prima facie* à la croyance au libre arbitre ne permet donc pas de dire que la croyance au libre arbitre est justifiée *tout court* (c'est-à-dire *ultima facie*) pour qui que ce soit ; il se pourrait même que cette justification *prima facie* soit toujours et nécessairement réfutée par d'autres aspects de notre système cognitif.¹⁸³ Une telle thèse est donc extrêmement modeste ; la thèse vraiment intéressante consistera à établir que cette justification *prima facie* est non réfutée. Et c'est cela qui occupera tout le restant de cette étude.

Il convient d'introduire ici une distinction très importante entre deux types de réfuteurs, la distinction entre « rebutting defeaters » et « undercutting (ou undermining) defeaters », que je traduis respectivement par « réfuteur direct » et « réfuteur fondationnel ».¹⁸⁴ Prenons un exemple simple. Dans un tribunal, le témoin n°1 vient dire à la barre qu'il a vu l'accusé tirer sur la victime. Ce premier témoignage donne aux jurés une justification *prima facie* à croire que l'accusé est coupable. Mais cette justification pourrait être réfutée par les témoins suivants, et ce de deux manières différentes. Première possibilité : le témoin n°2 vient dire à la barre qu'il a vu l'accusé sans arme à la main au moment où le coup de feu a retenti. Ce deuxième témoignage apporte une justification *prima facie* en sens contraire, et si cette deuxième justification est suffisamment « forte », elle réfutera la première.¹⁸⁵ Deuxième possibilité : le témoin n°3 vient dire à la barre que le témoin n°1 est un ennemi personnel de l'accusé et qu'il s'est déjà illustré par divers faux témoignages. Dans ce cas également, la justification *prima facie* apportée par le témoignage du témoin n°1 se trouvera réfutée, mais aucun élément d'information n'aura été apporté sur la question de fait (le meurtre). Ce deuxième type de réfuteur ne s'attaque pas directement à la vérité ou à la fausseté de la proposition antérieurement justifiée ; il s'attaque au fondement même de la justification. Pour parler métaphoriquement, il y a deux manières de « faire tomber » une croyance : on peut soit s'attaquer à son contenu (réfuteur direct), soit renverser son support justificationnel (réfuteur fondationnel). Il est crucial de remarquer qu'entre ces deux types de réfuteurs, il y a une différence de « niveau épistémique ». Un réfuteur fondationnel est une information à *propos* d'une autre information. Il s'agit donc d'un réfuteur de « deuxième ordre » qui suppose (chez le sujet) la maîtrise de concepts épistémiques. Un réfuteur direct en revanche est entièrement de « premier ordre » et ne requiert pas de maîtriser d'autres concepts que ceux engagés dans le contenu propositionnel justifié ou

“*prima facie*” and “defeating evidence” in such a way that only ordinary evidence of the sort employed by the man in the street and by the working scientist counts as defeating your *prima facie* justification. » (*ibid.*) Cette présentation des choses est, à mon sens, malheureuse. La compréhension intuitive de l'expression « justification *prima facie* » ne semble pas poser de contrainte sur le genre de considérations qui pourraient réfuter, et il est donc beaucoup plus clair de ne pas résoudre à coup de définitions la question de savoir si les considérations sceptiques réfutent ou non la justification *prima facie* des sens.

¹⁸³ Voici un exemple possible de justification *prima facie* qui serait toujours et nécessairement supprimée au niveau de la justification *ultima facie*. Un épistémologue pourrait défendre que si S a une expérience doxastique selon laquelle « toutes mes facultés cognitives sont défaillantes », alors cette expérience offre une justification *prima facie* en faveur de son contenu. Mais, de manière parfaitement consistante, le même épistémologue pourra également soutenir que dans toutes les situations possibles, la croyance correspondante serait *ultima facie* injustifiée parce que l'acceptation de cette croyance constituerait nécessairement une réfutation de la justification *prima facie*.

¹⁸⁴ Cette distinction a été introduite par John Pollock (voir notamment Pollock and Cruz 1999, 196). Plusieurs auteurs ont proposé des améliorations et subtilités sur cette distinction, notamment (Bergmann 2006, 158–159) qui distingue trois sortes de réfuteurs fondationnels. J'introduirai moi-même un peu plus loin une distinction entre deux types de réfuteurs fondationnels. Pour l'instant, la distinction classique suffira.

¹⁸⁵ Soit en amenant les jurés à renverser leur jugement et à tenir l'accusé pour innocent, soit en les amenant à suspendre leur jugement si les deux témoignages contradictoires sont aussi forts l'un que l'autre.

réfuté.¹⁸⁶ Plus précisément, un réfuteur fondationnel consistera ordinairement à montrer qu'une certaine source n'est pas *fiable* ou qu'elle ne peut pas conférer de *connaissance*. Ce type de réfuteur est donc au cœur des problèmes « d'intégration » ou de « méta-cohérence » dans lesquels un problème externaliste (un défaut de connaissance) se répercute en problème internaliste (en réfutation de justification). C'est donc au chapitre 3 que j'aborderai ce type de réfuteurs ; pour le chapitre présent, je ne considérerai que les questions de justification de *premier ordre*.

Quatrième et dernière remarque pour voir le caractère très modeste d'une revendication de justification *prima facie* : cette notion ne doit pas être confondue avec des notions telles que la justification « provisoire », ou « préthéorique », ou « préréflexive ». Toutes ces notions sont des notions de justification *ultima facie* pour un agent donné dans des circonstances données. On pourrait dire qu'un agent est justifié à croire que *p* « provisoirement », par exemple, lorsqu'il est justifié à *t* à croire que *p*, tout en sachant qu'il va acquérir après *t* d'autres informations susceptibles de réviser sa croyance. Dans la mesure où l'on peut dire à *t* qu'il *est justifié* (justifié « tout court ») à croire que *p*, cette notion est une notion de justification *ultima facie* – une notion qui prend en compte et fait la somme de *toutes* les données pro et contra dont il dispose à *t*. De même, on pourrait dire que les enfants sont justifiés « préthéoriquement » à croire à la physique naïve, tout en sachant que cette croyance ne sera plus justifiée lorsqu'ils apprendront en classe la théorie scientifique de la physique. Là encore, la notion de justification préthéorique est un cas particulier de justification *ultima facie* puisque la croyance peut être dite justifiée « tout court ».¹⁸⁷ Dans le courant de cette étude, je défendrai une thèse de ce genre à propos du libre arbitre : je défendrai que la croyance au libre arbitre est préthéoriquement justifiée (*ultima facie*), c'est-à-dire que la justification *prima facie* en sa faveur ne rencontre pas de réfutation préthéorique (si elle est réfutée, ce devra être par des considérations scientifiques ou philosophiques avancées). Mais dans ce chapitre, je défendrai *seulement* la thèse selon laquelle l'expérience du libre arbitre procure une justification *prima facie*. Cette thèse-là n'oblige pas à penser que qui que ce soit, où que ce soit, ait jamais été justifié à croire au libre arbitre, si brièvement et provisoirement que ce soit. La justification *prima facie* n'a aucune sorte de priorité *chronologique* par rapport à la justification *ultima facie* ; sa priorité est simplement logique.

Il est temps de conclure cette longue discussion sur l'infailibilisme. A la fin de la première section (celle sur l'objection de la « pure impression »), nous avons conclu qu'il

¹⁸⁶ La distinction entre *undercutting defeater* et *rebutting defeater* est parfois tirée dans une autre direction : au lieu d'être une distinction de niveau, elle serait une distinction dans l'*effet* produit par le réfuteur, effet de suspension du jugement antérieur vs effet d'inversion du jugement antérieur. Mais la distinction d'origine est logiquement indépendante de l'effet produit. Un réfuteur direct, comme on vient de le voir dans la note précédente, peut aussi bien produire la suspension que la révision, en fonction du « rapport de force » avec l'autre justification *prima facie*. Inversement, un réfuteur fondationnel, bien qu'il amène le plus souvent à la suspension, peut dans certaines circonstances précises amener à inverser son jugement. Supposez par exemple que Pierre et Jean m'apportent des informations contradictoires sur la proposition *p* ; comme Pierre est habituellement plus fiable, je crois ce qu'il me dit, à savoir *p*. Mais un troisième personnage m'apprend ensuite que Pierre a décidé de faire des farces aujourd'hui, ce qui m'apporte un réfuteur fondationnel contre son témoignage. Son témoignage étant ainsi réfuté, ma réaction sera d'*inverser* mon jugement (et non pas de le suspendre) parce que je dispose toujours du témoignage de Jean pour *~p*.

¹⁸⁷ On pourrait s'étonner que ces notions de justification *ultima facie* soient révisables « comme la justification *prima facie* », mais cet étonnement est une erreur de catégorie : la justification *prima facie* sert à *expliquer* la révision de la justification, mais ce qui est révisé, c'est toujours la justification *ultima facie* ; la justification *prima facie* (le justifieur, ou la force justifiante) est justement ce qui ne change pas lorsque la justification *ultima facie* change. Certes, pour que la justification *ultima facie* change, il faut bien qu'une *certaine* force justifiante soit apparue ou ait disparu, mais lorsque la croyance de Rémi (que la tour est ronde) se trouve réfutée, ce n'est pas le justifieur de *cette* croyance qui a changé, c'est un autre justifieur qui est apparu en faveur de « cette tour n'est pas ronde ».

fallait bien admettre le caractère justifiant de certaines impressions si l'on ne voulait pas être conduit à une régression sceptique des raisons. Mais peut-être n'est-on pas obligé d'admettre un caractère justifiant à *toutes* les impressions ? Peut-être l'expérience du libre arbitre a-t-elle une caractéristique problématique qui l'empêche de compter au nombre des expériences justifiantes ? Mais laquelle ? Dans cette section et dans la suivante, j'envisage deux candidats au titre de caractéristique problématique : l'expérience du libre arbitre aurait le défaut d'être une expérience *faillible*, et celui d'être une expérience *lourde*. Nous venons de voir que la première caractéristique ne pouvait pas soulever un gros problème : la faillibilité de l'expérience du libre arbitre ne poserait un problème de justification que si on acceptait une théorie infaillibiliste de la justification, qui mène tout droit au scepticisme de la perception.

Nous avons vu par ailleurs que l'infaillibilisme de la justification pouvait assez facilement être rejeté. Premièrement, parce que l'intuition infaillibiliste est d'abord et avant tout une intuition anti-erreur à propos de la *connaissance* et non pas à propos de la *justification*. Si on s'en tient à la notion externaliste de connaissance, on peut facilement résister à l'infaillibilisme parce qu'on peut faire droit à l'intuition anti-erreur en la reconduisant à l'intuition anti-hasard, et sans la pousser jusqu'à un principe maximaliste d'élimination de *toutes* les possibilités d'erreur. La faillibilité (non contestée) de l'expérience du libre arbitre ne constitue donc pas une bonne base pour lui refuser le statut de fondement de connaissance.

Deuxièmement, si on revient à la notion de justification qui est proprement notre objet, la faillibilité de l'expérience du libre arbitre ne pose, là non plus, aucun problème. Certes, le principe de méta-cohérence pourrait amener à « internaliser le problème externe », c'est-à-dire que la faillibilité pourrait devenir un problème de justification *si* elle nous obligeait à considérer notre propre croyance au libre arbitre comme n'étant pas une *connaissance*. Mais comme il n'y a pas de problème externe, le principe de méta-cohérence n'a rien ici à « internaliser ». Nous le retrouverons plus loin, pour d'autres problèmes externes que la faillibilité, mais la faillibilité en tout cas n'est pas source de problème.

Enfin, il est vrai que la faillibilité de l'expérience du libre arbitre amène naturellement à considérer la justification correspondante comme *révisable* du point de vue strictement internaliste. Mais la révisabilité peut (et doit) être assumée dans toute théorie non sceptique de la justification perceptive, au moyen de la notion de justification *prima facie*. La révisabilité de l'expérience du libre arbitre ne peut donc pas nous amener à refuser à cette expérience le statut d'expérience justifiante. Si on lui refuse ce statut, ce doit être à cause d'une *autre* caractéristique jugée problématique. À défaut d'un autre problème, l'introspectionniste aura toutes les raisons de soutenir que l'expérience du libre arbitre procure une justification *prima facie* en faveur de son contenu, au même titre que toute autre expérience perceptive. Nous allons voir à présent une autre raison de ne pas attribuer à l'expérience du libre arbitre tout statut justifiant, même celui de justification *prima facie*.

II.3) L'objection de l'impression lourde

À la fin de la section I) de ce chapitre, j'ai dit que les objections de l'impression faillible et de l'impression lourde reposaient toutes les deux sur la « possibilité » des scénarios d'erreur, mais dans deux sens différents de « possibilité ». Pour l'objection de l'impression faillible, c'est la possibilité *logique* des scénarios d'erreur qui pose problème. Mais nous avons vu que le problème était tout apparent : il n'y a rien de problématique à accepter à la fois « S est justifié à croire qu'il y a une table devant lui par son expérience visuelle » et « il est métaphysiquement possible que S ait cette expérience visuelle et qu'il n'y ait pas de table

devant lui ». Une des raisons pour lesquelles cette possibilité ne pose aucun problème à la justification, c'est qu'il s'agit d'une possibilité *externe* et que la justification est une notion *interne*. Mais dans la mesure où la possibilité *épistémique* est une notion interne au système cognitif du sujet, il est possible qu'elle pose un problème beaucoup plus redoutable que la possibilité métaphysique. En ce sens, peut-être que la possibilité *épistémique* d'une erreur (par exemple par manipulation martienne) empêche l'expérience du libre arbitre de fournir une justification *prima facie* ; ou peut-être qu'elle *réfute* automatiquement la justification *prima facie* apportée par cette expérience.¹⁸⁸ La caractéristique problématique de l'expérience du libre arbitre serait alors le fait qu'elle « engage » à trop de choses, qu'elle suppose l'acceptation *en interne* de propositions que nous ne sommes pas justifiés à exclure (à savoir les négations de scénarios d'erreur). Autrement dit, l'expérience du libre arbitre serait trop « lourde » (de contenu) pour fournir une justification *prima facie*, ou en tout cas pour que la justification *prima facie* puisse jamais se transformer en justification *ultima facie*. Seules les expériences « légères », anodines, qui n'engagent à rien (*il y a une table rouge*), pourraient fournir une justification *prima facie*.¹⁸⁹

Evidemment, quand on considère le concept de « libre arbitre » comme « trop lourd » pour pouvoir être justifié par l'expérience, on pense le plus souvent à l'*indéterminisme*, ou plus généralement à certaines conséquences lourdes qui résultent d'une certaine *analyse* du concept de libre arbitre : l'expérience du libre arbitre ne fournirait pas de justification à son contenu *parce que* ce contenu implique l'*indéterminisme* et serait donc trop lourd. Mais pour l'instant, nous nous situons encore dans le cadre de l'introspectionnisme « a-compatibiliste », c'est-à-dire en amont de toute analyse controversée du libre arbitre, et par conséquent, les seules possibilités épistémiques susceptibles de mettre en cause la justification par expérience à ce stade seront les scénarios *indéniablement* incompatibles avec le libre arbitre, quelque analyse qu'on en donne. C'est bien le cas du scénario des dispositifs martiens, imaginé par Peter van Inwagen. Voici la description précise du scénario telle qu'elle est donnée par van Inwagen :

(M) A chaque fois que naît un être humain, les martiens implantent dans son cerveau un petit dispositif – un dispositif indétectable par toutes les techniques d'observation dont nous disposons, bien qu'il ne soit pas indétectable *en principe* – qui contient un « programme » pour toute la durée de la vie de la personne : chaque fois que cette personne doit prendre une décision, le dispositif *cause* sa décision dans un sens ou dans l'autre en fonction des indications d'une table d'instructions qui a été incorporée dans la structure du dispositif avant que la personne ne soit conçue.
(van Inwagen 1983, 109)

Il est clair qu'*aucune* notion de liberté, si compatibiliste soit-elle, ne peut admettre ce genre de manipulation martienne ; si donc cette possibilité est épistémiquement possible pour nous, elle pose un problème pour toute forme d'introspectionnisme du libre arbitre, qu'il soit compatibiliste ou incompatibiliste. Si la vérité de ce scénario est pour nous une possibilité (épistémique) totalement ouverte, comment pourrions-nous donc être justifiés à affirmer que nous sommes libres ? Peut-on affirmer une proposition et laisser en suspens ses conséquences logiques ?

¹⁸⁸ Pour bien faire la distinction entre ces deux possibilités, il faut se rappeler qu'une justification *prima facie* ne « disparaît » pas lorsqu'elle est réfutée ou « défaite » : dans de telles circonstances, on dira que la croyance que *p* dispose *toujours* d'une justification *prima facie*, mais que, à cause du réfutateur, cette justification *prima facie* ne rend pas la croyance que *p* justifiée *ultima facie*.

¹⁸⁹ Ou en tout cas une justification *prima facie* suffisamment résistante pour résulter parfois en une justification *ultima facie*.

Nous allons voir une nouvelle fois que cette objection, si forte soit-elle, échoue à faire une distinction entre l'expérience du libre arbitre et les expériences perceptives en général, et qu'elle ne pose donc pas de problème particulier à l'introspectionnisme du libre arbitre.

II.3.a. L'objection et son principe de clôture

Comme j'ai introduit l'objection en termes de « possibilité épistémique », il convient de dire tout d'abord quelques mots sur cette notion.¹⁹⁰ On introduit généralement la notion de possibilité épistémique en faisant remarquer une ambiguïté dans les expressions de la forme « il est possible que p » (ou « p pourrait être le cas »). L'ambiguïté apparaît notamment si on compare les exemples suivants :

- (a) il est possible que César ne passe pas le Rubicon.
- (b) il est possible que la conjecture de Goldbach soit fausse.

Les énoncés (a) et (b) seront tous les deux corrects en un certain sens, mais clairement pas dans le même sens. L'énoncé (a) est correct dans le sens où l'on peut représenter ou concevoir un scénario où César ne passe pas le Rubicon, sans violer de loi logique ni de principe métaphysique. Mais tout en imaginant cette possibilité *métaphysique*, on sait pertinemment (si l'on est un minimum informé) qu'elle n'est pas le cas. En revanche, quand on énonce la phrase (b), on rapporte ou on exprime un manque d'information à propos de la conjecture de Goldbach : on ne sait pas si elle est vraie ou fausse ; ou plutôt, on ne se considère pas comme justifié à rejeter la possibilité que la conjecture soit fausse... mais cela ne veut pas dire qu'il y a là la moindre « possibilité métaphysique » puisqu'il s'agit d'un énoncé mathématique : s'il est vrai, il est métaphysiquement nécessaire ; s'il est faux, il est métaphysiquement impossible. La « possibilité » dont il s'agit ne se situe donc pas « dans la réalité », si l'on peut dire, mais plutôt dans le système cognitif du sujet : dire que la fausseté de la conjecture de Goldbach est possible pour S, c'est dire qu'elle n'est pas totalement exclue de son système cognitif (quand bien même elle serait métaphysiquement impossible). Telle est la notion de « possibilité épistémique ». Peut-on la définir plus précisément, en explicitant ce que signifie « être exclu du système cognitif de S » ? Il n'y a malheureusement pas de consensus sur cette question : certains considéreront que l'énoncé « il est possible que $\sim p$ » est correct seulement si S est *agnostique* sur la question, c'est-à-dire si $\sim p$ est compatible avec tout ce qu'il *croit*.¹⁹¹ D'autres diront que « il est possible que $\sim p$ » exprime l'*incertitude* de S à l'égard de p , tout en étant compatible avec la croyance que p .¹⁹² Pour d'autres, une proposition est épistémiquement possible pour S ssi elle est compatible avec ce qu'il *sait*.¹⁹³ Une dernière

¹⁹⁰ Ma présentation de la notion de possibilité épistémique est très proche de (Hawthorne 2004, 24–28), qui comme moi s'intéresse à cette notion dans son rapport au problème de la « clôture épistémique ». Pour une approche plus générale, voir notamment (DeRose 1991) et les références données par Hawthorne en note 70.

¹⁹¹ Une telle interprétation s'appuie sur l'étrangeté d'énoncés de la forme « p mais il est possible que $\sim p$ ». Si l'on estime qu'une assertion que p est fondamentalement l'expression de la *croyance* que p , alors on est porté à dire que « il est possible que $\sim p$ » est nécessairement inapproprié dès que l'on croit que p .

¹⁹² Il y a encore deux versions de cette thèse selon qu'on accepte ou non les énoncés de la forme « p mais il est possible que $\sim p$ ». Si on les refuse, alors on peut dire que l'énonciation de p est fondamentalement l'expression d'une *certitude* sur p et que c'est *pour cela* qu'elle est incompatible avec l'énonciation de « il est possible que $\sim p$ », prise comme expression de l'incertitude. Mais on pourrait aussi défendre que l'assertion de p est correcte en cas de croyance incertaine, et donc compatible avec l'assertion d'incertitude « il est possible que $\sim p$ ». Voir (Hawthorne 2004, 28).

¹⁹³ C'est l'interprétation privilégiée par (Hawthorne 2004, 26), définitions (1) à (4). Cette interprétation peut de nouveau reposer sur l'étrangeté de « p et il est possible que $\sim p$ » si on interprète l'assertion de p comme une *expression de connaissance* que p . Hawthorne peut clairement défendre cette voie, compte tenu du lien qu'il

interprétation possible consiste à dire que p est épistémiquement possible pour S ssi rien d'interne au système cognitif de S ne permet de croire que $\sim p$, c'est-à-dire si S n'est pas *justifié* à croire que $\sim p$. C'est cette dernière interprétation qui nous intéressera ici, car c'est dans ce sens que la « possibilité épistémique » d'une manipulation martienne peut poser un problème *strictement internaliste* à la croyance au libre arbitre.

La première prémisse de l'objection consiste à faire observer que c'est, pour nous, une possibilité épistémiquement ouverte que nous soyons les victimes d'une Manipulation Martienne (MM). Cette prémisse signifie donc que nous ne sommes pas *justifiés* à croire que nous ne sommes pas dans un tel scénario. Pourquoi ? Tout simplement parce que nous ne *pouvons pas* être justifiés à croire cela : il fait partie du scénario de van Inwagen que chaque dispositif martien est « indétectable par toutes les techniques d'observation dont nous disposons ». On peut formaliser cette prémisse de la manière suivante :

$$(1) \sim J \sim MM$$

La conclusion finale de l'objection, c'est bien évidemment que nous ne sommes pas justifiés à croire au libre arbitre (sur la base de l'expérience que nous en faisons). Autrement dit :

$$(C) \sim J LA$$

Et la raison pour laquelle l'absence d'une justification pour $\sim MM$ entraîne une absence de justification pour LA , c'est que le libre arbitre est logiquement incompatible avec le scénario de manipulation martienne. Autrement dit :

$$(2) LA \Rightarrow \sim MM$$

Le *principe* sur lequel repose l'objection est maintenant évident, compte tenu des prémisses et de la conclusion : il s'agit du principe qu'on appelle la *Clôture Epistémique* ; ou plus précisément, il s'agit du principe de la *Clôture de la Justification*, c'est-à-dire un principe précis à l'intérieur de la grande famille des principes de clôtures épistémiques. Dans sa version la plus simple, la clôture de la justification énonce que, si je suis justifié à croire que p , alors je suis justifié à croire toute conséquence logique de p .

$$(J\text{-Clot.}) [Jp \wedge p \Rightarrow q] \Rightarrow Jq$$

ou en contraposée :

$$(J\text{-Clot.}) [\sim Jq \wedge p \Rightarrow q] \Rightarrow \sim Jp$$

ou encore :

$$(J\text{-Clot.}) [p \Rightarrow q] \Rightarrow [Jp \Rightarrow Jq]$$

On dit qu'un opérateur (en l'occurrence la « justification ») est « clos » sous l'implication lorsque toute relation d'implication entre deux propositions p et q se répercute en une relation entre les mêmes propositions élevées par l'opérateur, soit Jp et Jq , ainsi que la troisième notation le met en évidence. Plus concrètement, un principe de clôture dans le domaine épistémologique est un principe qui met en relation les statuts épistémiques de plusieurs *croyances*, de plusieurs « parties de l'espace doxastique » : pour que ma croyance que p ait tel statut, il est nécessaire qu'ailleurs dans mon espace doxastique une autre croyance ait le même statut. On pourrait donc également le décrire comme un « principe de transfert » de statut

établit entre assertion et connaissance, à la suite de Williamson (voir Hawthorne 2004, 21–24). Mais il offre aussi un argument plus direct, partant de l'étrangeté de « je sais que p et il est possible que $\sim p$ ».

épistémique au sein de l'espace doxastique. Il s'agit donc d'un type de principe assez différent du principe inférentialiste (qui s'intéresse principalement à la *fondation* de la croyance que p) ou du principe irrévisabiliste (qui s'intéresse à la solidité diachronique du statut épistémique de p). Nous verrons par la suite un autre principe de transfert épistémique. Pour l'instant, tenons-nous en à la clôture elle-même, c'est-à-dire au transfert de la justification d'une proposition à ses conséquences logiques.

Deux remarques s'imposent pour mieux caractériser l'objection anti-introspectionniste de l'impression lourde et son utilisation du principe de clôture. Premièrement, le principe auquel il est ici fait appel est un principe qui concerne la *justification* et non pas la *connaissance* : dans la littérature épistémologique, les discussions sur le principe de clôture concernent le plus souvent la notion de connaissance. Les discussions sur la clôture de la justification (ou la clôture d'une notion internaliste de « connaissance ») utilisent plus souvent la terminologie de la « transmission » de justification ou de garantie.¹⁹⁴ Dans la mesure où nous nous intéressons ici à la *rationalité* ou à l'impeccabilité *interne* de la croyance introspective au libre arbitre, c'est seulement la clôture de la justification qui nous intéresse.¹⁹⁵ Et donc la première prémisse consiste bien à dire que la croyance à $\sim$ MM n'est pas *justifiée* (à cause de l'indéfectibilité interne des dispositifs martiens) et non pas qu'elle n'est pas une *connaissance*.

Deuxièmement, il faut sans doute apporter un léger correctif pour que l'objection soit plus forte : le principe de clôture *sous l'implication stricte* est trop fort pour être acceptable, car il implique que toute croyance vraie mathématique est nécessairement justifiée. En effet, puisque les vérités mathématiques sont nécessaires, elles sont logiquement impliquées par n'importe quelle proposition, par exemple la croyance que « j'existe » ; il s'ensuivrait donc que si je suis justifié à croire que j'existe, alors je suis justifié à croire toutes les vérités mathématiques (y compris des vérités mathématiques dont je n'aurais jamais vu la démonstration et que je me mettrais à croire par décision arbitraire). Autrement dit, ce n'est pas l'implication logique *en elle-même* qui semble importante pour le transfert de justification, mais plutôt la *conscience* que le sujet a d'une implication logique entre p et q . Une façon traditionnelle de régler ce problème est donc de dire que la justification (ou, plus souvent, la connaissance) n'est pas close sous l'implication stricte, mais sous l'implications stricte *justifiée* (ou, plus souvent, *connue*). On obtient alors le principe suivant :

(J-Clot.*) $[Jp \wedge J(p \Rightarrow q)] \Rightarrow Jq$
ou en (semi-)contraposée :
(J-Clot.*) $[\sim Jq \wedge J(p \Rightarrow q)] \Rightarrow \sim Jp$
ou encore :
(J-Clot.*) $J[p \Rightarrow q] \Rightarrow [Jp \Rightarrow Jq]$ ¹⁹⁶

¹⁹⁴ Sur la notion de transmission de garantie, voir notamment (Wright 2000; Pryor 2007). Comme le fait remarquer (Engel 2007, 159), le principe de transmission de garantie n'est pas directement et trivialement identifiable au principe de clôture épistémique. La terminologie de Wright oblige à faire quelques aménagements. Mais au-delà des aménagements terminologiques, il est clair que c'est bien le même problème qui est en jeu : Duncan Pritchard (2005, 93–95) n'hésite pas d'ailleurs pas à interpréter le diagnostic de Wright (d'échec de la transmission de garantie) comme un constat « d'échec d'une version internalisée de la clôture », et il reformule ainsi cette clôture internalisée : « For all S, ϕ, ψ , if S is internalistically justified in believing ϕ , and S knows that ϕ entails ψ , then S is internalistically justified in believing ψ . » (Pritchard 2005, 107).

¹⁹⁵ En tout cas, c'est la clôture interne qui nous intéresse *directement* : la clôture externe pourrait aussi, de manière indirecte, amener un problème de justification, mais à elle seule, elle permet seulement de conclure à un problème externe. Pour produire un problème indirect, il faudrait ajouter un principe qui répercute les problèmes externes en problèmes internes, c'est-à-dire un principe de métacohérence dont je discuterai au chapitre 3.

¹⁹⁶ Ce principe a un avantage considérable du point de vue formel : il s'agit en effet du fameux principe K de la logique modale, c'est-à-dire le principe le plus fondamental de *toute* logique modale (ou en tout cas de toute logique modale *normale*). Voir (Hughes and Cresswell 1996).

Bien évidemment, si l'on a recours à ce principe plus faible, il faut renforcer la prémisse (2) : il ne suffit plus que le libre arbitre *implique logiquement* l'absence de dispositif martien, il faut également que nous soyons *justifiés* à croire à cette implication logique. Mais ce renforcement ne pose aucun problème particulier. Toute personne qui réfléchit un minimum à la question sera prête à accorder non seulement (2) mais également :

$$(2^*) J [LA \Rightarrow \sim MM]$$

L'argument de l'impression lourde se présente donc sous la forme suivante :

Argument anti-introspectionniste de l'impression lourde

(J-Clot.*) $[\sim Jq \wedge J(p \Rightarrow q)] \Rightarrow \sim Jp$

(1) $\sim J \sim MM$

(2*) $J [LA \Rightarrow \sim MM]$

(C) $\sim J LA$

C'est-à-dire en langage naturel :

Argument anti-introspectionniste de l'impression lourde

(J-Clot.*) *S* est justifié à croire que *p* seulement si *S* est justifié à croire à toutes les conséquences logiques qu'il reconnaît à *p* de manière justifiée.

(1) Nous n'avons aucune justification internaliste à rejeter le scénario de manipulation martienne.

(2*) Nous sommes justifiés à croire que le libre arbitre est incompatible avec le scénario de manipulation martienne.

(C) Nous ne sommes pas justifiés (par notre expérience interne) à croire au libre arbitre.

L'expérience du libre arbitre échouerait donc à justifier son contenu au motif que ce contenu aurait des conséquences trop lourdes à justifier, et qu'on se doit de toujours justifier les conséquences de *p* lorsqu'on veut justifier *p*.

Ce type de raisonnement est souvent utilisé contre le libertarianisme, à cause de la conséquence indéterministe qu'entraîne une conception libertarienne du libre arbitre. Cette conséquence indéterministe serait trop lourde pour être justifiée ainsi par simple expérience. La justification par simple expérience ne pourrait fonctionner que pour les propositions légères et anodines, telles que « il y a une table en face de moi ». Mais on voit avec le scénario martien de Peter van Inwagen que l'argument de l'expérience lourde peut être construit en amont de toute analyse du libre arbitre : quel que soit le contenu exact de notre expérience d'être libres, ce contenu est assurément incompatible avec les scénarios martiens ; ce contenu est donc « lourd » pour tout le monde, compatibiliste ou incompatibiliste. Et c'est pourquoi, d'après van Inwagen, l'introspectionnisme se trouve réfuté dès l'étape a-compatibiliste.

L'objection de l'impression lourde a assurément une certaine force contre l'introspectionnisme a-compatibiliste. Mais ainsi que je l'ai montré dans (Guillon 2014), le grand problème est qu'elle a exactement la *même* force contre les expériences perceptives : les expériences perceptives sont « lourdes » également dans la mesure où elles impliquent que « je ne suis pas un cerveau dans une cuve » ou que « je ne suis pas manipulé par un malin génie », et elles sont par là vulnérables à l'argument sceptique de la clôture.

II.3.b. L'argument sceptique de la clôture

L'argument sceptique de la clôture est probablement l'argument sceptique le plus important dans la littérature contemporaine : la plupart des reconstructions de l'intuition sceptique par scénarios cartésiens (malins génies, cerveaux dans des cuves, etc.) font appel aujourd'hui, d'une manière ou d'une autre, à un principe de clôture. Mais le plus souvent, le principe utilisé est un principe de clôture de la connaissance. Avant de passer à l'argument anti-justification, il faut donc présenter un peu l'argument plus classique anti-connaissance. Je montrerai ensuite pourquoi il est possible, et même plus pertinent, d'utiliser cette même structure d'argument contre la *justification* plutôt que contre la *connaissance*.

L'argument sceptique de la clôture de la connaissance prend, dans sa version la plus simple, la forme suivante :

Argument sceptique de la clôture de connaissance

(K-Clot.) Si S sait que p , et S sait que $p \Rightarrow q$, alors S sait que q .

(1) Nous ne savons pas (nous ne pouvons pas savoir) que nous ne sommes pas des cerveaux dans des cuves (Hypothèse Sceptique HS).

(2) Nous savons que si nous avons deux mains (proposition Anodine A), alors nous ne sommes pas des cerveaux dans des cuves ($\sim$ HS)

(C) donc nous ne savons pas que nous avons deux mains.

C'est-à-dire en version formalisée :

Argument sceptique de la clôture de connaissance

(K-Clot.) $[\sim Kq \wedge K(p \Rightarrow q)] \Rightarrow \sim Kp$

(1) $\sim K \sim HS$

(2) $K [A \Rightarrow \sim HS]$

(C) $\sim K A$

Pourquoi cette forme d'argument a-t-elle pris tellement d'importance dans les dernières décennies ? Principalement à cause des discussions sur l'infailibilisme. Initialement, il était assez naturel de supposer que « le » problème posé par les scénarios sceptiques était simplement qu'ils représentaient des « possibilités (logiques) d'erreur », et donc la formalisation de « l'argument » sceptique reposait sur le principe infailibiliste... c'est-à-dire sur un principe qu'il semble assez aisé et intuitif de refuser. Ce diagnostic permettait une victoire facile contre le sceptique... une victoire trop facile, pouvait-on pressentir. Ce pressentiment a trouvé son expression dans le principe de clôture : grâce à ce principe, les scénarios sceptiques permettent de mettre en péril la connaissance commune *sans* recourir à un principe aussi fort que l'infailibilisme. Et l'argument sceptique ainsi construit est donc beaucoup plus convaincant.

La conclusion de l'argument de clôture de la connaissance correspond à ce qu'on appelle parfois le « scepticisme académique », c'est-à-dire l'affirmation positive que nous ne *connaissions pas* les vérités anodines que nous croyons habituellement connaître.¹⁹⁷ Cette conclusion, en tant que telle, ne conteste pas que nous *soyons justifiés* à croire à ces vérités anodines. Et elle ne nous oblige donc pas à changer quoi que ce soit à nos croyances

¹⁹⁷ Ce type de scepticisme s'oppose traditionnellement au scepticisme « pyrrhonien » qui repose fondamentalement sur l'attitude du doute, et n'entend pas *affirmer* que nous ne connaissons pas ceci ou cela (voir Klein 2011, sec. 2).

communes, ou en tout cas à nos croyances par expériences perceptives.¹⁹⁸ Autrement dit, elle concerne seulement le statut externe de nos croyances communes, pas leur rationalité ou leur statut interne. Pour cette raison, un certain nombre d’auteurs (notamment Pritchard 2005, 7; Pryor 2000, 523) ont estimé que l’argument de clôture de la connaissance manquait la substance proprement internaliste ou évidentialiste de l’intuition sceptique fondée sur les scénarios cartésiens. Ces scénarios posent d’abord et avant un problème interne à nos croyances perceptives. Pour mettre en évidence cette portée sceptique plus radicale que le scepticisme académique, il faut donc construire un argument dont la conclusion soit anti-justification. C’est précisément ce que permet de faire l’argument de la clôture de la justification.

L’argument sceptique de la clôture de la justification, présenté notamment par (Pritchard 2005, 106–107), est le parallèle exact de notre argument anti-introspection de l’impression lourde.

Argument sceptique de la clôture de justification

(J-Clot.*) $[\sim Jq \wedge J(p \Rightarrow q)] \Rightarrow \sim Jp$

(1) $\sim J \sim HS$

(2) $J [A \Rightarrow \sim HS]$

(C) $\sim J A$

Comme on vient de le voir, un premier intérêt de cet argument par rapport à l’argument anti-connaissance tient à sa *conclusion*. Mais cet argument semble aussi plus pertinent du fait de ses *prémisses*, en particulier de (J-Clot*) et de (1).¹⁹⁹ En effet, ces prémisses, outre qu’elles mènent à une conclusion plus forte, semblent aussi plus proches du raisonnement sceptique lorsqu’on les présente en version « justification » qu’en version « connaissance ». Pour le dire autrement, l’intuition de clôture, tout comme l’intuition d’indéfectibilité des hypothèses sceptiques, semblent être plutôt à l’origine des intuitions internalistes qu’externalistes.

Sur le caractère internaliste de l’intuition d’indéfectibilité (1), Pritchard fait la remarque suivante :

Jusqu’à présent, nous avons présenté le problème de la clôture en notant qu’il semble impossible de connaître les négations d’hypothèses sceptiques. Mais à l’origine, le problème est plutôt qu’il est impossible d’être *internalistiquement justifié* à croire aux négations d’hypothèses sceptiques. S’il n’y a, *ex hypothesi*, aucune différence phénoménologique disponible au sujet, qui puisse lui indiquer qu’elle n’est pas une victime de ce scénario, alors il s’ensuit qu’il n’y aura rien d’accessible réflexivement au sujet qui puisse suffire à lui indiquer que sa croyance dans cette proposition anti-sceptique est vraie. (Pritchard 2005, 44)

La manière même dont les scénarios sceptiques sont construits (la stipulation d’une indiscernabilité *interne* entre l’expérience du cerveau encuvé et celle du cerveau incarné) semble révéler une visée foncièrement internaliste. On peut contester (comme je le ferai un peu plus loin) la *priorité* que donne ici Pritchard à l’indéfectibilité de justification ($\sim J \sim HS$) sur l’indéfectibilité de connaissance ($\sim K \sim HS$), mais il est au moins évident que les scénarios sceptiques peuvent très naturellement motiver un problème de *justification* pour $\sim HS$.

¹⁹⁸ Elle oblige, certes, à réviser les quelques croyances de second ordre que nous pouvons avoir dans la vie commune.

¹⁹⁹ La prémisses (2) est tout aussi triviale dans la version anti-connaissance et dans la version anti-justification.

Pour ce qui est de la prémisse de clôture (J-Clot*), une de ses motivations principales semble également être fondamentalement internaliste (ou en tout cas, beaucoup plus forte en version internaliste). Je pense ici au « problème de l'inconséquent », que Hawthorne présente de la manière suivante :

Supposez que Q est une conséquence « lourde » de P et S sait que P et sait également que P implique logiquement Q. Je demande à S si elle est d'accord avec P. Sa réponse est affirmative : « Oui », dit-elle. Puis je demande à S si elle a conscience que Q s'ensuit logiquement de P. « Oui », dit-elle. Puis je lui demande si elle est d'accord avec Q. « Je ne suis pas d'accord avec ça », dit-elle. Je lui demande si elle souhaite à présent revenir sur une de ses déclarations précédentes. « Oh non », dit-elle. « Je maintiens ma déclaration que P, et ma déclaration que P implique logiquement Q. Simplement, je me refuse à soutenir que Q. » Notre interlocuteur ressemble très exactement à la Tortue de Lewis Carroll, ce célèbre objet de dérision qui était tout à fait prête à accepter les prémisses d'un argument en *modus ponens* mais se refusait à accepter la conclusion. (Hawthorne 2005, 32)

L'agent S d'un tel scénario présente une forme d'inconséquence qui, intuitivement, est inacceptable quels que soient les termes P et Q. On pourrait décrire ce problème de l'inconséquent en disant qu'il révèle la « clôture de l'engagement d'assertion » : si je m'engage (et reste engagé) à asserter P, et à asserter $P \Rightarrow Q$, il semble que je suis *de facto* rationnellement engagé à asserter Q. Dans le chapitre d'où ce texte est tiré, Hawthorne utilise l'argument de l'inconséquent pour défendre une clôture de la *connaissance*, et pour ce faire il ajoute à son raisonnement une prémisse qui fait le lien entre assertion et connaissance : le principe de la Connaissance comme Norme de l'Assertion.²⁰⁰ Mais qu'il soit possible ou non de passer ainsi à la clôture de la connaissance, il est clair que le problème de l'inconséquent repose *d'abord* sur une contrainte de rationalité interne au « système d'assertions » du sujet considéré... et plus généralement sur une contrainte de justification des *croyances*, qu'elles soient extérieurement assertées ou non.²⁰¹

L'argument de la clôture de la justification constitue donc un outil au moins aussi important dans l'attirail sceptique que l'argument plus courant (et pas toujours clairement distingué) de la clôture de la connaissance. Mais ce qui nous importe ici, c'est surtout qu'il constitue un argument exactement parallèle à l'argument anti-introspection de l'impression lourde, reposant sur le scénario Inwagenien de manipulation martienne. En effet, les scénarios sceptiques de cerveaux dans des cuves ou de malin génie sont tout aussi indétectables que le scénario de manipulation martienne, et ce pour les mêmes raisons (... c'est-à-dire pour la seule raison qu'on a *stipulé* l'indétectabilité dans le scénario). Par conséquent, si on soutient que la possibilité de manipulation martienne empêche à l'expérience du libre arbitre de fournir une justification à son contenu, *alors* on devra également soutenir que les expériences perceptives sont trop lourdes pour fournir une justification.

On ne peut donc pas faire une discrimination entre expérience du libre arbitre et expériences perceptives au motif que la première serait lourde et les secondes anodines : les expériences perceptives *sont* lourdes, et c'est même la source du plus célèbre des arguments

²⁰⁰ Ce principe est de plus en plus discuté ces dernières années, notamment depuis la discussion qu'en a donné Williamson (2000, chap. 11). Une défense très développée avait également été proposée, dès les années 70, par Unger (1975, chap. VI).

²⁰¹ Supposez que le sujet S se dise par devers soi « non, je ne suis pas d'accord avec Q, mais pour faire plaisir à mon interlocuteur, je vais faire comme si je le croyais » et qu'en conséquence S affirme « oui, en effet, j'accepte Q » : il semble évident qu'on aurait exactement la même intuition d'inconséquence bien qu'elle ne se manifeste plus au niveau des assertions extérieures.

sceptiques. Donc si on refuse aux expériences lourdes le statut justifiant, on ruine toute perception en même temps que l'introspectionnisme du libre arbitre.

II.3.c. Solution : transmission de justification *prima facie* par déduction compétente

Nous allons voir à présent comment on répond à l'argument sceptique de la clôture, et comment cette réponse s'applique à la défense de l'introspectionnisme. Mais avant cela, il faut clarifier un point que j'ai laissé en suspens concernant la conclusion exacte de l'argument sceptique de clôture : le sceptique entend-il montrer par cet argument que la lourdeur des expériences perceptives les empêche de conférer *la moindre* justification, ou bien, plus modestement, que cette lourdeur empêche la justification *prima facie* qu'elles confèrent de se « transformer » en justification *ultima facie* ? En un sens, la question est purement théorique : une justification *prima facie* qui resterait systématiquement virtuelle n'aurait pas beaucoup d'intérêt, et la gravité de la conclusion sceptique serait exactement la même dans les deux cas. Mais pour faciliter la discussion qui suit, je retiendrai la seconde lecture : en effet, si on admet que les expériences lourdes ne confèrent *aucune* sorte de justification, pas même *prima facie*, alors le sceptique gagne de manière triviale et directe, et la discussion ne pourra être qu'en tout ou rien. En revanche, si on dit que ces expériences fournissent une justification *prima facie* qui risque d'être réfutée par la lourdeur du contenu, alors on pourra discuter des conditions exactes de réfutation d'une croyance, voir si elles sont réunies ou non.²⁰² Une bonne partie de la discussion qui suit reviendra donc à réfléchir sur les conditions de réfutation des justifications *prima facie*. Parallèlement, pour l'argument anti-introspectionniste de l'expérience lourde, je l'interpréterai comme un argument qui prétend mettre en évidence un *réfuteur* de la justification *prima facie* conférée par l'expérience du libre arbitre, et je montrerai qu'on ne doit pas accepter le genre de réfuteur mis en évidence si l'on veut éviter le scepticisme.

Je classerai les réponses anti-sceptiques à l'argument de clôture en quatre catégories : les solutions contextualistes, les solutions qui refusent toute forme de clôture, les solutions qui acceptent une forme faible de clôture mais refusent la transmission de justification, enfin les solutions qui acceptent une forme de transmission de justification. Comme pour les objections précédentes, l'introspectionnisme pourrait répondre à l'objection de l'impression lourde en utilisant n'importe laquelle de ces solutions anti-sceptiques, mais je défendrai ultimement une version de la dernière catégorie, qui offre le cadre le plus adapté pour envisager les objections du chapitre suivant.

Nous avons déjà rencontré le contextualisme à propos de l'argument sceptique infailibiliste. Son utilisation est légèrement différente dans la discussion de l'argument de clôture, mais il reste non pertinent pour notre étude (et pour les mêmes raisons que celles exposées p.134). Le contextualisme sert généralement à réfuter les versions anti-connaissance de l'argument de clôture, en faisant observer que, s'il y a plusieurs sens (plus ou moins exigeants) de « connaître », $K_1, K_2, \dots K_n$, on peut soutenir que l'intuition anti-sceptique KA et l'intuition d'indéfectibilité $\sim K \sim HS$ reposent sur des sens différents de connaître, par

²⁰² Attention : la croyance qui a une justification *prima facie*, et qui est peut-être réfutée, c'est la croyance A elle-même, et pas le conséquent $\sim HS$. Savoir si le conséquent $\sim HS$ a aussi une justification *prima facie* en sa faveur est une question beaucoup plus complexe qui mérite argumentation. Je l'aborde dans la suite de cette section. Pour l'instant, j'affirme seulement qu'on peut raisonnablement présenter l'argument sceptique comme la mise en évidence d'un réfuteur pour la justification *prima facie* en faveur de A .

exemple K_1A et $\sim K_2\sim HS$, ce qui ne poserait alors aucun problème de clôture, ni pour K_1 ,²⁰³ ni pour K_2 .²⁰⁴ On pourrait donc continuer à défendre la clôture pour tout K_n . Comme nous traitons ici de la *justification* et non de la connaissance, il faudrait transposer la réponse et imaginer plusieurs sens possibles $J_1, J_2, \dots J_n$. Il n'est pas certain que le mot « justification » se prête à ce genre de variation contextuelle aussi facilement que le mot « connaître », mais quoi qu'il en soit du *mot* « justification », j'ai stipulé que dans cette étude je ne m'intéressais qu'à un seul de ses sens possibles, à savoir l'impeccabilité épistémique interne. Et si l'on estime encore qu'il pourrait y avoir des manières plus ou moins rigoureuses (des standards plus ou moins sévères) pour juger de cette impeccabilité, je rappellerai que les *seuls* standards qui nous intéressent ici sont les standards *communs* : ce que nous cherchons, encore une fois, c'est seulement de savoir si la croyance au libre arbitre peut être justifiée avec le même type et le même niveau de justification que nos croyances perceptives.

Une fois qu'on s'intéresse à un seul et unique sens de « justification », il n'y a que deux manières de préserver les justifications perceptives communes JA : il faut soit admettre $J\sim HS$ d'une manière ou d'une autre (au risque d'enfreindre notre intuition d'indéfectibilité), soit refuser la clôture d'une manière ou d'une autre (au risque d'enfreindre notre intuition de clôture). La première solution (admettre une certaine justification pour $\sim HS$) se subdivise encore en deux : soit la justification $J\sim HS$ dépend de notre justification JA , soit elle n'en dépend pas. Mais avant d'entrer dans cette distinction, je vais dans un premier temps expliquer pourquoi la solution par refus de la clôture est difficilement acceptable et non pertinente pour cette étude.

Le refus de la clôture, défendu notamment par Fred Dretske (1970) et Robert Nozick (1981, 204–211), est une des solutions qui a attiré le plus d'attention dans les trente dernières années. L'importance considérable qu'a pris cette solution – une solution dont il faut noter la très grande originalité dans l'histoire de la philosophie – est directement liée à l'essor des épistémologies externalistes, et plus particulièrement des épistémologies fiabilistes. Il ne serait jamais venu à l'esprit d'un philosophe antique ou moderne de remettre en cause l'idée que la connaissance se transfère par *modus ponens*. Et par conséquent, il ne serait jamais venu à l'esprit de ces philosophes d'essayer de répondre à l'argument sceptique autrement qu'en défendant, d'une manière ou d'une autre, la justification de $\sim HS$. Et pourtant, les analyses fiabilistes du mot « connaître » ont permis de mettre au jour des concepts de connaissance qui ne respectent *pas* la clôture, et elles ont ainsi offert pour la discussion du scepticisme une réponse possible qui a le très grand avantage d'être motivée *indépendamment* de tout projet anti-sceptique. Je n'entrerai pas ici dans le détail de ces solutions parce qu'elles reposent entièrement sur l'analyse d'une notion *externaliste* de connaissance : l'idée est qu'un certain type de non-hasard, un certain type de dépendance externe à la vérité, ne se préserve pas à travers l'implication logique.²⁰⁵ Par conséquent, on pourrait soutenir que nous connaissons les propositions anodines A et que néanmoins nous ne connaissons pas les propositions anti-sceptiques $\sim HS$. John Hawthorne (2005) a soulevé des objections assez considérables contre le refus de la clôture *même* pour ces notions externalistes, mais quand bien même la notion externaliste de connaissance serait non close, il ne s'ensuit pas absolument pas que la

²⁰³ Il suffirait d'*admettre* $K_1\sim HS$, c'est-à-dire qu'il y ait un sens dans lequel nous connaissons les négations d'hypothèses sceptiques.

²⁰⁴ Il suffirait de *refuser* K_2A , c'est-à-dire que dans un certain sens fort (et non commun) de connaître, nous ne connaissons pas les propositions d'observation commune.

²⁰⁵ Le type de dépendance à la vérité dont il s'agit, comme nous le verrons au chapitre suivant (p.225 sqq.), est la condition de *sensibilité* ou de *traçage* contrefactuels : « Si p était fausse, S ne croirait pas que p ». Dretske montre, à travers l'exemple maintenant célèbre de la « mule savamment déguisée » que ma croyance que « ceci est un zèbre » peut tracer la vérité bien que la croyance en sa conséquence logique « ceci n'est pas une mule savamment déguisée » ne trace pas la vérité. (Dretske 1970, 1015–1016)

justification soit également non close. Comme on l'a vu avec le « problème de l'inconséquent », il est naturel de penser que la clôture est à l'origine une contrainte de cohérence internaliste ; dire que la dépendance externe au vrai n'est pas close sous l'implication logique, c'est une chose ; dire qu'on peut être justifié à croire que p , apercevoir que p implique logiquement q et... n'avoir aucune justification à conclure que q , c'est une toute autre chose qui met directement en cause le principe le plus basique du raisonnement rationnel. Le fait de refuser absolument la clôture de la *justification* semble donc un déni d'intuition difficilement acceptable. Il faut remarquer, cependant, que si on acceptait ce genre de solution pour le scepticisme, on pourrait tout à fait l'utiliser pour l'introspectionnisme du libre arbitre également. Un anti-sceptique dirait ceci : j'affirme que j'ai deux mains, et cette croyance est justifiée par mon expérience visuelle d'avoir deux mains, et j'ai bien conscience que cela implique qu'il y a un monde extérieur, mais je ne prétends pas pour autant avoir la moindre justification à croire qu'il y a un monde extérieur. L'introspectionniste, de la même façon : j'affirme que je suis libre, et cette croyance est justifiée par mon expérience d'être libre ; être libre suppose également qu'il n'y a pas de dispositif martien dans mon cerveau, mais je ne prétends pas être justifié à croire cela. On pourrait même utiliser cette réponse pour défendre l'introspectionnisme *libertarien* contre la possibilité du déterminisme : j'affirme que je suis libre, sur la base de mon expérience, et que la liberté implique l'indéterminisme, mais pour ce qui est de l'indéterminisme, je ne prétends pas avoir une justification à l'affirmer. Ce dernier exemple permet sans doute de mieux apprécier pourquoi un refus catégorique de la clôture internaliste est en soi inacceptable, et pourquoi il rendrait absolument trivial le problème épistémologique du libre arbitre.

Il semble donc évident que, dans le cas de la justification au moins, *une certaine version* du principe de clôture doit être vraie, pour éviter le problème de l'inconséquent.²⁰⁶ Encore faut-il voir précisément quelle version est vraiment requise. En effet, pour éviter le problème de l'inconséquent, il n'est pas nécessaire d'avoir recours à des versions fortes de la clôture telles que :

(Clôture Forte) Si S croit de manière justifiée que p , et si $p \Rightarrow q$, alors S croit de manière justifiée que q .

Un premier problème de la Clôture Forte, dont nous avons déjà parlé, c'est que la clôture ne doit pas s'appliquer à toutes les conséquences logiques de p , mais seulement à celles qui sont reconnues par S (de manière justifiée) : si S n'a aucunement conscience que p implique q , l'absence de justification pour q ne pose pas de problème à sa justification pour p . Un deuxième problème, qui repose également sur l'idée générale que S doit « avoir conscience du problème », vient du fait que la plupart de nos croyances sont non occurrentes, et qu'on peut très bien, par conséquent, n'avoir jamais songé à « réunir » (ou à concevoir ensemble) deux croyances p et $p \Rightarrow q$ pour s'apercevoir qu'elles impliquent q . L'idée qu'on devrait croire positivement toutes les conséquences q , pour lesquelles on croit p et $p \Rightarrow q$ de manière justifiée, est donc tout à fait invraisemblable. Et par ailleurs, on n'a pas besoin de ce principe radical pour faire apparaître le « problème de l'inconséquent » : ce qu'on reproche à l'inconséquent, c'est de ne pas tirer la conséquence *alors même* qu'il a réuni consciemment les deux prémisses du *modus ponens*. L'idée n'est donc pas que nous devrions positivement croire toutes les connaissances reconnues de nos croyances, mais plutôt que lorsque nous apercevons la conséquence d'une croyance nous devons bien avoir un fondement potentiel de

²⁰⁶ Pour reprendre la sentence de Richard Feldman : « the idea that no version of the closure principle is true—that we can fail to know things that we knowingly deduce from other facts we know—is among the least plausible ideas to gain currency in epistemology in recent years » (Feldman 1999, 95)

justification pour former la croyance q (mais nous n'avions pas forcément cette croyance auparavant).

(Clôture Ded.) Si S croit que p de manière justifiée et occurrente à t , et si S reconnaît à t qu'il peut déduire q de p , alors S a à t une justification propositionnelle à croire que q .²⁰⁷

Ce principe de « clôture sous la déduction compétente » permet de bien cibler ce qui pose problème chez « l'inconséquent » : le problème de l'inconséquent c'est qu'au même moment t , il prétend conserver sa justification à croire que p (et que $p \Rightarrow q$), et il prétend pourtant n'avoir aucune justification propositionnelle à croire que q (aucune raison d'accepter que q). Mais de deux choses l'une : soit sa justification à croire que p (et que $p \Rightarrow q$) lui confère une justification (déductive) à croire que q ; soit il n'acquiert effectivement aucune justification à affirmer que q , mais alors il n'est plus justifié non plus à continuer d'affirmer que p (et que $p \Rightarrow q$). Dans le deuxième cas, sa justification antérieure pour p (ou pour $p \Rightarrow q$) se trouve alors « défaite » ou réfutée. Pour être plus explicite, il faudrait donc formuler le principe de clôture déductive sous la forme d'une disjonction :

(Clôture Disj.) Si S croit que p de manière *prima facie* justifiée à t et si S reconnaît à t qu'il peut déduire q de p , alors

(a) soit S est en mesure d'avoir à t la croyance justifiée que q sur la base

(i) d'une justification antécédente (indépendante de p) à croire que q , ou bien

(ii) d'une justification déductive pour q résultant de la justification *prima facie* pour p

(b) soit la justification *prima facie* pour p se trouve par là même réfutée (pour S à t).

Si on applique ce principe aux scénarios sceptiques, il signifie qu'il n'est pas indispensable (pour éviter le scepticisme) de croire déjà que nous ne sommes pas des cerveaux dans des cuves lorsque nous voyons nos deux mains (A). L'expérience peut conférer une justification *prima facie* (et même sans doute une justification *ultima facie*) en l'absence de toute considération consciente des scénarios sceptiques. Mais *une fois* qu'on considère une hypothèse sceptique, alors de deux choses l'une : soit (a) on est en mesure d'affirmer qu'elle est fausse de manière justifiée (justifiée d'une manière ou d'une autre, que ce soit (i) par une justification indépendante ou (ii) par déduction), soit (b) la conscience de l'impossibilité à tirer toutes les conséquences de A réfute la justification *prima facie* en faveur de A .²⁰⁸ De même pour l'introspectionnisme du libre arbitre : *une fois* qu'on a considéré les scénarios martiens, soit (a) on est en mesure d'affirmer leur négation de manière internalistiquement impeccable, soit (b) on perd toute possibilité d'affirmer l'existence du libre arbitre de manière justifiée, c'est-à-dire que la justification introspective *prima facie* se trouve réfutée. L'alternative (b) est évidemment celle à laquelle veut conclure l'objecteur de l'impression lourde... mais c'est aussi la conclusion recherchée par le sceptique dans l'argument parallèle. Par conséquent, toute stratégie anti-sceptique « conséquente » (c'est-à-dire toute stratégie anti-sceptique qui évite le « problème de l'inconséquent » en acceptant un principe minimal

²⁰⁷ Sur la notion de « justification propositionnelle » par opposition à la « justification doxastique », voir p.93. Cette formulation du principe est largement inspirée de l'affaiblissement de la clôture proposé par Hawthorne grâce à la notion de « déduction compétente » (Hawthorne 2005). Dans (2004, 36–46), ce principe s'appelle « single premise closure » car il est formulé dans un contexte où l'alternative essentielle est entre clôture à une prémisse et clôture à plusieurs prémisses. Mais le rôle de la déduction compétente et la motivation de « conscience du problème » restent fondamentalement les mêmes. Je n'aborderai pas ici les problèmes posés par la clôture à plusieurs prémisses.

²⁰⁸ C'est-à-dire que cette conscience même rend impossible de continuer à affirmer (ou à croire) A d'une manière internalistiquement impeccable.

de clôture) revient à soutenir que (a) nous sommes en mesure, après réflexion, d'affirmer la négation des hypothèses sceptiques d'une manière qui ne viole aucun principe de rationalité interne, c'est-à-dire d'une manière *justifiée*, quelle que soit la nature exacte (i ou ii) de la justification proposée pour $\sim HS$. Selon le type de justification qu'on accepte pour $\sim HS$, il existe donc deux stratégies anti-sceptiques conséquentes. Ce sont ces deux stratégies que je vais discuter à présent.

Les deux réponses « conséquentes » à l'argument de clôture, dans la mesure où elles *préservent* l'intuition de clôture, courent le risque de ne pas faire droit à l'autre intuition de l'argument sceptique, à savoir l'intuition d'indéfectibilité (que le sceptique et le négateur de la clôture expriment sous la forme forte $\sim J \sim SH$). L'enjeu sera donc de voir si l'on peut (et/ou doit) faire droit à cette intuition d'indéfectibilité. La prémisse d'indéfectibilité telle que la formule le sceptique, $\sim J \sim SH$, repose comme on l'a vu sur l'idée que les scénarios sceptiques sont *par stipulation* des scénarios où toutes les données internes sont indiscernables du scénario normal. Je rappelle l'argument de Pritchard :

S'il n'y a, *ex hypothesi*, aucune différence phénoménologique disponible au sujet, qui puisse lui indiquer qu'elle n'est pas une victime de ce scénario, alors il s'ensuit qu'il n'y aura rien d'accessible réflexivement au sujet qui puisse suffire à lui indiquer que sa croyance dans cette proposition anti-sceptique est vraie. (Pritchard 2005, 44)

Si l'on admet que la justification (interne) d'une croyance doit reposer sur un justifieur (interne), c'est-à-dire sur un certain élément de la vie mentale du sujet, et si, par stipulation, les hypothèses sceptiques empêchent qu'il y ait dans la vie mentale du sujet le moindre justifieur, alors les hypothèses sceptiques rendent logiquement impossible toute forme de justification interne. Mais ce raisonnement contient en fait une prémisse implicite qui est tout à fait problématique. On apercevra mieux cette prémisse si on présente l'argument sous la forme suivante :

- (1) Aucun élément interne ne permet de faire une différence entre HS et $\sim HS$.
- (2) Donc aucun élément interne ne peut être un justifieur pour croire que $\sim HS$.
- (3) Donc il n'est pas possible d'être justifié à croire que $\sim HS$.

La transition de (2) à (3) n'est pas problématique, si du moins on accepte le fondationnalisme internaliste pour la notion de justification : seul un élément interne peut être justifieur, et seule une croyance dotée de justifieur positif peut être dite justifiée. Mais la transition de (1) à (2) suppose un principe bien plus contestable, à savoir que la seule manière pour un élément de la vie mentale de S de justifier à croire que p , c'est de « faire une différence » entre le scénario où p est vraie et le scénario où $\sim p$ est vraie.

Or une stratégie anti-sceptique traditionnelle, celle défendue notamment par Peter Klein (1995, 226–228), consiste précisément à attribuer le rôle de justifieur pour $\sim HS$ à un élément interne qui « ne fait pas de différence », à savoir à la croyance que A dont $\sim HS$ peut être déduite. Si la croyance que A constitue un justifieur pour croire $\sim HS$, c'est tout simplement parce que son contenu (A) implique logiquement $\sim HS$, et non pas parce l'occurrence de cette croyance ferait une différence entre les deux scénarios.²⁰⁹ De fait, puisque *tout* élément

²⁰⁹ Peut-être faut-il faire ici une clarification sur la notion de « faire une différence » : dans un sens faible, la croyance que A fait bien une différence en faveur du scénario $\sim HS$, c'est-à-dire qu'elle favorise (de par son contenu) la croyance que $\sim HS$ plutôt que la croyance que HS . Mais dans le sens pertinent pour la discussion sceptique (dans le sens où (1) est vraie par stipulation), dire que la croyance que A ne fait pas de différence, c'est dire que cet élément interne est présent *aussi bien* dans le scénario sceptique que dans le scénario normal.

interne du sujet normal est reproduit (par stipulation) dans l'esprit du sujet encuvé, il s'ensuit que la croyance que *A* sera également présente chez le sujet encuvé. Et donc le sujet encuvé aura *également* une justification (déductive) à croire qu'il n'est pas encuvé. On pourrait s'étonner que cette stratégie anti-sceptique en vienne à attribuer aux deux sujets (encuvé et normal) les mêmes mérites épistémiques : n'y a-t-il pas clairement une différence de mérite épistémique entre les deux sujets ? Bien sûr que si, mais il s'agit d'une différence *externaliste*, une différence en termes de « connaissance ». Pour ce qui est en revanche de la notion de « justification », l'idée que les deux sujets sont à égalité (pour le meilleur ou pour le pire) n'est pas un résultat malencontreux : c'est au contraire l'intuition de départ sur la question. Comme on l'a vu en Introduction (p. 24), cette intuition de départ est suffisamment forte pour constituer la prémisse principale du meilleur argument contemporain en faveur de l'internalisme de la justification, le « Nouvel Argument du Malin Génie ». Contrairement à ce que pense Pritchard, la véritable intuition d'indétectabilité est donc une intuition fondamentalement *externaliste* et non pas internaliste, une intuition du type $\sim K \sim HS$ et non pas $\sim J \sim HS$. Certes les scénarios sceptiques sont construits pour générer une *indiscernabilité* de nature fondamentalement interne, mais de cette indiscernabilité interne, c'est une indétectabilité *externe* qui découle. Pour résumer, la stratégie « déductiviste » de Peter Klein fait apparaître (en dépit de l'argument de Pritchard) un possible justifieur interne pour $\sim HS$ – ce qui invalide donc le principe $\sim J \sim HS$. Et le fait d'accepter un justifieur interne $J \sim HS$ n'est pas un pur déni d'intuition indétectabiliste, puisqu'il apparaît que cette intuition peut être « canalisée » comme étant une intuition fondamentalement externaliste (du type $\sim K \sim HS$). Or ce principe externaliste, même s'il était vrai, ne serait pas la prémisse requise pour faire fonctionner l'argument de la clôture anti-justification. Certes, si ce principe était vrai *et cru de manière justifiée* par un sujet, cela pourrait soulever une question de méta-cohérence, et peut-être entraîner une réfutation fondationnelle de la justification (déductive) à croire $\sim HS$. C'est-à-dire qu'on peut construire un argument sceptique de deuxième ordre fondé sur le principe de méta-cohérence *plus* la prémisse d'indétectabilité externe. J'aborderai cet argument au chapitre 3. Pour ce chapitre, je m'en tiens aux arguments sceptiques de premier ordre, et l'argument de clôture à proprement parler est strictement de premier ordre : il interroge seulement notre capacité croire de manière justifiée $\sim HS$, pas notre capacité à croire de manière justifiée que « je sais que $\sim HS$ ». Quand on s'en tient aux considérations de premier ordre, les seuls réfuteurs possibles d'une justification sont les « réfuteurs directs » : en l'occurrence, ce serait une justification à croire *positivement* qu'un certain scénario sceptique est le cas ; or personne – pas même les sceptiques – n'a jamais soutenu que nous avons une justification positive à croire un scénario sceptique.

Pour revenir aux deux solutions anti-sceptiques « conséquentes » envisagées plus haut – à savoir (i) une justification $J \sim HS$ indépendante de la justification JA , ou (ii) une justification $J \sim HS$ reposant sur JA – la stratégie de Peter Klein représente clairement la deuxième voie ; et cette deuxième voie n'a pas à craindre l'intuition d'indétectabilité *absolue* exprimée par Pritchard, parce que cette intuition ne concerne que le niveau externaliste. Pourtant, si un certain nombre de philosophes ont préféré la première voie, c'est à cause d'une autre intuition d'indétectabilité, plus limitée mais du coup plus dangereuse pour la stratégie déductiviste : quand bien même la croyance que je ne suis pas un cerveau dans une cuve pourrait être « détectée » ou justifiée d'une manière ou d'une autre, il semble qu'elle ne pourrait pas être détectée *de cette manière-là*, c'est-à-dire sur la base d'une expérience perceptive *A* (qui pourrait aussi bien avoir été provoquée par l'ingénieur qui s'occupe de la cuve). L'intuition, ici, est qu'il y aurait une sorte de circularité ou de pétition de principe, à accepter la non-manipulation de nos expériences *sur la base même* de nos expériences. Donc en admettant que nos expériences fournissent une justification pour leur contenu *A*, cette justification ne pourrait pas se « transmettre » à la conséquence $\sim HS$. Cette intuition « anti-transmission » a

été particulièrement développée par Crispin Wright (2000). Si on l'accepte, elle permet de montrer que, en fin de compte, la réponse « déductiviste » ne sauve le principe de clôture qu'au prix d'un déni d'intuition du côté de l'indéfectabilité (limitée). C'est ce problème qui a amené un certain nombre d'anti-sceptiques conséquents à emprunter la voie (i), c'est-à-dire à soutenir que la croyance $\sim HS$ est internalistiquement impeccable ou justifiée *autrement* que par une justification dépendant ultimement de JA. Je vais donc envisager rapidement cet autre type de réponse anti-sceptique (conséquente) : comme aux étapes précédentes de mon raisonnement, je montrerai que cette réponse anti-transmission, si on l'acceptait, pourrait être transférée sans problème à la défense de l'introspectionnisme du libre arbitre, mais qu'elle n'est ni la plus plausible ni la plus efficace pour formuler la suite du problème du libre arbitre. Ultimement, je vais défendre que cette solution ne présente pas l'équilibre le plus plausible entre intuition de clôture et intuition d'indéfectabilité, parce qu'elle affaiblit trop l'intuition de clôture et que le cœur de l'intuition anti-transmission peut en fait être satisfait par les solutions déductivistes à la Peter Klein.

Les stratégies anti-transmission comme celle de Crispin Wright comportent un versant négatif – qui vise à satisfaire l'intuition d'indéfectabilité du sceptique – et un versant positif – qui vise à satisfaire l'intuition de clôture. Le versant négatif est constitué par la critique des stratégies déductives, sur la base de l'intuition anti-transmission que je viens de présenter. Pour des raisons historiques, cette critique des stratégies déductives part généralement de la très fameuse « preuve du monde extérieur » de G. E. Moore (1939). Selon la présentation la plus courante aujourd'hui, Moore prétendrait répondre au sceptique sur la base d'une « preuve parfaitement rigoureuse » dotée d'une seule prémisse : « Voici une main » (cette prémisse étant prononcée par Moore en même temps qu'il lève la main droite devant son auditoire). Sur la base de cette prémisse – qui est au moins aussi certaine que n'importe quelle autre prémisse qu'on puisse apporter pour n'importe quelle autre preuve – Moore déduit validement la conclusion recherchée : « Il y a un monde extérieur ». ²¹⁰ La preuve de Moore, dans cette reconstruction au moins, a été très largement considérée comme un échec patent, comme une réponse naïve qui passe totalement à côté des inquiétudes du sceptique. Le constat d'échec est si consensuel qu'il ne semble y avoir débat que sur le diagnostic exact des raisons de l'échec, et sur les intentions profondes qui pouvaient être celles de Moore en proposant un argument si clairement fallacieux. ²¹¹ En partant de ce consensus, l'idée de Crispin Wright est ensuite de faire remarquer que les stratégies déductivistes à la Peter Klein ne font finalement rien d'autre que proposer une preuve mooréenne comme fondement de la justification anti-sceptique $J\sim HS$. La preuve de Moore est donc prise à la fois comme archétype et comme réduction à l'absurde de toute stratégie anti-sceptique dans laquelle la justification $J\sim HS$ dépend ultimement de justifications aux croyances anodines de type JA. Et pour rendre compte de l'échec de Moore (et de toute stratégie déductiviste), Wright avance le diagnostic de circularité ou de pétition de

²¹⁰ Crispin Wright (2007, 25–26), par exemple, donne cette présentation classique de l'argument de Moore. En réalité, ainsi que l'ont montré (Sosa 2007, 52–54) et (Coliva 2010, 26–28), l'adversaire auquel Moore entendait répondre par cette preuve n'était pas vraiment le *sceptique* (qui refusera la prémisse qu'il existe une main, ou qui insistera sur le fait qu'elle n'est pas connue) mais plutôt l'*idéaliste* (qui acceptera la prémisse qu'il existe une main, et ne sera en désaccord que sur le caractère *extérieur* de l'objet « main » - toute la première partie de l'article de Moore est d'ailleurs consacrée à l'analyse de ce qu'est un objet *extérieur*).

²¹¹ (Wright 2007, 25–27) et (Pryor 2004, 349–351) partent tous deux de ce consensus qu'il y a un problème dans l'argument de Moore. (Coliva 2008) donne une recension particulièrement intéressante des différents diagnostics contemporains de l'origine du problème, en insistant sur le fait qu'il est *immédiatement évident* pour tout un chacun qu'il y a un problème. Ce caractère absolument évident de l'échec a amené (Soames 2003, 23) à conclure que Moore lui-même ne pouvait pas considérer sérieusement son argument comme un véritable argument, mais qu'il avait dû le proposer au contraire pour faire apparaître l'inanité même de toute tentative d'argumentation contre le scepticisme.

principe exposé au paragraphe précédent, c'est-à-dire le diagnostic le plus classique, ou « conservateur » selon la terminologie de (Pryor 2004) :

La Preuve échoue parce que sa prémisse serait sans support [evidence] si sa conclusion ne faisait déjà partie de l'information du sujet. (Wright 2007, 27)

Pour qu'une preuve déductive réussisse, il faut non seulement que la prémisse implique logiquement la conclusion, mais il faut aussi que le support justificationnel sur lequel repose la prémisse puisse devenir (indirectement) le support justificationnel de la conclusion. Or c'est précisément cela qui est impossible dans la « Preuve » de Moore : le canal déductif $A \Rightarrow \sim HS$ transmet valablement la vérité de la prémisse à la conclusion, mais il ne laisse pas passer la justification de la prémisse à la conclusion.²¹² Et par conséquent, il faut refuser l'idée que l'implication logique préserve *toujours* la justification des prémisses. C'est-à-dire qu'il faut refuser le principe suivant :

(J-Transmission) Si S a à t une justification à croire que p et si S reconnaît à t qu'il pourrait déduire valablement q de p , alors S a à t une justification déductive (potentielle) à croire que q sur la base de sa justification à croire que p .

La Preuve de Moore est un exemple patent d'échec du principe de transmission de la justification ; et pour que sa réponse ne semble pas *ad hoc*, Wright propose d'autres exemples, notamment celui d'un spectateur dans un stade qui voit un homme en maillot shooter dans un ballon et l'envoyer dans une cage de but, et qui produit la croyance :

(i) un but vient d'être marqué.

Wright (2000, 142) remarque qu'il serait illégitime pour ce spectateur d'utiliser sa croyance (i) pour déduire la conséquence (tout à fait valide) que

(ii) c'est une partie de foot qui est en train de se jouer [par opposition à un simulacre de partie de foot dans le tournage d'une série de télévision].

La justification dont le spectateur dispose pour (i) ne se transmet pas à (ii) car (ii) doit déjà faire partie de l'information dont il dispose pour pouvoir être justifié à croire que (i). A cause de tels exemples, Wright estime que l'on a des raisons indépendantes de croire que la justification ne se transmet pas toujours d'une croyance à ses conséquences reconnues, et en particulier, qu'elle ne se transmet pas des croyances anodines aux croyances anti-sceptiques, de telle sorte que la stratégie « déductiviste » échoue à nous donner une justification pour $\sim HS$.

Mais la stratégie de Crispin Wright comporte également un versant positif : après avoir refusé le principe de transmission et la justification *déductive* pour $\sim HS$, Wright montre qu'il peut continuer de défendre le principe de clôture et une autre sorte de « justification » pour $\sim HS$. Wright (2000, 140–141) met clairement en évidence la différence entre clôture et transmission : l'un et l'autre principe consistent à dire qu'une croyance que p , si elle est

²¹² Pour être exact, il faut souligner que Wright ne parle pas de la transmission de la *justification* mais de la transmission du *warrant* : son intention est ainsi de rester neutre entre analyses internaliste et externaliste du *warrant*, avec l'idée que son propos marche aussi bien contre la transmission dans l'une et l'autre analyses. Cependant, dans la mesure où les stratégies déductivistes qu'il critique sont d'abord et avant tout des réponses au problème de la clôture *internaliste*, c'est presque toujours dans le cadre d'une discussion sur le *warrant* internaliste (donc la *justification* dans ma terminologie) que ses thèses se trouvent discutées. Voir notamment (Engel 2007, 156–159).

justifiée, est toujours accompagnée d'une justification à croire toute conséquence reconnue q . Mais là où la transmission affirme l'existence d'une justification *déductive* en faveur de q , la clôture en revanche peut admettre n'importe quelle sorte de justification pour q – l'important est seulement qu'il y ait *une* justification pour q , sinon la justification pour p se trouverait réfutée. Or Wright accepte la clôture de la justification, au moins dans le sens minimal que j'ai défini plus haut : c'est-à-dire qu'il considère que l'anti-sceptique confronté aux hypothèses sceptiques est effectivement obligé d'affirmer $\sim HS$, et il considère qu'il lui est *possible* d'affirmer cela sans violer aucune norme de rationalité interne, donc de manière *justifiée* dans ma terminologie. Seulement, pour Wright, ce n'est absolument pas la justification de A qui rend compte de l'affirmation légitime de $\sim HS$: au contraire, la rationalité de cette affirmation doit être conçue comme antérieure et indépendante à la justification de A . Pour Wright, les hypothèses sceptiques disposent, en amont de toute information empirique, et même en amont de toute réflexion *a priori*, d'un type particulier de rationalité qu'il appelle « Autorisation » (*Entitlement*).

Avant de voir les difficultés que pose cette réponse, soulignons tout d'abord qu'elle pourrait évidemment s'appliquer à la défense de l'introspectionnisme du libre arbitre si on l'acceptait comme solution anti-sceptique. En effet, si nous avons une autorisation wrightienne à présumer que nous ne sommes pas les victimes d'un malin génie ou d'un scientifique manipulant nos cerveaux dans des cuves de formol, il est clair que nous serons tout autant autorisés à présumer que nous ne sommes pas dotés d'un dispositif de manipulation martienne déterminant nos actions. Par conséquent, un wrightien pourrait répondre à l'argument de l'impression lourde la chose suivante : certes, en affirmant l'existence du libre arbitre sur la base de notre expérience E , nous sommes engagés à croire $\sim MM$, et sans doute ne pouvons-nous pas croire $\sim MM$ sur la base de notre expérience E , mais nous sommes justifiés à le croire d'une *autre* manière, indépendante de l'expérience E , par simple autorisation épistémique. La réponse anti-sceptique de Wright pourrait donc suffire à défendre l'introspectionnisme contre l'objection de Peter van Inwagen. Mais il me semble que cette réponse, et son application au libre arbitre, n'est ni la plus convaincante, ni la plus utile pour envisager la suite de la discussion sur le libre arbitre.

L'intention de Wright est de concilier à la fois l'intuition d'indétectabilité (sous la forme du refus de la transmission) et l'intuition de clôture (grâce à l'autorisation épistémique), mais le compromis trouvé n'est finalement pas satisfaisant, et ce pour trois raisons. Premièrement, Wright ne parvient à préserver la clôture qu'en introduisant une notion d'autorisation épistémique qui consiste en fait à revenir sur le principe évidentialiste présenté plus haut (p. **Erreur ! Signet non défini.** sqq.). La définition même de l'autorisation épistémique met clairement ce point en évidence. L'autorisation de Wright est :

un type de garantie [warrant] rationnelle dont la possession ne requiert pas l'existence de données [evidence] – dans le sens le plus large, comprenant à la fois les considérations *a priori* et empiriques – pour la vérité de la proposition garantie. (Wright 2004, 174–175)

Dans la stratégie de Wright, la croyance $\sim HS$ est donc internalistiquement impeccable – « beyond rational reproach » (2004, 175) et même « all things considered, epistemically appropriate » (2007, 30) – sans qu'il y ait le moindre justifieur positif en sa faveur. Cette stratégie est donc victime des problèmes du « prisonnier » ou « d'isolation » considérés plus haut. Deuxièmement, le refus du principe de transmission de justification est en soi un coût très élevé à payer lorsqu'on se concentre sur une notion internaliste de justification. En effet, tout raisonnement déductif doit d'une manière ou d'une autre reposer sur le fait qu'une certaine croyance offre à une autre croyance logiquement impliquée une justification

dérivée ; par conséquent, il est évident qu'un certain principe de transmission *doit* être vrai si l'on accepte des déductions. Wright pourrait certes préserver un principe de transmission *restreint*, mais si l'essence de la justification déductive tient à l'implication logique entre les deux contenus, alors la restriction à *certaines* implications plutôt que d'autres semble arbitraire. Il semble évident, du point de vue internaliste, que quand je sais que $p \Rightarrow q$, ma justification pour p donne au moins un certain degré de justification *prima facie* pour q ; on peut certes discuter des circonstances dans lesquelles cette justification *prima facie* pourra devenir *ultima facie*, mais refuser absolument la moindre justification interne, même *prima facie*, en faveur de q lorsque je suis justifié à accepter p et que je déduis q de p , c'est violer l'esprit même de la « clôture sous la déduction compétente » tout en respectant la lettre. Troisièmement, le compromis de Wright ne semble pas le meilleur parce que les deux sacrifices que je viens de mentionner (évidentialisme et transmission de la justification) sont motivés par des *desiderata* qu'on peut satisfaire sans payer un tel coût : Wright fait légitimement apparaître qu'une bonne part de l'intuition indétectabiliste réside dans le sentiment très largement partagé que « quelque chose ne va pas dans la preuve mooréenne » ; mais comme on va le voir, on peut donner un diagnostic du problème mooréen qui préserve le principe de transmission. Pour résumer, s'il s'agit de trouver un compromis entre intuition de clôture et intuition indétectabiliste, le compromis de Wright n'est pas satisfaisant parce qu'il viole le cœur de l'intuition internaliste de clôture sous la déduction, et ce dans le seul but de rendre compte d'une intuition anti-mooréenne (indétectabiliste) dont on peut rendre compte autrement. L'enjeu est donc à présent d'exposer une solution « déductiviste » qui respecte le principe de transmission tout en rendant compte de l'échec de Moore. Et c'est très exactement cette tâche que s'est proposée Jim Pryor dans sa réponse anti-sceptique qu'il a baptisée « dogmatisme ».

La solution « dogmatiste » de Pryor consiste d'abord à sauver partiellement la preuve de Moore, en distinguant trois niveaux d'échecs possibles pour un argument : niveau logique, niveau épistémologique et niveau dialectique. Nul ne conteste (pas même Wright) que la preuve de Moore soit valide au niveau logique, c'est-à-dire que ses prémisses sont vraies et impliquent logiquement la conclusion. Mais Wright situe le problème au niveau épistémologique : la structure logique de l'argument ne peut pas, d'après lui, être utilisée comme structure de justification appropriée pour les croyances d'un sujet. Et d'après Wright, c'est *pour cela* que l'argument est aussi dialectiquement inopérant. Pryor, quant à lui, fait apparaître qu'un argument peut être dialectiquement inopérant même s'il repose sur une structure justificationnelle tout à fait appropriée. Imaginons par exemple que Conrad est victime d'un complot extrêmement sophistiqué qui a pour but de le faire inculper dans une affaire de meurtre ; ses ennemis ont réuni un nombre considérable de faux indices accablants de telle sorte que la police et le tribunal, bien qu'ils soient honnêtes, ne peuvent pas rationnellement douter de la culpabilité de Conrad. Le jour du procès, on demande à Conrad s'il était sur les lieux du crime (un hôtel) à midi ; il répond qu'il ne pouvait pas y être puisqu'il se souvient pertinemment avoir passé toute la journée à la plage. Il est évident que son témoignage ne donnera aux juges aucune justification à croire qu'il n'était pas sur les lieux du crime étant donné tous les faux indices qu'ont réuni les ennemis de Conrad. Donc son argument sera dialectiquement inopérant, mais cela n'empêche pas que sa croyance « je n'étais pas sur les lieux du crime » soit parfaitement légitimement fondée sur sa croyance « j'étais à la plage », elle-même fondée sur le souvenir net d'avoir été à la plage. L'argument de Conrad a une excellente structure justificationnelle, mais il ne peut pas *communiquer* à d'autres cette structure, parce que ses interlocuteurs ne sont pas dans la même « situation évidentielle » que lui. Les juges ont trop de réfuteurs évidentiels en défaveur de Conrad

pour pouvoir le croire sur parole. Pryor fait voir que la preuve de Moore souffre du même problème *lorsqu'elle est adressée au sceptique* :

Il est clair que l'argument de Moore n'est pas très efficace dialectiquement contre le sceptique. Mais la raison devrait être claire à présent. Le sceptique a des doutes qui empêchent l'argument de Moore de le persuader rationnellement. Il n'y a rien de problématique dans la structure justificationnelle que l'argument développe, ni avec le raisonnement de Moore lui-même. Ce qui ne va pas, c'est que le sceptique a des doutes qu'il ne devrait pas avoir. (Pryor 2004, 368)

Si Moore ne peut pas convaincre le sceptique, c'est donc simplement parce que ce dernier entretient d'entrée de jeu des suspicions (sur la fiabilité de nos sens) que Moore lui-même n'entretient pas. Ces suspicions, lorsqu'elles sont présentes, constituent un réfutateur de la justification à croire *A* (et donc également à croire $\sim HS$). Mais *en l'absence* de ces suspicions (c'est-à-dire dans la situation évidentielle où se trouve Moore, ou Pryor, ou toute personne dans son bon sens), il n'y a rien de problématique à ce que la justification *prima facie* pour *A* devienne justification *ultima facie*... ni au fait qu'elle se transmette à $\sim HS$. En fait, le cadre conceptuel des « justifications *prima facie* » et « réfuteurs » permet de dire que la justification *prima facie* se transmet de *A* à $\sim HS$ même pour le sceptique : le sceptique a une justification *prima facie* pour *A*, et cette justification se transmet (après réflexion) en justification *prima facie* pour $\sim HS$; le problème est seulement que (pour le sceptique) l'une et l'autre justifications *prima facie* sont réfutées par la présence (dans son système épistémique) des suspicions sceptiques (lesquelles suspicions sceptiques sont excessives, inappropriées, et donc injustifiées, mais un élément interne peut constituer un réfutateur de justification même s'il est lui-même injustifié). Par conséquent, rien dans l'intuition anti-Moore ne s'oppose à une version « *prima facie* » du principe de transmission, qu'on pourrait formuler ainsi :

(J-Transmission*) Si *S* a à *t* une justification *prima facie* à croire que *p* et si *S* reconnaît à *t* qu'il pourrait déduire valablement *q* de *p*, alors *S* a à *t* une justification *prima facie* à croire que *q* (sur la base de sa justification à croire que *p*).

Toutes les réticences de Wright lui-même peuvent s'exprimer en termes de réfuteurs éventuels sans violer ce principe. La position de Wright pourrait se reformuler de la manière suivante : la justification déductive *prima facie* en faveur de propositions anti-sceptiques $\sim HS$ sera systématiquement réfutée et ne se transformera donc jamais en justification déductive *ultima facie*. C'est donc sur la question des réfuteurs qu'il y a véritablement désaccord entre les anti-sceptiques déductivistes et les anti-sceptiques qui réclament une justification antécédente pour $\sim HS$: un déductiviste comme Pryor estime qu'il n'y a réfuteur que chez un esprit sceptique qui entretient (positivement) des suspicions (injustifiées) ; tandis que Wright considérera qu'il y a réfuteur chez tout sujet qui considère la question sérieusement, et que Moore lui-même ne peut donc pas obtenir une justification déductive *ultima facie* pour $\sim HS$.

Etant admis qu'il y a transmission au moins d'une certaine justification *prima facie* en faveur de $\sim HS$, Wright a-t-il raison de penser qu'il existe un réfuteur systématique de cette justification *prima facie* ? Pour répondre à cette question, il faut d'abord considérer la nature du réfuteur que Wright met en avant. Nous allons voir qu'il s'agit d'un réfuteur fondationnel ou de deuxième ordre, et non pas d'un réfuteur direct ou de premier ordre. On peut voir cela en partant d'un point qui fait consensus entre Wright et Pryor, à savoir le fait que l'échec de la preuve de Moore n'est pas absolu mais « relatif au fondement ». Pryor met en évidence que le problème Mooréen, même si on le décrit avec un diagnostic wrightien anti-transmission, disparaît si on imagine certains types de sources pour la prémisse :

Par exemple, si Dieu vous disait que p [A] est vraie, vous pourriez être capable d'inférer de manière justifiée que non- U [$\sim HS$] à partir de p , mais quand votre fondement pour croire p est perceptif, vous ne pouvez pas inférer de manière justifiée non- U à partir de p . (Pryor 2000, n. 35)

Et Wright est d'accord sur ce point, bien qu'il l'exprime dans une terminologie différente :

La transmission de garantie n'est pas nécessairement une caractéristique absolue d'un argument valide. Il se peut qu'un argument particulier soit tel que un certain type de garantie, w_1 , pour ses prémisses est transmissible [...] tandis qu'un autre, w_2 , ne l'est pas, mais peut seulement être possédé par un sujet dont l'information contient déjà une garantie pour la conclusion. (Wright 2000, 141)

Deuxièmement, une fois qu'on a remarqué que le problème mooréen était variable selon le fondement épistémique, il faut remarquer que cette question de fondement peut être un fait parfaitement externaliste. Le scénario suivant permet de s'en rendre compte :

Scénario de circularité purement externe :

Sur un champ de bataille, Don Diègue rencontre le comte de Gormas, dont il ignore l'identité. Il se lie d'amitié avec lui et, après plusieurs semaines, il constate que c'est un homme fiable. Au cours d'une conversation, le comte de Gormas, sans révéler son identité, lui dit que tous les comtes de Castille, notamment les Gormas, sont des hommes d'honneur. Comme il estime (légitimement) que son interlocuteur est fiable, Don Diègue accepte cette croyance. Quelques années plus tard, Don Diègue a toujours en mémoire cette croyance que les comtes de Gormas sont fiables, mais il ne sait plus du tout de quelle source elle lui vient. Un jour, dans une réception, il rencontre le comte de Gormas qu'il n'a pas vu depuis toutes ces années, et on lui apprend enfin son identité de comte de Gormas. Don Diègue se fait alors la remarque suivante : « j'ai toujours cru que je pouvais me fier à cet homme, mais voici une nouvelle raison de le penser : c'est un Gormas ».

Il est clair que le « renfort épistémique » que Don Diègue croit trouver dans ce raisonnement souffre du type de circularité qui inquiète Crispin Wright : Don Diègue a sans doute une « garantie » pour croire que « les comtes de Gormas sont fiables » (car il le tient d'un homme dont il a éprouvé directement la fiabilité), et cette croyance (en conjonction avec l'information non problématique que cet homme est comte de Gormas) implique logiquement que « cet homme est fiable », mais il semble que la « garantie » ne se transmette pas de la prémisse à la conclusion, car la prémisse n'a de garantie *que parce que* Don Diègue avait une garantie antécédente pour la conclusion. Seulement, ce qui apparaît ici, c'est que ce problème de circularité est fondamentalement *externaliste* : au moment où Don Diègue opère sa déduction, il n'y a plus dans son esprit la moindre trace des faits qui génèrent la circularité. Donc si cette circularité pose un problème épistémique (un problème de « garantie ») ce ne pourra être que dans un sens externaliste, c'est-à-dire qu'elle constituera éventuellement un obstacle à la *connaissance*, pas un réfuteur de la *justification*. Pour passer du problème de connaissance à un réfuteur de justification, il faudra que la circularité en question soit *positivement représentée* dans l'esprit de Don Diègue. Par exemple, si on modifie le scénario précédent en disant simplement que Don Diègue se rappelle parfaitement d'où lui vient cette

information,²¹³ alors il est évident que le problème externaliste se répercutera en problème internaliste, c'est-à-dire que la déduction de Don Diègue sera alors injustifiée, à cause de la présence d'un réfuteur. Et le réfuteur en question sera clairement un réfuteur fondationnel, ou de second ordre, c'est-à-dire un réfuteur qui suppose la maîtrise de concepts épistémiques et s'attaque à la manière dont la croyance est produite plutôt qu'à son contenu.

De la même façon, si l'on cherche à défendre, avec Wright, que la déduction de $\sim HS$ (à partir de A) est justificationnellement réfutée, on devra dire que le réfuteur en question est un réfuteur de second ordre, à savoir la *conscience* (interne) que telle *source* pour A ne peut pas transmettre une *connaissance* (externe) pour $\sim HS$. Autrement dit, le réfuteur proposé repose sur les questions de méta-cohérence que j'étudierai au chapitre 3. Comme je l'ai dit un peu plus haut (p.161), l'objection de clôture à proprement parler est strictement une objection de premier ordre (qui interroge notre justification à croire $\sim HS$ et non pas notre justification à croire que « je sais que $\sim HS$ »). Dans le contexte de l'intuition d'intédictabilité absolue de $\sim HS$ (l'intuition de Pritchard), j'avais répondu que cette indédictabilité, si elle était effective, était externe et ne pouvait donc poser pour la justification, au pire, qu'un problème de méta-cohérence. Ce qu'on vient de voir, c'est que l'intuition d'intédictabilité relative (l'intuition de Wright qu'on ne peut pas détecter $\sim HS$ par la déduction à partir de A) est exactement dans la même situation : elle provient d'un souci externaliste et ne peut entraîner un problème de justification qu'au moyen d'un principe de méta-cohérence. Dans la mesure où l'argument de clôture est strictement un argument de premier ordre, la réponse déductiviste (qui ne pose aucun problème de premier ordre) est donc parfaitement satisfaisante à cette étape de la discussion, en dépit des réticences de Wright ; et elle présente l'énorme avantage, par rapport à la solution de Wright, de préserver l'intuition forte de la transmission *universelle* de justification *prima facie* lors d'une déduction compétente. Par ailleurs, pour anticiper sur les résultats du chapitre suivant, nous verrons que cette solution déductiviste ou dogmatiste résiste également aux objections de second ordre ou de méta-cohérence, que ces objections reposent sur une indédictabilité externe de principe (Pritchard) ou sur un problème de circularité externe (Wright).

Ainsi, la solution de Pryor semble bien remplir le cahier des charges initial : elle permet de rendre compte de l'échec mooréen, tout en évitant les deux aspects contre-intuitifs de la solution de Wright. D'une part, elle permet de préserver l'intuition internaliste très forte qu'en apercevant la conséquence logique d'une croyance justifiée, j'acquiesce par là une certaine justification, au moins *prima facie*, à croire cette conséquence. D'autre part, en conférant aux négations d'hypothèses sceptiques une justification de type « déductif », elle permet de préserver l'évidentialisme, c'est-à-dire de préserver l'idée qu'une croyance que p ne peut être internalistiquement impeccable que s'il y a *quelque chose* dans l'esprit du sujet (en l'occurrence la croyance anodine JA) qui *favorise* d'une manière ou d'une autre p plutôt que $\sim p$.²¹⁴

²¹³ C'est-à-dire qu'il entretient la croyance mémorielle « c'est cet homme qui m'a fait savoir que les comtes de Gormas étaient fiables ».

²¹⁴ Le caractère « déductif » de la justification pour $\sim HS$ est à prendre dans un sens large : il est probable que personne d'entre nous n'a jamais consciemment opéré une *déduction* mooréenne pour arriver à la croyance que nous ne sommes pas des cerveaux dans des cuves. Ce qui se passe plutôt, lorsqu'un scénario sceptique est présenté à un homme sain, c'est qu'il tient spontanément ce scénario pour faux sans opérer aucune « déduction » consciente. Mais il le tiendra pour faux *sur la base du fait* que ce scénario est exclu par une écrasante majorité de choses qu'il sait, c'est-à-dire en fait sur la base justificationnelle ultime d'une collection d'expériences perceptives qui (chacune individuellement) implique logiquement la fausseté du scénario évoqué. Autrement dit, la croyance ne sera pas explicitement produite comme une déduction, mais ce qui la rend épistémiquement appropriée sera tout de même la justification que nous avons pour ces autres croyances perceptives. Cette description plus réaliste de la manière dont nous formons nos croyances $\sim HS$ peut donner les ressources pour

Il est clair que la réponse de Pryor peut parfaitement être transposée en défense de l'introspectionnisme contre l'objection de l'impression lourde. Un introspectionniste pourrait en effet dire ceci : l'expérience du libre arbitre nous donne une justification *prima facie* à croire à son contenu indexical ou a-compatibiliste. Comme ce contenu, avant tout analyse compatibiliste ou incompatibiliste, implique logiquement que nous ne sommes pas manipulés par des martiens, il faut bien (lorsque nous considérons ce scénario) que nous ayons une justification à le rejeter (à le tenir pour faux), faute de quoi nous perdrons toute légitimité à continuer de croire que nous sommes libres ; mais le simple fait de prendre conscience de l'implication logique $LA \rightarrow \sim MM$ suffit à nous donner une justification déductive pour rejeter le scénario de van Inwagen *sur la base* de notre justification *JLA*. Mais peut-on vraiment avoir une *connaissance* de $\sim MM$ sur la base d'une telle déduction ? Et même, n'est-il pas problématique de nous attribuer la moindre connaissance de $\sim MM$ puisqu'on a dit que ce scénario était stipulativement indétectable pour nous ? Ces questions sont tout à fait pertinentes, mais la façon la plus claire de les aborder est de remarquer qu'il s'agit de questions de second ordre, qui reposent essentiellement sur un principe de méta-cohérence du type $Jp \rightarrow JKp$. Et donc, il est plus clair de les présenter comme des éventuels *réfuteurs* fondationnels d'une justification *prima facie* déjà acquise au premier ordre (c'est-à-dire au niveau non réflexif), et acquise par simple transmission de la justification *prima facie* par déduction compétente. La solution de Pryor est donc non seulement la plus plausible, mais aussi la plus claire pour aborder ensuite les questions de méta-cohérence.

Pour conclure cette section sur l'objection de l'impression lourde, je rappellerai tout d'abord que la défense de l'introspectionnisme ne *dépend pas* de l'acceptation d'une solution déductiviste à la Pryor plutôt que d'une solution anti-transmission à la Wright : le simple parallèle entre argument de l'impression lourde et argument sceptique de la clôture suffit à montrer que *si* on veut éviter le scepticisme de la perception, on doit d'une manière ou d'une autre refuser la conjonction de principes qui génère l'objection de l'impression lourde. C'est-à-dire qu'on doit d'une manière ou d'une autre refuser que la justification soit close sous la déduction compétente, ou refuser que les négations d'hypothèses soient injustifiables. D'un point de vue strictement argumentatif, ce simple constat suffit à défendre l'introspectionnisme contre l'objection de van Inwagen, sans avoir besoin d'explorer la *réponse* anti-sceptique la plus appropriée et son application au cas de l'introspectionnisme. Mais pour pouvoir aborder les questions suivantes (les questions de justification de second ordre), il était utile d'exposer la solution de Pryor qui offre le cadre théorique le plus plausible et le plus efficace. Dans cette réponse, la justification *prima facie* se transmet toujours lors d'une déduction compétente, et par conséquent, les conséquences « lourdes » de l'expérience du libre arbitre sont justifiées *prima facie* au premier ordre, quitte à faire ensuite l'objet d'une réfutation de second ordre. Cette démarche, qui sépare les réfutations de premier ordre et celles de deuxième ordre, sera

répondre à une objection de (Coliva 2008, 237) contre la solution de Pryor : pour Coliva, si la solution de Pryor était correcte, on devrait s'attendre à ce que la preuve de Moore ne semble problématique *que* lorsqu'on pense au sceptique, alors que l'intuition d'échec est absolument directe et évidente pour tout un chacun, sans avoir besoin de faire un détour « par procuration » (vicariously) pour imaginer si elle va convaincre ou non quelqu'un d'autre (le sceptique). Pryor pourrait répondre que la preuve de Moore, tout en exhibant *une* structure déductiviste *possible* (celle reposant sur *cette* croyance particulière et nouvelle « voici une main »), échoue à exhiber un fondement nouveau et pertinent : c'est-à-dire que toute personne anti-sceptique écoutant son argument aura *déjà* refusé l'hypothèse sceptique sur la base d'une collection de croyances perceptives extrêmement vaste et impossible à identifier consciemment, et par conséquent, le fait de présenter *cette* nouvelle proposition perceptive semble entraîner l'implicature pragmatique (erronée) qu'il y aurait en elle quelque chose de particulier ou de nouveau permettant d'apporter à la croyance $\sim HS$ un renfort justificatif d'un nouveau type. Par conséquent, Pryor peut tout à fait accorder à Coliva que la preuve de Moore est dialectiquement inopérante *même* pour un interlocuteur non sceptique, car cette preuve échouera à lui apporter une justification significativement nouvelle pour la conclusion, ce qui est le but dialectique d'une preuve.

utile tout d'abord pour la justification « a-compatibiliste » et ses conséquences lourdes du type $\sim MM$; mais elle se révélera encore plus utile lorsque j'aurai défendu une analyse libertarienne du contenu de l'expérience E. A l'issue de cette analyse, on pourra conclure que la justification *prima facie* en faveur de LA transmet une justification *prima facie* à sa conséquence (lourde) qu'est l'indéterminisme, et la question sera ensuite de savoir s'il y a des réfuteurs directs (des preuves du déterminisme) ou des réfuteurs fondationnels (une impossibilité de connaître l'indéterminisme de cette manière-là). En attribuant une justification *prima facie* déductive en faveur de l'indéterminisme, mon intention ne sera certainement pas d'éluder ces questions, mais au contraire de leur donner le cadre théorique le plus clair possible.

III. Conclusion : la justification indexicale de premier ordre en faveur du Libre Arbitre

Dans le chapitre 1, j'ai donné une définition « descriptive » de l'expression « libre arbitre » en fixant son sens au moyen de l'expérience commune d'être libres E. Une fois que ce contenu conceptuel est fixé (mais pas encore *analysé*), la question est de savoir si nous avons (ou pourrions avoir) une justification à croire que ce contenu est instancié dans la réalité. C'est cette question que j'ai abordée dans ce chapitre 2, en partant d'une réponse naïve qui vient spontanément à l'esprit : nous sommes justifiés à accepter que le contenu de cette expérience est instancié... simplement par le fait que nous *avons* cette expérience. Et tout l'enjeu du chapitre a consisté à préciser cette réponse « introspectionniste », et à montrer que son discrédit contemporain était tout à fait illégitime. Les objections classiques à cette épistémologie introspectionniste sont fondamentalement des objections sceptiques, qui amènent à refuser non seulement la justification de la croyance introspective au libre arbitre, mais aussi la justification de nos croyances perceptives anodines. Certes ces objections ont une réelle force intuitive... mais c'est la force même du scepticisme ; et donc, si elles marchaient, elles permettraient seulement de prouver que la croyance introspective au libre arbitre est « aussi mal justifiée » que nos croyances perceptives communes, ce qui n'est pas une conclusion très problématique pour l'introspectionniste. Pour voir s'il y a des objections graves contre la croyance introspective au libre arbitre, il faut envisager les objections qui demeurent si le scepticisme général échoue, et donc il faut *présupposer* l'échec du scepticisme.

Une première objection anti-introspection repose sur l'idée qu'une croyance ne peut pas être justifiée par une « pure impression » ; mais j'ai montré que cette objection n'était en fait rien d'autre que l'acceptation d'une contrainte inférentialiste sur la justification qui génère l'argument sceptique de la régression. Toute réponse à l'argument de la régression permettra donc de répondre à l'objection de la pure impression. Cependant, la réponse la plus plausible semble être une forme de fondationnalisme, qui conserve une contrainte évidentialiste (internaliste) en considérant que les raisons d'une croyance ne sont pas toujours d'autres croyances mais parfois des expériences. Cela suffit pour répondre à l'objection de la pure impression, mais cela ne permet pas directement d'affirmer que l'expérience du libre arbitre donne une justification pour son contenu, car le fondationnalisme n'oblige pas à ce que *toutes* les expériences justifient. Il pourrait donc y avoir d'autres raisons de penser que *cette* expérience, celle du libre arbitre, n'est pas assez honnête pour fournir une justification. Les deux objections suivantes consistaient justement à trouver dans l'expérience du libre arbitre un point faible qui l'empêcherait d'entrer dans le clan des expériences justifiantes.

L'objection de l'impression faillible consiste à dire que l'expérience du libre arbitre n'apporte pas de justification à son contenu parce qu'il est *logiquement possible* qu'elle ne

soit pas véridique. Là encore, cette objection correspond à un argument sceptique très célèbre : l'argument infaillibiliste, qui n'admet comme expériences justifiantes que les sources infaillibles (introspection ? raison ?), et ne permet pas (*pace* Descartes) de récupérer le monde extérieur. Toute réponse au scepticisme infaillibiliste permettra de répondre à l'objection de l'impression faillible. J'ai montré cependant que la réponse la plus plausible consistait à adopter une forme de révisabilisme de la justification, qui s'exprime en termes de justification *prima facie* : une expérience faillible peut fournir une justification *prima facie* en faveur de son contenu, laquelle justification pourra parfois être réfutée par d'autres éléments évidentiels, soit des justifications plus fortes en sens contraire, soit des justifications qui attaqueront le fondement même de la justification initiale. Ainsi, l'expérience du libre arbitre peut fournir une justification *prima facie* en faveur de son contenu, et cette justification *prima facie*, si elle n'est pas réfutée, nous rendra *ultima facie* justifiés à croire que nous sommes libres.

L'objection de l'impression lourde, qui repose sur le scénario vaninwagenien de manipulation martienne, consiste à dire que le contenu de l'expérience E a des conséquences trop lourdes pour pouvoir apporter une justification à son contenu ; soit qu'elle n'apporte aucune justification, pas même *prima facie* ; soit plus probablement que la justification *prima facie* qu'elle apporte soit ultimement réfutée par la prise en considération des conséquences injustifiables qu'elle implique. Cette objection revient donc à ne considérer comme expériences (ultimement) justifiantes que les expériences (éventuellement faillibles) qui sont « légères », qui n'engagent à rien. Malheureusement, si on applique un tel critère de sélection pour retenir les croyances justifiantes, il faudra exclure en même temps les croyances perceptives : c'est précisément ce que fait voir l'argument sceptique de la clôture. Et donc, toute réponse à l'argument de clôture sera également une réponse à l'objection de van Inwagen. Mais lorsqu'on lit l'argument de clôture en termes internalistes (comme un argument de clôture de la *justification*), les solutions qui reviennent d'une manière ou d'une autre à refuser la clôture de justification sont peu plausibles, car il semble clair que la déduction compétente transmet toujours une forme de justification, fût-elle seulement *prima facie* et dialectiquement inopérante, de la prémisse à la conclusion déduite. Rien n'empêche, par conséquent, que l'expérience du libre arbitre apporte une justification *prima facie* à croire que nous sommes libres, laquelle croyance apporte ensuite une justification *prima facie* à croire la conséquence « lourde », à savoir que nous ne sommes pas manipulés par des martiens.

Rien de ce qui précède ne permet d'établir positivement que l'expérience du libre arbitre apporte une justification à son contenu : ce que j'ai montré, c'est qu'il n'y a aucune objection sérieuse contre l'idée très naturelle que notre expérience du libre arbitre nous donne au moins une certaine raison de croire à son contenu. Or si nous cherchons une raison possible de croire que « nous sommes libres » (dans le sens fixé par l'expérience), il semble que cette réponse introspectionniste est non seulement la plus simple et la plus spontanée, mais aussi la seule qui permette réellement de défendre, en l'état, notre croyance commune au libre arbitre, puisque le fondement *doxastique* de cette croyance se trouve manifestement dans notre expérience E ; pour que cette croyance soit *doxastiquement* justifiée (ou « bien fondée »), et non pas seulement *propositionnellement* justifiée (ou « justifiable »), il faut donc défendre que ce fondement doxastique constitue également son fondement justificatif... c'est-à-dire qu'il faut qu'une certaine version (bien comprise) de l'introspectionnisme soit vraie. Pour toutes ces raisons, il semble légitime de partir d'une conception introspectionniste de la justification du libre arbitre, aussi longtemps qu'on n'aura pas donné des raisons sérieuses de penser que l'expérience E ne *peut pas* conférer une justification. Or nous venons de voir qu'il n'y avait pas de raisons de penser cela.²¹⁵

²¹⁵ Tout du moins pas de raisons de premier ordre, nous verrons au chapitre suivant qu'il n'y a pas non plus de raisons de deuxième ordre ; et je traiterai les éventuelles objections de second ordre comme des arguments visant

Peut-on aller plus loin et argumenter *positivement* que l'expérience du libre arbitre confère une justification (*prima facie*) en faveur de son contenu ? Il y aurait deux manières de faire cela : soit établir que l'expérience du libre arbitre a une certaine caractéristique positive qui lui donne droit d'entrée dans le clan privilégié des expériences justifiantes ; soit établir que *toutes* les expériences doxastiques, quelles qu'elles soient, fournissent une justification *prima facie* en faveur de leur contenu. Si on suit le premier type de stratégie, on pourra par exemple mettre en avant le caractère commun de l'expérience E, son caractère naturel ou inéliminable, sa force intuitive, etc. La deuxième stratégie, défendue par Michael Huemer (2007) sous le nom de « conservatisme phénoménal charitable » (*compassionate phenomenal conservatism*), repose sur l'idée qu'on ne peut pas faire de discriminations entre les expériences pour ce qui est de la justification *prima facie*. Bien qu'aucune de ces deux stratégies ne soit indispensable dialectiquement à mon propos (pour les raisons que je viens d'exposer), je vais tout de même expliquer pourquoi la deuxième stratégie me semble la plus prometteuse *pour ce qui est de la justification*.

Le type de caractéristiques positives mises en avant par la première stratégie (notamment la naturalité) me semblent pertinentes lorsqu'il s'agit de déterminer (ou de présumer) quelles expériences constituent des sources de *connaissance* : on les retrouvera donc au chapitre suivant. Si S a une expérience doxastique qui n'a aucune caractéristique positive de ce genre (et qui n'est indirectement garantie par aucune source de connaissance), alors il ne sera peut-être pas légitime pour un observateur extérieur de tenir cette expérience comme une source de connaissance pour S. Mais encore une fois, cela ne pose de problème de *justification* que si S lui-même conçoit cette absence de caractéristiques positives. Autrement dit, la présence ou l'absence de ces caractéristiques positives constitue seulement à mon sens un possible renfort ou une possible réfutation *de second ordre* par rapport à une justification *prima facie* de premier ordre qui doit déjà être en place. Si on exigeait d'un sujet S qu'il conçoive les caractéristiques positives de ses expériences comme *condition préalable* pour que ces expériences fournissent la moindre justification de premier ordre, on tomberait dans un nouveau type de régression, dont je montrerai au chapitre suivant qu'il mène au scepticisme.

Par conséquent, il me semble que la seule stratégie viable, si l'on veut établir positivement que l'expérience du libre arbitre confère une justification *prima facie* en faveur de son contenu, c'est la stratégie « charitable » ou « anti-discrimination » de Michael Huemer. Il convient tout d'abord de distinguer le conservatisme phénoménal charitable et le dogmatisme, car les deux thèses sont étroitement apparentées et parfois confondues. Le dogmatisme est une thèse sur la *structure* de nos justifications *prima facie* : il revient à dire que certaines expériences peuvent conférer une justification *prima facie* en faveur de leur contenu (et en faveur des conséquences impliquées) *en amont* de toute question de second ordre sur la fiabilité de ces expériences. Une justification *prima facie* n'a pas besoin, structurellement, de reposer sur une raison de second ordre. Le conservatisme phénoménal charitable est une thèse sur l'étendue du domaine des expériences justifiantes : *toutes* les expériences fournissent une justification *prima facie*. Dans la version défendue par Michael Huemer, le conservatisme phénoménal implique et présuppose le dogmatisme, c'est-à-dire que toutes les expériences apportent une justification *en amont de toute question de second ordre*.²¹⁶ Mais l'implication inverse n'est pas vraie : un dogmatiste peut limiter sa thèse à une

à mettre en évidence un *réfuteur* (fondationnel) de la justification de premier ordre apportée par l'expérience E, et non pas comme des arguments visant à montrer que l'expérience E n'apporte *aucune sorte* de justification.

²¹⁶ De fait, une stratégie charitable et *non* dogmatique serait difficilement défendable, quoique logiquement possible : elle reviendrait à soutenir que toutes nos expériences sont (factuellement) dotées de caractéristiques de second ordre suffisamment solides pour pouvoir apporter une justification reposant sur ces caractéristiques. Mais

certaines catégories d'expériences (par exemple les expériences perceptives), en soutenant que les *autres* catégories d'expériences requièrent quant à elles un renfort préalable de second ordre avant de pouvoir conférer une justification, même *prima facie*. Il faut donc un argument supplémentaire pour passer du dogmatisme au conservatisme phénoménal.

Le premier argument de Huemer en faveur de la non discrimination des expériences doxastiques est très simple : il consiste à dire qu'il n'existe aucune manière satisfaisante de tracer la démarcation entre expériences justifiantes et expériences non justifiantes.²¹⁷ Une démarcation devrait avoir deux caractéristiques pour être satisfaisante : elle devrait être *non arbitraire* d'une part (c'est-à-dire que la démarcation proposée soit intuitivement pertinente pour la justification),²¹⁸ et elle ne devrait pas conduire au scepticisme (à cause de la démarche anti-sceptique déjà présentée). Nous avons déjà considéré deux critères de démarcation possibles, le critère d'infailibilité et le critère de « légèreté », mais nous avons vu que ces critères menaient tout deux au scepticisme de la perception. Une troisième démarcation, proposée par Bonjour et discutée par (Huemer 2007, 32–36), consisterait à privilégier les expériences introspectives en tant que telles, tout en admettant leur caractère faillible ; mais Huemer montre que, dans la mesure où elle accepte un faillibilisme de l'introspection, cette solution revient à faire une différence parfaitement arbitraire entre les expériences perceptives (non justifiantes, dans cette solution) et certaines expériences introspectives qui seront justifiantes quoiqu'elles soient fausses :

Le problème consiste à identifier un facteur pertinent-pour-la-justification qui puisse être présent dans le cas d'une croyance introspective fautive mais n'est pas présent dans le cas des croyances perceptives typiques. Il est difficile de voir ce qui pourrait justifier les croyances introspectives *fautes* si ce n'est le simple fait qu'il semble au sujet qu'elles sont vraies. (Huemer 2007, 35)

Une quatrième démarcation possible consisterait à limiter le dogmatisme aux expériences doxastiques qui proviennent de nos sens ; ce serait une version de la doctrine empiriste.²¹⁹ L'empirisme, pris comme doctrine de la justification interne, pose trois types de difficultés. Premièrement, il pose une difficulté symétrique à celle du rationalisme cartésien : en se limitant à l'introspection et à la raison, le rationalisme cartésien a toutes les peines du monde à « récupérer » une justification pour les croyances perceptives ; inversement, en limitant la justification aux expériences perceptives, un empiriste strict aura toutes les peines du monde à « récupérer » une justification pour nos croyances logiques ou mathématiques, qui semblent reposer sur des intuitions rationnelles et non pas sur des expériences perceptives. En refusant le moindre caractère justifiant aux intuitions rationnelles, l'empirisme strict semble mener à un scepticisme mathématique qui n'est pas plus acceptable que le scepticisme du monde extérieur. Un empiriste pourrait essayer de montrer que *certaines* intuitions rationnelles sont ultimement dérivées d'expériences perceptives – par exemple, l'intuition que « $1+1=2$ » à partir de perceptions du type « il y a un zèbre ici et un zèbre là, et en tout il y a deux « zèbres » (?) – et restreindre le caractère justifiant aux intuitions rationnelles ainsi dérivables.

il est évident que nous avons parfois des expériences doxastiques qui ne résistent pas une seconde à la prise en compte de leurs caractéristiques de second ordre.

²¹⁷ Ce type d'argument a été défendu un peu plus tôt par Joel Pust (2000, 5.3) pour défendre le caractère évidentiel des *intuitions* en philosophie. Pust rappelle pertinemment l'origine Reidienne de cette critique en « partialité illégitime » (*Undue Partiality*).

²¹⁸ « I take it that the critic who believes that only some appearances should be accepted as sources of justification incurs a burden of identifying an epistemically relevant difference between those appearances that do and those appearances that do not confer justification on their contents. » (Huemer 2007, 32).

²¹⁹ Dans la version plus traditionnelle, l'empirisme soutient que les *sens* sont les seules sources possibles de *connaissance*, et non pas que les *expériences sensibles* sont les seules sources possibles de *justification interne*.

Mais, c'est le deuxième problème, une telle discrimination *entre les intuitions rationnelles* serait tout à fait arbitraire, car le caractère dérivable ou non d'une intuition à partir de la perception est une question certainement absconse, qui échappera presque toujours à notre appréhension, et donc il semble que si on accepte un caractère internalistiquement justifiant à certaines intuitions rationnelles, leur seule caractéristique pertinente pour ce rôle internalistiquement justifiant ne pourra être que le fait que leur contenu *semble vrai*, une caractéristique qu'elles partagent avec toutes les autres intuitions rationnelles. Ou pour le dire autrement, pour sauver certaines intuitions rationnelles, la seule voie non arbitraire est celle défendue par George Bealer (1992; 1996), qui consiste à sauver toutes les intuitions rationnelles *en tant que telles*. Le troisième argument, et le plus célèbre, est l'argument de l'auto-réfutation : il est largement admis que le déclin de l'empirisme logique au XX^{ème} siècle a été lié, en partie, à un argument d'auto-réfutation. Le problème de l'empirisme logique est que la contrainte vérificationniste sur la justification est inconsistante avec l'*acceptation justifiée* de cette même contrainte ; l'empirisme logique est donc rendu injustifié par ses propres critères.²²⁰ Pour toutes ces raisons, il ne semble pas possible non plus de restreindre la catégorie des expériences doxastiques justifiantes aux seules expériences issues des sens. Et plus généralement, il ne semble pas y avoir de manière non arbitraire de tracer la limite entre expériences justifiantes et expériences non justifiantes, sauf à prendre des critères qui mènent au scepticisme. Donc, à moins d'être sceptique, il faut admettre que toute expérience doxastique, comme telle, confère au moins un certain de degré de justification *prima facie* en faveur de son contenu.

Huemer (2007, 39–44) donne un deuxième argument plus complexe en faveur du conservatisme phénoménal, un argument qui consiste à généraliser l'objection d'auto-réfutation habituellement adressée à l'empirisme logique. L'idée est que *toute* théorie alternative au conservatisme phénoménal aura pour conséquence de rendre injustifiée l'acceptation de cette théorie elle-même. De prime abord, cet argument semble trop fort : certes une théorie sera auto-réfutante si elle rejette comme injustifiantes le type d'expériences doxastiques sur lesquelles elle est elle-même fondée (ce qui semble bien être le cas de l'empirisme logique), mais ne peut-on pas imaginer une théorie qui accepte comme expériences justifiantes, par exemple, les types T1, T2, T3, exclusivement, tout en reposant elle-même sur une expérience de type T3 ? L'idée principale de Huemer est qu'une telle théorie, quoiqu'elle soit logiquement possible, ne serait pas empiriquement défendable à cause du *genre* de fondement doxastique sur lequel reposent nos croyances méta-épistémiques. Toute croyance en une théorie méta-épistémique (une théorie sur la nature des expériences qui confèrent une justification) sera doxastiquement fondée sur le fait que cette théorie nous *semble vraie*. Si l'empiriste a échoué à fonder sa propre croyance empiriste sur des expériences perceptives, ce n'est pas tellement parce que les expériences perceptives sont en défaut à cet égard, mais parce que le fondement effectif de la croyance empiriste est clairement ailleurs, à savoir dans le simple fait que cette théorie *semble vraie* à celui qui l'accepte. Et la croyance empiriste n'a de ce point de vue rien de particulier par rapport à d'autres croyances méta-épistémiques telles que nous pouvons *réellement* les entretenir.²²¹ Par conséquent, toute théorie méta-épistémique autre que le conservatisme phénoménal arrivera à la conclusion que son proprement fondement doxastique n'est pas un fondement

²²⁰ Ce type d'argument est présenté sous une forme assez traditionnelle par (Alston 2003). Les trois arguments anti-empiristes proposés par Bealer dans son article « The Incoherence of Empiricism » (Bealer and Strawson 1992) sont également présentés comme des réélaborations de l'idée que « denying that intuitions have evidential weight leads one to epistemic self-defeat. » (Bealer 1996, 124)

²²¹ Sauf, dit (Huemer 2007, n. 14), si quelqu'un les entretient par vœu pieux, par acte de foi, ou par dysfonctionnement cérébral, mais il est peu probable qu'un philosophe veuille défendre une méta-épistémologie de *cette* manière-là.

justificationnel ; cela ne permet pas de dire que la théorie en elle-même est proportionnellement injustifiée (ou injustifiable), mais cela implique que la *croyance* du théoricien est bel et bien doxastiquement injustifiée (ou « mal fondée »).²²²

Huemer montre donc que toute alternative au conservatisme phénoménal est probablement auto-réfutante, et qu'elle est en tout cas *soit* arbitraire dans son principe de discrimination, *soit* non arbitraire mais menant au scepticisme. Pour l'économie de ce travail, la conclusion la plus intéressante est celle du premier argument, qu'on peut aussi présenter ainsi : toute alternative sérieuse (non arbitraire) au conservatisme phénoménal mène au scepticisme. Cette formulation met en évidence l'apport du raisonnement de Huemer dans la démarche anti-sceptique que nous suivons ici. On pourrait en effet résumer ce raisonnement et l'appliquer à une défense directe de l'introspectionnisme, sous la forme de l'argument anti-sceptique suivant :

Argument anti-sceptique pour la Justification *prima facie* de l'expérience du libre arbitre

- (1) Nos croyances perceptives et mathématiques sont justifiées. (prémisse antisceptique)
- (2) Si nos expériences *comme telles* ne conféraient pas une justification *prima facie* en faveur de leur contenu, alors nos croyances perceptives et/ou mathématiques seraient injustifiées. (pas de critère non arbitraire pour sauver *telles* expériences)
- (3) Donc toute expérience (comme telle) confère une justification *prima facie* en faveur de son contenu. (conclusion intermédiaire au conservatisme phénoménal)
- (4) Donc l'expérience d'avoir le libre arbitre confère une justification *prima facie* à croire au libre arbitre.

L'idée introspectionniste selon laquelle notre expérience E nous donne une justification *prima facie* en faveur de son contenu n'est donc pas seulement une théorie qui vient spontanément à l'esprit, une théorie (la seule !) qui permet de satisfaire la « motivation doxastique » (et donc sauver notre croyance commune « en l'état »), et enfin une théorie non réfutée par les objections traditionnelles : c'est aussi une théorie pour laquelle on peut donner un argument *indépendant* de toute motivation liée au libre arbitre, à savoir l'argument du conservatisme phénoménal. Pour toutes ces raisons, j'admettrai positivement cette idée dans le restant de mon étude. Il est crucial de remarquer que cette conclusion intermédiaire n'est *pas* la thèse introspectionniste proprement dite : ce que je soutiens pour l'instant, c'est *simplement* que l'expérience du libre arbitre confère une justification *prima facie* positive en faveur de son contenu, tandis que la thèse introspectionniste défend une justification *ultima facie*. Pour conclure ce chapitre, je soulignerai donc une dernière fois le caractère très modeste de cette conclusion intermédiaire.

L'idée que nous avons une justification *prima facie* introspective en faveur du libre arbitre est une idée extrêmement modeste dans la mesure où une justification qui n'est que *prima facie* peut être réfutée de toutes sortes de façons, éventuellement de manière triviale et évidente, éventuellement même de manière constante. Plus précisément, la modestie dialectique de cette idée se remarque au fait que toute réticence qu'on pourrait avoir à accepter la justification pour telle croyance, peut être reformulée en termes de *réfutation* de la justification *prima facie*. Prenons un exemple : supposons que j'aie, de manière fugitive, vague et inexplicable, l'expérience doxastique que « il y a de la vie sur une planète située à un million d'années lumières de la terre » (proposition ET) ; et supposons que sur la base de cette

²²² Dans la suite de la discussion, Huemer considère une théorie méta-épistémique qui reviendrait à légitimer *de manière singulière* l'expérience doxastique qui fonde cette théorie. Huemer répond qu'une telle théorie ne serait probablement pas auto-réfutante, mais qu'elle souffrirait

expérience, je me mette à croire (ET). N'est-il pas évident que cette croyance serait parfaitement injustifiée pour n'importe quel sujet ? C'est effectivement évident, mais cela ne s'oppose absolument pas à l'idée que mon expérience doxastique confère « un certain degré de justification *prima facie* » en faveur de cette croyance. L'idée évidente que cette croyance est injustifiée provient de considérations qui sont très puissantes (le caractère fugitif et vague, le fait qu'il semble impossible d'imaginer que cette expérience apporte une *connaissance*, etc.) mais cela montre seulement qu'on est en présence d'une justification *prima facie* qui est en présence de réfuteurs très puissants (et ce probablement pour *toute* personne saine d'esprit). De même, dire que l'expérience E nous donne une justification *prima facie* en faveur de son contenu reste compatible avec toutes sortes de réfuteurs, aussi nombreux et aussi forts qui se puissent imaginer.

Mais malgré sa modestie, cette thèse est une base fondamentale pour aborder le reste de la discussion, c'est-à-dire justement la question de savoir s'il y a ou non des réfuteurs pour cette justification *prima facie*. Et pour aborder cette question de manière plus précise, j'ai posé deux distinctions. Tout d'abord la distinction entre *justification indexicale* (ou « a-compatible ») et *justification descriptive* (ou « analysée ») : dans la mesure où l'on peut être justifié à croire 'ceci est le cas' (en se référant au contenu non analysé de notre expérience E) sans être justifié à croire tel contenu conceptuel analysé de E, il est évident que les réfuteurs possibles ne seront pas les mêmes dans l'un et l'autre cas. Deuxièmement, j'ai distingué deux catégories de réfuteurs : *réfuteurs directs* (ou de premier ordre) et *réfuteurs fondationnels* (ou de second ordre). A partir de ces deux distinctions, on peut classer les réfuteurs possibles de la manière suivante :

Réfuteurs	direct	fondationnel
de la justification indexicale	ex : raison positive de croire que <i>MM</i> est le cas	ex : doutes positifs sur la possibilité de <i>KLA</i>
de telle justification analysée (ex : analyse libertarienne)	ex : preuve scientifique du déterminisme	ex : doutes positifs sur la possibilité de <i>KIndét.</i>

Dans ce chapitre 2, je me suis strictement attaché à la justification indexicale, et j'ai défendu qu'il n'y avait pas de raison de ne pas reconnaître une certaine justification *prima facie* de type indexical. Mais j'ai fait un peu plus que cela.

En défendant le principe de transmission de la justification *prima facie*, j'ai également posé les bases d'un passage possible de la justification indexicale à la justification analysée : en effet, si quelqu'une reconnaît que l'analyse libertarienne est l'analyse correcte du contenu de son expérience E, alors cette prise de conscience associée à la justification *prima facie* indexicale, lui donnera *ipso facto* (grâce au principe de transmission) une justification *prima facie* en faveur du contenu analysé. Notez qu'en sens inverse, en vertu du principe de clôture, cette analyse pourra aussi transformer un réfuteur de la justification analysée en réfuteur de la justification indexicale.²²³

Par ailleurs, j'ai également soutenu que la justification indexicale *prima facie* fournie par E n'était menacée par aucun réfuteur direct. Si l'on convient d'appeler « justification

²²³ Imaginez par exemple quelqu'un qui aurait des preuves imparables du déterminisme, et n'aurait pas encore analysé le contenu de l'expérience E : sa justification *prima facie* en faveur de E pourrait être non réfutée, et donc procurer une justification *ultima facie*. Mais imaginez à présent qu'il analyse le contenu de E, et l'analyse de manière incompatible : le principe de transmission nous dit que, dans ces circonstances, il acquerra une justification *prima facie* à croire à l'indéterminisme... mais cette justification *prima facie* restera virtuelle car elle sera dès le départ en présence d'un réfuteur direct (la preuve du déterminisme), et de même que la justification *prima facie* se transférera du contenu indexical au contenu indéterministe, de même l'absence de justification *ultima facie* se transférera en sens inverse du contenu indéterministe au contenu indexical.

pré-réflexive » une justification *prima facie* qui n'est pas réfutée au premier ordre (mais pourrait l'être au second), la conclusion précise de ce chapitre est donc la suivante :

(J_{pr}-Index-1) La croyance introspective que « nous sommes libres » (au sens indexical) est pré-réflexivement justifiée.

Le reste de mon étude consistera tout simplement à montrer qu'il n'y a pas non plus de réfuteurs des trois autres catégories : au chapitre 3, je soutiendrai qu'il n'existe pas de réfuteur de deuxième ordre pour la justification indexicale, de telle sorte que la conclusion générale de ma première partie sera que notre justification introspective à croire au libre arbitre n'est aucunement réfutée en amont d'une analyse conceptuelle particulière du libre arbitre. Et donc les seules réfutations possibles de l'introspectionnisme *devront* reposer sur une analyse particulière du contenu de l'expérience E. A partir de la deuxième partie, je m'appuierai donc sur une analyse particulière du contenu de E, en défendant une analyse libertarienne, pour montrer que cette analyse ne rencontre pas non plus de réfuteurs. La fin de la deuxième partie consistera à montrer qu'il n'y a pas de réfuteur direct du libertarianisme ; et la troisième partie consistera à montrer qu'il n'y a pas non plus de réfuteur fondationnel de la croyance introspective au libertarianisme, c'est-à-dire qu'il est possible de concevoir l'expérience du libre arbitre comme fournissant une connaissance du libertarianisme (et donc de l'indéterminisme). C'est donc cette dernière partie qui achèvera ma réponse au « Problème d'Intégration », mais les étapes précédentes auront apporté les ingrédients indispensables pour comprendre cette solution et sa place dialectique.

Chapitre 3 : Justification *prima facie* à croire que nous savons que nous sommes libres

Sir we *know* our will is free, and *there's* an end on't.
(Boswell 1946 AD 1769, *Ætat.* 60)

Why, sir, should I believe the faculty of reason more than that of perception; they came both out of the same shop.

Reid, *Inquiry*, VI, 20, "Of Perception in general"

I. Deux types d'objections méta-épistémiques à l'introspectionnisme	181
II. L'objection de l'impression non étayée	189
II.1) L'objection et son principe de méta-cohérence	189
II.2) Principe de méta-cohérence et scepticisme de la méta-régression	190
II.3) Solution : Méta-cohérence disjonctive et circularités épistémiques vertueuses	196
III. L'objection de l'impression non étayable	217
III.1) L'objection et son principe de détectabilité externaliste	217
III.2) L'argument sceptique du double engagement	220
III.3) Solution : sensibilité des expériences et sûreté des conséquences	222
IV. Conclusion : la justification indexicale de second ordre et le « point de départ épistémologique » du problème du libre arbitre	235

Dans le chapitre 2, j'ai défendu que notre expérience commune du libre arbitre nous donne une justification *prima facie* à croire que nous sommes libres, et que cette justification *prima facie* ne rencontre aucun réfutateur direct si l'on se place en amont de toute analyse conceptuelle du « libre arbitre ». En effet, si l'on se place en amont de la discussion compatibilistes/incompatibilistes, la seule chose qui pourrait contredire directement notre justification *prima facie* serait une indication (ou au moins une suspicion) que nous nous trouvons dans un scénario de manipulation, du type du scénario *MM* de van Inwagen. Mais il est évident que nous n'avons pas de réfutateur de ce type : rien ne nous justifierait à croire que nous sommes dans un scénario de manipulation, et il y a tout lieu de penser que nous sommes même positivement justifiés à croire le contraire. Mais il reste une deuxième manière dont notre justification pourrait être réfutée : si nous apercevions que notre expérience du libre arbitre *ne peut pas* fournir une *connaissance* de son contenu, alors cette considération de second ordre annulerait notre justification *prima facie* à croire au libre arbitre... ou en tout cas, il semble bien qu'elle l'annulerait, car comment pourrions-nous être justifiés à affirmer à *la fois* « nous sommes libres » et « nous ne savons pas que nous sommes libres » ? L'intuition sous-jacente est le principe de méta-cohérence, dont on a vu qu'il était au cœur du problème d'Intégration du libertarianisme : si le libertarien ne peut pas dire qu'il *connaît* l'indéterminisme, comment peut-il continuer à soutenir qu'il est libre (en un sens qui implique l'indéterminisme) ? Il semble qu'un principe de méta-cohérence l'*oblige* à admettre une

forme de connaissance de l'indéterminisme, ou bien à abandonner sa croyance libertarienne. Dans ce chapitre, je n'étudierai le principe de méta-cohérence que dans son application à la justification *indexicale*, donc en amont de toute analyse compatibiliste ou incompatibiliste. Même au niveau de la justification *indexicale*, il semble qu'un défenseur du libre arbitre (et notamment un introspectionniste « a-compatibiliste ») est obligé d'une manière ou d'une autre d'admettre que sa croyance au libre arbitre constitue une connaissance, ou d'abandonner finalement sa croyance. C'est donc cette question que je vais aborder à présent.

Pour le dire formellement, je vais passer des questions de type *Jp* aux questions de type *JKp*. Il faut noter que ces dernières ne sont pas seulement des questions « de deuxième ordre » ; ce sont aussi des questions « externe-interne », dans la mesure où elles abordent le problème (externe) de la connaissance *du point de vue* de la justification (interne) que nous avons à y croire.²²⁴ Et ce qui m'intéressera plus particulièrement, ce sera d'une part la *possibilité* de défendre une thèse du type *JKLA* (pour la justification *indexicale* introspective du libre arbitre), et d'autre part la *nécessité* de défendre une telle thèse pour pouvoir préserver la justification *JLA*. La question de nécessité correspond au principe de méta-cohérence dont on peut donc donner comme première formulation :

$$(MC) Jp \Rightarrow JKp$$

Cette discussion sur la justification *JKLA* *indexicale* aura deux enjeux fondamentaux. Le premier enjeu concerne la justification *JLA* *indexicale* ; le second concerne la justification *JKLA* analysée.

Le premier enjeu d'une discussion (et éventuellement d'une défense) de *JKLA* *indexicale*, c'est qu'une *absence* de justification *indexicale* de second ordre pourrait constituer comme on l'a vu une *réfutation* de la justification *indexicale* *prima facie* *JLA*. Autrement dit, un introspectionniste a-compatibiliste pourrait être amené à abandonner sa croyance au libre arbitre parce qu'il ne serait pas en mesure de soutenir qu'elle constitue une connaissance. Donc notre défense de la justification *prima facie* *JLA* restera bien fragile (ou en tout cas sans grande conséquence) s'il y a un réfuteur de second ordre évident. Or nous avons vu au chapitre 2 que certaines intuitions présentes dans les arguments anti-introspection suggéraient l'existence d'un problème externe-interne : nous avons vu qu'une partie de la motivation anti-introspectionniste semblait venir de la difficulté à concevoir que nous *savons* (externalistiquement) que nous ne sommes pas manipulés par des martiens (intuition de van Inwagen), ou en tout cas de la difficulté à concevoir que nous le savons *sur la base de notre expérience* (intuition de Wright). Dans les deux cas, il semble donc y avoir une intuition selon laquelle l'introspection ne peut pas donner une *connaissance* du libre arbitre (c'est-à-dire une intuition du type $\sim KLA$) et si l'on a une conscience justifiée de ce fait (c'est-à-dire si l'on a $J\sim KLA$), on aura certainement un réfuteur (fondationnel) de la justification *prima facie* (introspective) *JLA*, avant même toute analyse compatibiliste ou incompatibiliste. Le premier enjeu de la discussion de la justification *indexicale* *JKLA* est donc de défendre *JLA* elle-même contre une dernière salve d'objections, les objections de second ordre. Et ma réponse suivra très exactement le type de raisonnement du chapitre 2 : les objections de second ordre contre la justification (*indexicale*) *JLA* ont une certaine force intuitive, mais il est aisé de montrer

²²⁴ Il existe également des questions « de second ordre » qui sont interne-interne (sommes-nous justifiés à croire que notre expérience du libre arbitre nous donne une justification à croire au libre arbitre ?) et d'autres qui sont externe-externe (pouvons-nous avoir une connaissance – une croyance fiable – du fait que notre expérience du libre arbitre constitue une source de connaissance ?). Si je pars de la question *JKp* plutôt que des questions *JJp* et *KKp*, c'est d'abord parce qu'elle est au cœur du problème d'Intégration : le problème d'Intégration est un problème de *rationalité* lié à des considérations sur la possibilité de *savoir* un certain contenu. Mais une deuxième raison apparaîtra au cours de ce chapitre : c'est que les questions *JKp* sont aussi au cœur des arguments sceptiques, bien plus que les questions *JJp* et *KKp* sur lesquelles j'aurai l'occasion de revenir.

qu'elles correspondent à des objections sceptiques radicales, et par conséquent, elles ne *peuvent pas* constituer des objections pertinentes contre la justification introspective (indexicale) du libre arbitre. En suivant la démarche anti-sceptique, le défenseur de l'introspectionnisme « a-compatibiliste » pourra se satisfaire de n'importe quelle solution anti-sceptique et l'appliquer au cas particulier de l'expérience du libre arbitre. J'envisagerai en fait deux types d'objections anti-introspectionniste de second ordre (et leurs objections sceptiques parallèles). Mais fondamentalement, les deux grandes solutions anti-sceptiques consisteront soit à refuser ou minimiser la méta-cohérence (et soutenir que la *JLA* n'est *pas* réfutée par les problèmes de second ordre), soit à refuser ou minimiser les problèmes de second ordre (en préservant l'idée que *si* il y en avait, la *JLA* serait menacée). J'expliquerai ensuite pourquoi le premier type de réponse (le refus de la méta-cohérence) me semble tout à fait problématique en lui-même ; si ce refus est problématique en lui-même, il l'est surtout pour l'économie de cette étude dont le problème central, le problème d'Intégration, repose entièrement sur l'acceptation du principe de méta-cohérence : si l'on refuse la méta-cohérence, alors un libertarien pourrait très bien se contenter de croire à l'existence d'actes indéterminés sans se soucier aucunement de la possibilité (éventuelle ou établie) que cette croyance soit totalement hasardeuse du point de vue épistémique externaliste. Notez qu'on aurait alors réussi à fournir une défense de l'épistémologie libertarienne, une défense reposant sur l'idée que les problèmes d'Intégration, d'une manière générale, ne sont pas des problèmes du tout puisque le principe de méta-cohérence est faux. Toute mon étude repose sur la conviction que les problèmes d'intégration, et notamment l'intégration du libertarianisme, sont des problèmes réels qui méritent une réponse, et donc la solution anti-sceptique qui repose sur le refus de méta-cohérence sera clairement non pertinente pour la suite de la discussion.

Le deuxième enjeu d'une discussion de la justification indexicale *JKLA*, c'est justement d'apporter un premier élément à l'intégration du libertarianisme. L'intégration du libertarianisme suppose une justification de deuxième ordre de type *JKLA*, mais pour un contenu *analysé* (et analysé en termes incompatibilistes). De la même façon que la justification indexicale *J_iLA* constitue dans ma démarche une étape fondamentale pour rendre compte ensuite de la justification *J_DLA* en faveur du contenu après analyse conceptuelle, de même la justification indexicale *J_iKLA* me semble une étape décisive pour rendre compte d'une justification de second ordre *J_DKLA*. Pour le dire simplement, si avant toute analyse conceptuelle nous sommes justifiés à tenir notre introspection du libre arbitre comme source de *connaissance*, alors nous aurons, même après analyse du contenu, une certaine justification au moins *prima facie* à considérer que nous connaissons par là le contenu analysé. Bien sûr, cette justification *J_DLA* pourra à son tour être réfutée de différentes façons, mais le point fondamental est de reconnaître qu'elle aura *au moins* une justification *prima facie*.

Pour résumer, la question de la justification indexicale *J_iKLA* a deux enjeux indirects sur la question de savoir si le libertarianisme est ultimement rationnel, c'est-à-dire sur *J_DLA*. Le premier passe par le chemin de *J_iLA* : en effet, *J_iKLA* semble utile pour défendre *J_iLA* (méta-cohérence), et *J_iLA* semble utile pour établir *J_DLA* (par transmission). Le second passe par le chemin de *J_DKLA* : en effet, *J_iKLA* semble utile pour établir *J_DKLA*, et *J_DKLA* semble utile pour défendre *J_DLA* (problème de l'Intégration).

Dans ce chapitre, je m'attacherai principalement au premier enjeu, c'est-à-dire à la manière dont la question méta-épistémique pourrait réfuter la justification *prima facie* *J_iLA* établie au chapitre précédent, et ainsi ruiner l'introspectionnisme avant même toute analyse particulière du concept de libre arbitre. Pour ce faire, je distinguerai d'abord deux versions de l'objection anti-introspectionniste méta-épistémique : l'objection de l'impression non-étayée et l'objection de l'impression non-étayable. La première objection consiste à dire que nous manquons de raisons de penser que notre expérience du libre arbitre est externalistiquement fiable, et conclut que nous ne sommes pas justifiés à accepter son contenu. La seconde

objection part de la prémisse plus forte que nous avons des raisons positives de penser que notre expérience du libre arbitre ne peut pas constituer une source de connaissance, et en tire la même conclusion que notre justification *prima facie* est ainsi (fondationnellement) réfutée. Je montrerai ensuite comment ces objections correspondent à deux formes classiques d'arguments sceptiques radicaux, ce qui leur enlève toute valeur dialectique contre l'introspectionniste « a-compatibiliste », tout du moins dans la démarche anti-sceptique adoptée ici. Je montrerai enfin quelle réponse anti-sceptique est la plus plausible à mes yeux, et la plus utile pour une application à la question du libre arbitre. La conclusion générale de ce chapitre sera donc que la justification *prima facie* indexicale fournie par l'expérience du libre arbitre n'est pas plus réfutée au deuxième ordre qu'elle n'est réfutée au premier ordre, c'est-à-dire qu'elle n'est pas réfutée *du tout* en amont de l'analyse précise du concept de libre arbitre. Cette « conclusion » n'est à mes yeux que le *point de départ* épistémologique de toute discussion philosophique bien posée sur le libre arbitre.

I. Deux types d'objections méta-épistémiques à l'introspectionnisme

Notre justification *prima facie* à croire au libre arbitre n'est-elle pas réfutée en son fondement ? N'est-elle pas réfutée par une incapacité à constituer un fondement de *connaissance* ? Ces objections de type fondationnel à l'introspectionnisme sont certainement parmi les plus importantes. Nous les avons vues apparaître, déjà, dans la discussion sur les objections de premier ordre : si une impression révisable peut fournir une justification (*prima facie*), ne faut-il pas tout de même une fiabilité externe pour qu'il y ait *connaissance* ? si nous sommes sans doute *justifiés* à admettre la conséquence lourde de notre impression (que nous ne sommes pas manipulés par des martiens), n'y a-t-il pas tout de même un problème à dire que nous en avons la *connaissance* ? Si je me suis efforcé de montrer que dans leur sens obvie les objections de premier ordre ne posait pas réellement problème, il n'en reste pas moins qu'elles « pointent » vers des problèmes de second ordre plus difficiles. Même l'objection de la « pure impression » pourrait être réinterprétée comme une objection de second ordre : une « pure impression » serait alors une impression qui n'est « étayée » par aucune raison de penser qu'elle est fiable. Dans sa forme la plus évidente, le problème fondationnel peut apparaître dans un dialogue extrêmement simple entre un introspectionniste (I) et une anti-introspectionniste (AI) :

AI : Qu'est-ce qui te permet de dire que tu es libre ?

I : C'est que j'en fais l'expérience.

AI : Mais cette expérience n'a aucune raison d'être une expérience fiable, et on a même toutes les raisons de penser le contraire : donc quand bien même tu aurais raison, ce serait par pur hasard.

Mais si ce dialogue est simple en apparence, et soulève manifestement un problème, ou une question légitime, il n'est pas si simple d'identifier la nature exacte du problème soulevé. Aussi, dans cette section, je vais commencer par identifier précisément le problème en éliminant certaines interprétations erronées ; une fois que le problème fondationnel aura été clairement identifié, je proposerai une classification des différentes versions du problème.

Il y a principalement deux manières de *mal* comprendre le problème fondationnel, tel qu'il est soulevé par le petit dialogue ci-dessus. C'est-à-dire qu'il y a deux manières de mal

identifier le principe épistémologique sur lequel s'appuie implicitement l'anti-introspectionniste AI pour mettre l'introspectionniste I dans l'embarras.

La première interprétation erronée, c'est de croire qu'il s'agit d'un simple problème de connaissance (d'un problème externe de premier ordre). Apparemment, la deuxième remarque de AI se situe simplement à ce niveau : AI fait observer que, selon toute apparence, l'expérience du libre arbitre ne fournit pas un fondement de connaissance. Autrement dit, sa remarque est, formellement, du type $\sim KLA$. Et il est clair que AI prétend avoir ainsi soulevé un problème qui *doit remettre en question* la croyance de I au libre arbitre, c'est-à-dire qu'il prétend avoir mis en évidence un embarras pour la *justification* de cette croyance. On doit donc avoir une conclusion de la forme $\sim JLA$. Si l'on s'en tient à cette analyse minimale, il pourrait sembler que AI passe de la prémisse $\sim KLA$ à la conclusion $\sim JLA$ au moyen du principe épistémologique suivant :

$$(JK) Jp \Rightarrow Kp$$

c'est-à-dire

(JK) S ne peut être justifié à croire que p (sur un fondement épistémique F) que si S sait que p (sur la base de F).

Mais ce principe est évidemment faux, compte tenu de la nature internaliste de la justification et de la nature externaliste de la connaissance. Certes, on *pourrait* employer le mot « justifié » dans un sens qui vérifie (ou impose) cette condition, mais si l'on parle de « l'impeccabilité épistémique interne », il est évident que ce statut peut être possédé par une croyance quand bien même elle serait produite par des mécanismes systématiquement trompés. C'est le cas notamment, comme nous l'avons vu en introduction (p.24), des cerveaux manipulés dont toutes les activités épistémiques seraient internalistiquement identiques aux nôtres. Si l'argument de AI était de cette forme, un troisième interlocuteur (T) pourrait fort bien lui tirer la manche et lui dire :

T : Admettons que l'expérience de I soit sans connexion aucune avec le fait d'être libre, cela ne l'empêche pas, du point de vue interne de I, de fournir une *justification*.

La réponse de T est parfaitement légitime, et montre que l'argument anti-introspectionniste ne peut par reposer sur le principe (JK). Mais cette réponse permet aussi de faire apparaître la vraie nature du problème : car, s'il est clair que la réponse est pertinente *lorsqu'elle est assertée par un troisième interlocuteur*, il est clair aussi que cette réponse serait extrêmement problématique, et même inacceptable, *si elle était assertée par I lui-même*. I ne pourrait pas répondre :

I : Admettons que mon expérience soit sans connexion aucune avec le fait que je sois libre, cela ne l'empêche pas (de mon point de vue interne) de me justifier à croire que je suis libre.

Si la réponse de I est inacceptable alors que celle de T ne l'est pas, c'est parce que cette réponse fait apparaître un *nouvel* élément dans le « point de vue interne » de I, à savoir la *croyance* que « l'expérience du libre arbitre est hasardeuse ». Cette croyance est une croyance méta-épistémique, ou de second ordre, qu'on peut noter : $B\sim KLA$. Autrement dit, le problème que AI essaye de soulever est *essentiellement* un problème de second ordre : en faisant remarquer $\sim KLA$ son intention n'est pas de mettre en évidence une proposition logiquement

incompatible avec *JLA*, mais de produire dans le système cognitif de I la croyance de second ordre $B \sim KLA$; c'est cette *croyance*, ou peut-être seulement l'absence d'une croyance positive en sens contraire ($\sim BKLA$), qui doit donner lieu à un embarras pour l'introspectionniste.

La deuxième interprétation erronée du problème méta-épistémique, c'est de croire que la croyance de second ordre $B \sim KLA$ est problématique pour des raisons de *connaissance*, et qu'elle soulève ainsi un problème de *connaissance* de second ordre (de type KK). Dans cette interprétation, l'anti-introspectionniste AI chercherait par exemple à faire remarquer que I, quand bien même son expérience serait fiable, n'est pas en mesure de connaître cette fiabilité, donc il ne sait pas qu'il sait : $\sim KKLA$. Et pour entraîner un embarras, il s'appuierait ensuite sur un principe épistémologique qu'on appelle la « luminosité » de la connaissance :

$$(KK) Kp \Rightarrow KKp$$

c'est-à-dire

(KK) Savoir, c'est savoir qu'on sait.

ou encore (pour être plus précis dans notre contexte) :

(KK') La croyance de *S* que *p* (sur un fondement épistémique *F*) ne peut constituer ultimement une connaissance que si *S* sait que son fondement *F* est une source capable de délivrer de la connaissance.

Si l'objection anti-introspectionniste était de cette nature, elle serait très faible, pour plusieurs raisons. D'abord parce que le principe (KK) est notoirement problématique en épistémologie.²²⁵

Mais surtout parce qu'il reste strictement externaliste : à supposer que AI parvienne à établir $\sim KKLA$, toute proposition de la forme $K\phi$ ou $\sim K\phi$ est une thèse sur les mérites épistémiques *externes* de la croyance ϕ (quelle que soit la nature et le degré de cette croyance). Or si nous sommes passés du premier ordre ($\sim KLA$) au second ordre ($B \sim KLA$), c'était précisément pour faire apparaître un problème *interne* dans l'attitude épistémique de I. La proposition $\sim KKLA$ n'est donc pas encore le bon candidat : pour que ce nouveau problème externe pose un problème interne, il faudrait à son tour qu'il soit *cru* par I, c'est-à-dire qu'il faudrait passer au troisième ordre $B \sim KKLA$, qui à son tour serait sans doute interprété comme un problème de *connaissance* du troisième ordre $\sim KKKLA$, etc. Il semble alors qu'on s'engage dans une régression vicieuse des niveaux de connaissance, vicieuse non pas pour I, mais pour l'objecteur, qui n'arrivera jamais à un niveau capable de fonder son objection de manière internaliste. Compte tenu de ces deux objections, un troisième interlocuteur pourrait de nouveau intervenir pour dire à AI :

²²⁵ Le problème du principe (KK) est qu'il semble insatisfaisable pour deux raisons principales. La première raison est que, dans sa version la plus simple, il entraîne de manière évidente une régression à l'infini des niveaux de connaissance. La deuxième raison, plus complexe, a été mise en avant par (Williamson 2000, sec. 4.3 & 5.1). L'argument de Williamson est difficile à résumer ; disons simplement qu'il repose sur la notion de « vague » telle qu'on la rencontre dans les problèmes de sorite. D'après Williamson, si l'on acceptait le principe (KK), on ne pourrait pas rendre compte du fait qu'un agent peut savoir, à vue d'œil, qu'un arbre mesure plus d'un millimètre, et qu'il ne peut pas savoir, à vue d'œil, que cet arbre mesure exactement 5m 45cm 23mm. En l'absence d'un élément de discontinuité entre 1mm et 5,4523m, le principe (KK) entraînerait une « propagation » de l'ignorance des 5,4523m, et faire ainsi que l'agent « ignore » également que l'arbre mesure plus d'un millimètre. Pour le détail de l'argument, je ne peux que renvoyer au texte de Williamson.

T : Admettons que I ne *sache pas qu'il sait* (que LA), cela ne l'empêche probablement pas de *savoir* (que LA) – puisque (KK) est très problématique – et même si cela l'empêchait de savoir, cela ne l'empêcherait pas d'être justifié.

Cette nouvelle réponse est parfaitement légitime, mais elle passe de nouveau à côté du vrai problème soulevé par AI : le vrai problème est que cette réponse n'est pas possible *pour I* lui-même. C'est-à-dire que le vrai problème est non seulement un problème de second ordre (qui engage la croyance de second ordre BKLA), mais surtout un problème externe-interne – qui concerne la *justification* de cette croyance de second ordre (JKLA), et non son statut de connaissance (KKLA). En somme, le vrai problème soulevé par le dialogue tout simple entre l'introspectionniste I et l'anti-introspectionniste AI repose bel et bien sur quelque chose comme le principe de méta-cohérence annoncé au début de ce chapitre :

(MC) $Jp \Rightarrow JKp$

Les distinctions qui précèdent peuvent sembler inutilement subtiles, mais elles ont pour but de rejeter comme non pertinentes deux réactions assez courantes à ce type de dialogue : la réponse « mais vous confondez notions internes et notions externes » (or il n'y a ici nulle confusion, mais une *intégration* des considérations externes comme contenu dans la sphère interne) et la réponse « mais on n'a pas besoin de *savoir qu'on sait* » (or l'anti-introspectionniste ne demande pas à son interlocuteur de « savoir qu'il sait », ce qui serait sans doute une exigence trop forte, mais seulement d'être prêt à *croire de manière justifiée* qu'il sait).

Une manière équivalente de conceptualiser le problème méta-épistémique est d'utiliser la distinction de Duncan Pritchard (2005, chap. 6) entre deux types de « hasard épistémique », le hasard épistémique *véritique* et le hasard épistémique *réflexif*. Le hasard épistémique *véritique* est celui que nous avons défini en introduction (p. 19) comme constituant le cœur du concept de connaissance. Comme on l'a vu, il s'agit de l'idée (purement externaliste) d'une absence de dépendance entre le fait qui est cru (p) et le fait qu'est la croyance (Bp). On peut l'exprimer de manière assez simple :

Hasard épistémique *véritique*

C'est un pur hasard si la croyance de S que p est vraie.²²⁶

Dans la mesure où ce type de hasard est externaliste, un agent S peut avoir une croyance qui n'est pas hasardeuse (véritiquement), mais *ignorer* qu'elle n'est pas hasardeuse, ou même être convaincu qu'elle est hasardeuse. Dans ce cas, la croyance de S sera considérée par Pritchard comme non affectée de hasard *véritique*, mais affectée d'un *autre* type de hasard épistémique, internaliste cette fois, qu'on peut exprimer ainsi :

Hasard épistémique *réflexif*

Du point de vue de S lui-même, c'est (ou ce pourrait être) un pur hasard que sa croyance que p soit vraie.²²⁷

Remarquez que, dans cette définition, le mot « hasard » du *definiens* correspond au hasard *véritique* (c'est-à-dire fondamentalement la question de *connaissance*) qui se trouve

²²⁶ Voir (Pritchard 2005, 146).

²²⁷ Voir (Pritchard 2005, 175).

« internalisé » du point de vue du système épistémique de S (et pose donc potentiellement un problème de *justification*).²²⁸

La distinction entre hasard véritique et hasard réflexif sera plus claire en prenant des exemples. Comme on vient de le voir, la distinction est particulièrement utile pour les cas de croyances qui sont fiables (ou véritiquement non hasardeuses) *à l'insu* du sujet qui les croit (et qui, partant, sont réflexivement hasardeuses). Les exemples les plus célèbres pour ce genre de situation sont les exemples de « clairvoyance » développés par Laurence Bonjour (Bonjour 1985, sec. 3.3; Bonjour and Sosa 2003, 2.2) :

Scénario d'Amanda

Imaginez qu'Amanda est douée d'une capacité fiable de clairvoyance dans certaines conditions spécifiques à l'égard d'un certain domaine de questions. Grâce à certains processus causaux qui ont pour l'instant échappé à l'investigation scientifique, des croyances dans ce domaine lui arrivent de manière très vive et spontanée dans ces conditions, et ces croyances sont généralement, ou même peut-être constamment, vraies. Amanda, pourtant, bien qu'elle accepte en général les croyances en question, n'a jamais cherché empiriquement si ces croyances étaient vraies, et elle n'a pas non plus reçu d'autrui des données empiriques indiquant la vérité de ces croyances, et donc indiquant sa fiabilité générale. (Bonjour and Sosa 2003, 28)

Un peu plus loin, Bonjour radicalise le scénario de la manière suivante :

*Scénario d'Amanda**

Imaginez qu'Amanda – en plus de n'avoir aucune raison à sa disposition qui soutienne positivement la thèse que les croyances en question sont produites de manière fiables ou ont des chances d'être vraies – possède en outre un ensemble de données qui semblent prouver de manière convaincante la thèse *contraire*. (*ibid.*)

Dans la terminologie de Pritchard, les croyances clairvoyantes d'Amanda sont véritiquement fiables mais réflexivement hasardeuses (parce qu'elle n'a aucune indication pour, voire des indications contre, leur fiabilité véritique).²²⁹ Dans la terminologie adoptée jusqu'ici, le cas serait décrit ainsi : la justification *prima facie* qu'Amanda reçoit de ses impressions doxastiques (vives et spontanées) se trouve réfutée par les considérations de second ordre. Les objections anti-introspection méta-épistémiques reviennent donc à dire que notre croyance introspective au libre arbitre est aussi problématique que les croyances clairvoyantes d'Amanda : quand bien même elle serait fiable du point de vue externaliste, elle reste

²²⁸ Au sens strict, il n'y a donc pas « deux types de hasard », mais un seul type de hasard pris soit en réalité, soit comme objet d'une croyance ou d'un système représentationnel. De la même façon, il n'y a pas « deux types de Parthénon », le Parthénon qui est sur l'Acropole, et celui qui est dans l'esprit de Socrate ; il y a un seul type de Parthénon qui peut être considéré soit en lui-même, soit comme représenté dans l'esprit de Socrate. Ainsi, une croyance « réflexivement hasardeuse » n'est pas *réellement* hasardeuse, elle est seulement *tenue pour* hasardeuse (ou possiblement hasardeuse) par S. Aussi, lorsque j'utiliserai le vocabulaire du « hasard » épistémique sans plus de précision, j'aurai toujours en vue le hasard « véritique »... c'est-à-dire le hasard véritable.

²²⁹ On peut aussi donner des exemples du cas inverse, c'est-à-dire des croyances qui sont véritiquement hasardeuses quoique réflexivement fiables : c'est précisément le cas des cerveaux manipulés qui auraient les mêmes données sensorielles et les mêmes croyances que nous. Ils auraient très certainement des raisons de penser que leurs facultés sont fiables, bien qu'en fait elles ne le soient pas. Cet exemple n'est pas sans poser de problème cependant, car il *présuppose* que nous avons des raisons de croire à la fiabilité de nos facultés ; or c'est précisément ce que conteste le sceptique lorsqu'il met en avant le cas des cerveaux dans des cuves. Je ne l'avance donc pas ici *pour répondre au sceptique*, mais seulement *pour donner un exemple* de ce qui pourrait être décrit, de manière au moins concevable, comme un cas de hasard véritique sans hasard réflexif. Pour la discussion avec le scepticisme proprement dit, je l'aborde dans les sections suivantes de ce chapitre.

réflexivement hasardeuse parce que nous n'avons pas les moyens de justifier (et nous avons peut-être les moyens de réfuter) sa fiabilité externe.

Il y a une troisième (et dernière) manière de décrire le problème méta-épistémique (et cette troisième description a l'avantage de bien souligner le caractère internaliste du problème), c'est de le comparer avec le principe de clôture de la justification. Au chapitre 2 (p.150), j'ai présenté le principe de clôture de la justification comme un « principe de transfert » du statut épistémique (interne) d'une croyance p à une croyance q . C'est-à-dire qu'au sein de l'espace épistémique d'un agent S , pour qu'une certaine « position épistémique » ait un certain statut épistémique interne (par exemple : *ultima facie* justifiée), il est nécessaire qu'une *autre* « position épistémique » ait un certain statut épistémique (éventuellement le même statut que la première). Je noterai $\langle p \rangle^S$ la position épistémique correspondant à la proposition p dans l'espace épistémique d'un agent S ²³⁰, pour ne pas la confondre avec la proposition p elle-même, ni avec la *croyance* que p (notée Bp) : la même position $\langle p \rangle^S$ peut faire l'objet de croyance ou pas (Bp ou $\sim Bp$), de justification ou pas (Jp ou $\sim Jp$), etc. Dans le cas de la clôture, le transfert s'effectue entre deux positions $\langle p \rangle^S$ et $\langle q \rangle^S$ lorsque S a conscience de pouvoir déduire q de p . De même, le principe de méta-cohérence peut être décrit comme un principe de transfert entre la position épistémique $\langle p \rangle^S$ et la position $\langle Kp \rangle^S$ au sein de l'espace épistémique de S : pour que la première ait un certain statut interne positif (tout spécialement : la justification), il faut que la seconde ait aussi un certain statut interne particulier (par exemple : ne pas être tenue pour fausse). Les différentes versions du principe de méta-cohérence se distinguent ensuite en fonction des statuts exacts de $\langle p \rangle^S$ et $\langle Kp \rangle^S$ qu'ils mettent en corrélation. On pourrait entrer dans une classification très détaillée de ces différents principes, mais la distinction la plus importante pour notre propos est la distinction entre une version *positive* et une version *négative* du principe. La version négative du principe dit simplement que lorsque la position $\langle Kp \rangle^S$ est occupée par la croyance justifiée que $\sim Kp$, c'est-à-dire lorsqu'on a $J\sim Kp$, alors la position $\langle p \rangle^S$ ne peut pas avoir le statut « justifié » (*ultima facie*) : quand bien même S aurait une justification *prima facie* à croire que p , celle-ci se trouve (fondationnellement) réfutée. Autrement dit :

$$\begin{aligned} &(\text{MC-}) J\sim Kp \Rightarrow \sim Jp \\ &\text{ou encore} \\ &(\text{MC-}) Jp \Rightarrow \sim J\sim Kp \end{aligned}$$

Cette contrainte est « négative » dans le sens où elle peut être satisfaite par une simple *absence* de justification en faveur de $\sim Kp$. La contrainte positive va plus loin et soutient que le statut justifié de $\langle p \rangle^S$ requiert *positivement* une certaine justification pour la position $\langle Kp \rangle^S$, c'est-à-dire qu'en l'absence de ce statut positif en $\langle Kp \rangle^S$, un tenant du principe positif considèrera la croyance que p comme réfutée. Autrement dit :

$$\begin{aligned} &(\text{MC+}) \sim JKp \Rightarrow \sim Jp \\ &\text{ou encore} \\ &(\text{MC+}) Jp \Rightarrow JKp \end{aligned}$$

²³⁰ Si l'on avait une conception hyper-réaliste de l'espace épistémique d'un agent – par exemple, sous la forme d'un système neuronal dont différentes parties représentent différentes propositions, on pourrait dire que la position épistémique $\langle p \rangle^S$ est cette partie du système neuronal de S qui représente spécifiquement la proposition p . Ce genre de théorie a peu de chance d'être vrai ; par conséquent, la notion de « position épistémique » doit plutôt être comprise comme une fiction théorique pour décrire l'ensemble des attitudes ou dispositions épistémiques qu'un agent S peut avoir à l'égard d'une proposition p .

(On s'aperçoit ici que la première formulation proposée pour le principe de méta-cohérence correspondait en fait à ce principe *positif* de méta-cohérence, c'est-à-dire à une version déjà assez forte du principe.) Cette distinction entre méta-cohérence positive et méta-cohérence négative est au cœur de la classification des objections méta-épistémiques que je vais proposer à présent.

Parmi les différentes objections méta-épistémiques qu'on peut soulever contre l'introspectionnisme du libre arbitre, il faut donc d'abord distinguer les (tentatives de) réfutations *positives* et les (tentatives de) réfutations *négatives*. Les premières partent de la prémisse (forte) que nous sommes positivement justifiés à croire que notre expérience E n'est pas fiable (soit $J \sim Kp$), et peuvent ainsi réfuter l'introspectionnisme à l'aide du principe (faible) de méta-cohérence négative. Les réfutations négatives en revanche, partent de la prémisse plus faible que nous *manquons* d'une justification à croire à la fiabilité de notre expérience E (soit $\sim JKp$), et doivent donc recourir au principe plus fort de méta-cohérence positive pour arriver à la conclusion anti-introspectionniste. Pour un objecteur positif, l'introspectionniste ressemble à Amanda*, qui a toutes les raisons de penser que ses expériences doxastiques (clairvoyantes) sont hasardeuses, et ne devrait donc pas continuer à s'y fier. Pour un objecteur négatif, l'introspectionniste ressemble seulement à Amanda, qui n'a simplement aucune raison (*pro* ou *contra*) concernant la fiabilité de ses expériences doxastiques, mais (d'après l'objecteur) c'est bien suffisant pour constituer une réfutation fondationnelle de sa justification.

Le deuxième type d'objection, l'objection négative, je l'appellerai « objection de l'impression non étayée ». Je dirai qu'une expérience doxastique (ou une impression) est « étayée » lorsqu'elle est accompagnée d'une justification positive à croire que « cette expérience est fiable (ou peut constituer une source de connaissance) ». C'est cette objection de l'impression non étayée que j'aborderai dans la section suivante de ce chapitre.

Pour ce qui est de l'objection positive, l'objection qui part de l'idée que nous avons des raisons positives de tenir notre expérience du libre arbitre comme non fiable ($\sim Kp$), on peut encore en distinguer trois sortes, selon la nature des raisons apportées pour $\sim Kp$. L'objecteur pourrait avancer les raisons suivantes :

- 1) il pourrait apporter des données empiriques indiquant que notre expérience E est produite par une fonction cérébrale qui n'est en rien corrélée avec le contrôle de nos actions : la fonction cérébrale en question n'est pas « orientée à la vérité » (par exemple, qui sert seulement à la satisfaction psychologique), ou bien elle est corrélée à un des faits biologiques qui n'ont rien à voir avec le contrôle de l'action (par exemple, la pression sanguine), ou bien elle n'arrive pas au bon moment pour pouvoir être considérée comme détecteur du contrôle de l'action (par exemple, suivant une interprétation des fameuses expériences de Benjamin Libet (1999), elle arriverait « trop tard », après l'événement cérébral qui contrôle réellement l'action).
- 2) l'objecteur pourrait aussi soutenir que notre expérience E se révèle, à l'analyse conceptuelle, une expérience extrêmement engageante, et notamment une expérience dont le contenu implique logiquement l'indéterminisme causal ; et que, par ailleurs, ce contenu lourd est clairement tel qu'il ne *pourrait pas* être connu sur la base de l'expérience E : quelle que soit la manière exacte dont fonctionne cérébralement l'expérience E, elle ne *pourrait pas* être fiablement corrélée à l'indéterminisme causal.
- 3) enfin, l'objecteur pourrait estimer que, en amont de toute analyse compatibiliste ou incompatibiliste, le contenu de E a déjà incontestablement des conséquences logiques que nous ne pourrions pas *connaître* sur la base de l'expérience E, par exemple la conséquence que nous ne sommes pas manipulés par des martiens : quand bien même nous serions justifiés *prima facie* à supposer que nous ne sommes pas manipulés, un

problème de second ordre apparaîtrait dès que nous prendrions conscience que nous ne pourrions pas *savoir* cela.

Dans ce chapitre, je n'aborderai que le troisième type d'objection positive. Le deuxième type repose sur une analyse conceptuelle particulière du contenu de E, et ne concerne donc pas la question de la justification *indexicale* qui fait l'objet de ma première partie. Il est possible que l'analyse conceptuelle révèle un contenu trop lourd pour être connu, et en ce sens, il est possible que l'analyse conceptuelle *apporte* un réfuteur fondationnel à la croyance introspective au libre arbitre ; j'aborderai cette question à partir de la deuxième partie, mais pour l'instant, je m'intéresse à la question de savoir si elle est (ou était) justifiée *avant* cette entreprise philosophique spécialisée qu'est l'analyse conceptuelle. Le premier type d'objection pourrait quant à lui être formulé indépendamment de toute analyse conceptuelle du contenu de E : en effet, si les neurobiologistes nous apprenaient demain que notre expérience « d'être libre » est un pur épiphénomène d'un processus cérébral qui sert strictement à réguler la pression sanguine, il est clair que notre réaction ne serait pas de dire « donc 'être libre' veut dire 'avoir une bonne pression sanguine' », mais plutôt « donc cette expérience 'd'être libre' n'a absolument aucune connexion avec le fait d'être libre, quel que soit le sens exact de ce contenu ». Donc des données empiriques de ce type (des « preuves de décorrélations ») peuvent constituer un réfuteur efficace, dès le niveau de la justification *indexicale*. Je ne les aborderai pourtant qu'en troisième partie, parce qu'elles reposent sur des informations empiriques spécialisées : il me semble évident qu'Epicure ou Aristote, par exemple, n'avaient aucun réfuteur de ce type.²³¹ Or ce qui m'intéresse dans cette première partie, c'est fondamentalement de retrouver la situation épistémique « de sens commun », celle que l'on a en amont des réfuteurs éventuels apportés par les sciences spécialisées. Il est *possible*, encore une fois, que les sciences spécialisées apportent des réfuteurs fondationnels par preuve de décorrélations, mais cela n'enlèverait pas tout intérêt à l'idée suivante (que je défends dans cette partie) : en amont des réfuteurs éventuels apportés ensuite par les sciences spécialisées, les êtres humains peuvent être absolument justifiés (justifiés *ultima facie*) à se fier à leur expérience du libre arbitre, c'est-à-dire à la fois à accepter son contenu et à la considérer comme une source de connaissance. Je reviendrai à la fin de ce chapitre sur cette thèse de la justification (*ultima facie*) « pré-scientifique ». Pour l'instant, il est clair que seul le troisième type d'objection (l'inconnaissabilité de la conséquence évidente $\sim MM$) pourrait constituer une raison positive de penser que notre expérience du libre arbitre ne confère pas une connaissance. C'est donc cette objection que j'aborderai ici sous le nom « d'objection de l'impression non étayable » : en effet, cette objection revient à dire que l'expérience E n'est pas seulement en manque d'un possible support méta-épistémique, mais tout bonnement impossible à étayer, compte tenu de la « lourdeur » de son contenu.

Examinons donc à présent ces deux objections « pré-scientifiques » : l'objection de l'impression non étayée et l'objection de l'impression non étayable (car trop lourde).

²³¹ Il serait *possible*, en droit, que nous ayons des « preuves de décorrélations » qui ne reposent pas sur des données empiriques spécialisées. Par exemple, on pourrait imaginer le scénario suivant : l'expérience d'agir librement apparaît systématiquement (et uniquement) lorsque nous avons une très grande faim et pour les actes précis qui visent la satiété, et par ailleurs tous les hommes de bon sens se font naturellement cette observation (sur eux-mêmes et sur autrui), et comprennent que cette corrélation avec la faim et la satiété n'a donc rien à voir avec le moindre contrôle de l'action. Dans cette société, c'est même devenu un fait de connaissance commune (que les enfants apprennent sur les genoux des nourrices) que cette expérience est une illusion (comparer : pour nous, c'est une illusion reconnue par le sens commun que le temps « a l'air » de passer plus vite lorsqu'on s'amuse que lorsqu'on s'ennuie). Dans un tel scénario, la justification *indexicale* (procurée par l'expérience du libre arbitre) serait réfutée par une preuve de décorrélations *de sens commun* et non pas spécialisée. Mais en réalité, il n'y a pas de preuve de décorrélations de sens commun : toutes les preuves de décorrélations qu'on peut plausiblement considérer proviennent de sciences avancées (et même relativement tardives à l'échelle de l'histoire de l'humanité).

II. L'objection de l'impression non étayée

II.1) L'objection et son principe de méta-cohérence

L'objection de l'impression non étayée consiste donc à reprocher à l'expérience du libre arbitre de n'être accompagnée d'aucune de raison de penser que cette expérience peut fournir une information fiable concernant son contenu. D'après l'objecteur, l'expérience du libre arbitre ressemble donc exactement à l'expérience de la clairvoyante Amanda : elle n'a peut-être pas (comme Amanda*) de raisons positives de penser que son expérience est illusoire, mais elle manque de toute raison de penser qu'elle est fiable. On pourrait encore comparer cette expérience à une lettre contenant certaines informations mais dont l'auteur serait totalement inconnu et pourrait aussi bien être un auteur de fiction, qu'un journaliste véreux, ou qu'un honnête témoin. Si nous n'avons *aucune* raison de penser que cette lettre provient d'un auteur fiable qui cherche à nous transmettre une information sur la réalité, nous n'irons pas *croire* le contenu de la lettre.

J'ai déjà présenté, dans la section précédente, les deux prémisses fondamentales de ce raisonnement. La première est la prémisse de « non étalement » (NE) : nous ne sommes pas justifiés à tenir une croyance introspective au libre arbitre comme une *connaissance* ($\sim JKLA$). La deuxième prémisse est le principe de méta-cohérence positive : pour être justifiée, une croyance doit être étayée ($Jp \Rightarrow JKp$). La conclusion s'ensuit directement.

Argument anti-introspection de l'impression non-étayée

(NE) $\sim JKLA$

(MC+) $Jp \Rightarrow JKp$

(C) $\sim JLA$

ou en version plus explicite :

Argument anti-introspection de l'impression non-étayée

(NE) Nous n'avons pas de justification à croire que notre expérience du libre arbitre peut fournir une connaissance du libre arbitre.

(MC+) Si *S* a une justification *prima facie* à croire que *p* sur la base d'un fondement épistémique *F*, mais n'a pas de justification à croire que ce fondement peut constituer un fondement de *connaissance*, alors sa justification à croire que *p* est fondationnellement réfutée.

(C) La justification *prima facie* apportée par notre expérience du libre arbitre est réfutée fondationnellement.

Dans cette formulation explicite, il est important de noter que la justification *prima facie* de premier ordre n'est *pas* mise en cause par la version de la méta-cohérence que j'utilise ici. Certaines versions du principe de méta-cohérence pourraient être utilisées pour dire que les expériences non étayées ne fournissent pas le début d'une justification, que *seules* les raisons méta-épistémiques peuvent *conférer* à telle ou telle expérience la moindre force justificationnelle. Ces versions plus fortes poseraient évidemment un problème plus radical contre l'introspectionnisme, mais elles seraient aussi moins plausibles et surtout elles remettraient en cause ce qui a été établi dans les discussions précédentes sur le problème de premier ordre, à savoir qu'il *faut bien* que certaines expériences *en elles-mêmes* confèrent de la justification (au moins *prima facie*) si l'on veut, d'une manière ou d'une autre éviter le scepticisme. Dans l'argument tel que je le formule ici, on voit que le problème de second

ordre ne dépend pas d'un principe si fort, et qu'il se pose *même* si l'on accepte (au premier ordre) des justifications *prima facie* par simple expérience (c'est-à-dire le fondationnalisme). Il s'agit simplement de remarquer que les considérations de second ordre peuvent *réfuter* une justification *prima facie* de premier ordre, en soulevant contre elles un problème de (méta)rationalité. La prémisse de méta-cohérence positive repose sur l'intuition forte qu'il y aurait un problème de rationalité à maintenir que « *p* » si l'on n'est pas en mesure d'accepter (de manière justifiée) « je sais que *p* ».

Une remarque également sur la prémisse de « non étalement » : c'est peut-être une prémisse de ce genre que visaient les objecteurs de la « pure impression » étudiés au chapitre 2, même s'ils la formulaient malencontreusement en termes inférentialistes. Il y a en effet deux manières de reprocher à une expérience d'être une « pure expérience », c'est-à-dire dénuée d'un support méta-épistémique. Une première façon consiste à dire qu'une expérience *en elle-même* ne peut pas apporter de justification, et qu'on ne pourrait donc accepter son contenu que par *inférence* à partir de la croyance méta-épistémique (« cette expérience est fiable ») et d'une croyance méta-expérientielle (« j'ai une expérience que *p* »). Cette première façon conduit directement à la régression inférentialiste. Mais une deuxième façon de formuler le même genre de reproche, c'est de dire qu'une expérience en tant que telle *peut* fournir une justification *non inférentielle*, mais qu'elle ne peut faire cela que dans certains contextes favorables, et les contextes où aucun support méta-épistémique n'est présent seraient des contextes défavorables. Pour répondre au premier reproche, il n'est même pas utile de considérer si l'expérience du libre arbitre est étayée ou non : il suffit de souligner l'erreur inférentialiste et d'accepter les justifications basiques. Mais pour répondre au deuxième reproche, nous allons devoir envisager si, oui ou non, l'expérience du libre arbitre est aussi peu étayée que le soutient l'objecteur.

D'où vient l'intuition que l'expérience du libre arbitre n'est absolument pas étayée ? Je suppose que cette intuition provient implicitement d'une prémisse plus profonde, à savoir l'idée que nous n'avons pas d'indications sur la *manière* dont notre expérience du libre arbitre constitue une connaissance (sur la manière dont notre expérience du libre arbitre dépend du *fait* du libre arbitre). Comparons avec le sens de la vue : nous avons des indications sur la manière dont nos expériences visuelles dépendent (via les réunions lumineuses) des formes et couleurs que nous expérimentons visuellement ; et ces indications nous fournissent sans doute une justification à croire que nos expériences visuelles apportent des connaissances. Dans le cas du libre arbitre, il semble que nous ne pouvons pas dire *que* nos expériences sont fiables, parce que nous ne pouvons pas dire *comment* elles sont fiables. Je reviendrai plus loin sur ce type de motivation de l'intuition (NE), et sur les problèmes qu'elle soulève pour la perception elle-même.

A présent, je ne vais considérer que les problèmes (sceptiques) que soulève l'argument de l'expérience non étayée, quelles que soient les motivations qu'on apporte pour ses prémisses.

II.2) Principe de méta-cohérence et scepticisme de la méta-régression

J'adopterai contre l'argument de l'impression non étayée la même méthode anti-sceptique employée jusqu'ici, c'est-à-dire que je me demanderai si elle permet de faire apparaître contre l'introspectionniste un problème *qui ne ruinerait pas en même temps* la justification des croyances perceptives. Or le principe de méta-cohérence est une source bien connue d'arguments sceptiques contre la justification perceptive. En tirant les leçons de ce

type d'argument sceptique, on pourra répondre de manière satisfaisante à l'argument de l'impression non étayée.

A première vue, il semble pourtant que l'objecteur ait enfin trouvé un argument qui attaque l'introspectionnisme sans attaquer la perception : en effet, comme je l'ai exposé dans la section précédente, il semble qu'on puisse faire un contraste entre le sens de la vue, pour lequel nous avons des indications sur la manière dont nos expériences dépendent des faits observés, et l'expérience du libre arbitre, pour lequel nous *n'avons pas* d'indications de ce genre. Autrement dit, il ne semble pas qu'on puisse mettre en parallèle de l'argument anti-introspection un argument anti-perception vraiment convaincant, tel que :

Argument anti-perception de l'impression non-étayée

(NEP) $\sim JKA$

(MC+) $Jp \Rightarrow JKp$

(C) $\sim JA$

Dans cet argument anti-perception, une prémisse est particulièrement faible, c'est la prémisse de « non étalement perceptif ». Soit en version explicite :

(NEP) Nous n'avons pas de justification à croire que nos expériences perceptives peuvent fournir une connaissance de leur contenu.

Cette prémisse ne semble pas du tout intuitive en elle-même, et il ne semble pas non plus qu'on puisse lui donner une motivation comparable à celle de (NE) : on ne peut pas dire que « nous n'avons aucune indication de la *manière* dont nos expériences perceptives pourraient constituer une connaissance ». Il semble au contraire, de prime abord, que nous *avons* des indications sur cette manière : nous savons que la vue repose sur l'effet causal des rayons lumineux, que l'ouïe repose sur l'effet causal des ondes sonores dans l'air, que l'ensemble des sens reposent sur la transmission d'informations des récepteurs externes jusqu'au système nerveux central en passant par les nerfs, etc. Mais pour le libre arbitre, même si l'on se place en amont de toute analyse compatibiliste ou incompatibiliste (et à vrai dire *surtout* si l'on se place en amont d'une analyse précise de ce que désigne le concept de « libre arbitre »), il est clair que nous n'avons pas de théorie comparable à disposition. Donc l'objecteur semble avoir trouvé ce qu'il lui fallait : une différence pertinente entre expérience perceptive et expérience du libre arbitre, qui permette de dire que les unes nous justifient à accepter leur contenu, tandis que les autres, même si elles peuvent justifier *prima facie*, ne permettent pas (toutes choses considérées) de croire de manière justifiée à leur contenu.

Et pourtant, si l'on réfléchit attentivement à cette manière d'éviter le scepticisme en étayant les expériences perceptives, on s'apercevra qu'elle n'est pas sans poser un certain nombre de problèmes.

Premièrement, cette manière d'étayer les expériences perceptives n'est disponible que pour une toute petite portion de l'humanité : la connaissance des rayons lumineux, des ondes sonores dans l'air, du fonctionnement des neurones, n'étaient pas disponibles, par exemple, pour Aristote ou Epicure, et ne le sont toujours pas pour un grand nombre d'êtres humains aujourd'hui. Si la réponse à l'argument sceptique anti-perception *dépendait* de la possession de telles théories, cela voudrait dire que pour Aristote, Epicure, et une portion très importante de l'humanité, l'argument marche : eux du moins n'ont pas de manière d'étayer leurs expériences perceptives et donc la simple considération des questions méta-épistémiques devraient réfuter leur justification à croire leurs sens. Mais c'est certainement une conclusion

étrange, et déjà radicalement sceptique, de dire que Aristote n'aurait pas dû (épistémiquement) se fier à ses sens.

Deuxièmement, il semble que cette solution pour sauver la perception d'un argument par « non-étalement » ne se généralise pas à d'autres types de croyances qu'un anti-sceptique voudrait sauver, et en particulier aux croyances mathématiques : ainsi que je l'évoquais en introduction, le célèbre « problème de Benacerraf » fait apparaître la difficulté où nous sommes pour offrir une théorie sur la manière dont nos croyances mathématiques (ou non expériences doxastiques mathématiques) pourraient dépendre des vérités qu'elles décrivent. Par conséquent, l'anti-introspectionnisme de l'impression non-étayée ne conduit peut-être pas clairement au scepticisme *de la perception*, mais il semble qu'il conduirait toujours au scepticisme *mathématique*.

Troisièmement, même pour la perception, le type d'étalement proposé présente après réflexion un aspect de circularité qui le rend, au moins, problématique : en effet, comment connaissons-nous l'existence et le fonctionnement des rayons lumineux, des ondes sonores, des récepteurs sensoriels ? N'est-ce pas ultimement sur la base d'observations empiriques au moyen de nos sens eux-mêmes ? Notons que la circularité dont il s'agit ici n'est certainement pas une *circularité logique* : il ne s'agit pas de dire que, en partant de la croyance « notre sens de la vue est fiable », nous aurions *inféré* un ensemble de croyances perceptives (par ex. « donc cette lentille est éclairée de telle et telle façon »), dont nous aurons ensuite inféré une théorie de la vue (« donc il y a des rayons de lumières qui transmettent des informations visuelles »), dont enfin nous aurions enfin inféré que « notre sens de la vue est fiable ». Dans le contexte fondationnaliste où nous travaillons depuis le chapitre précédent, la première étape est mal décrite car les croyances perceptives ne sont pas inférées du tout : *étayer* une croyance perceptive par une méta-justification, ce n'est pas l'*inférer* de cette méta-justification. Et donc le type d'étalement scientifique proposé en faveur de nos sens externes ne sera pas logiquement circulaire... mais il sera néanmoins frappé de *circularité épistémique* dans le sens suivant :

Lorsque l'on dépend d'une source de croyance pour établir la croyance que cette même source de croyance est fiable, alors cette croyance est épistémiquement circulaire. (Bergmann 2006a, 198)²³²

Ce qui est frappé de circularité épistémique, ce ne sont pas les croyances perceptives, mais la croyance méta-épistémique KA (« notre source visuelle est fiable, *i.e.* les croyances qu'elle délivre sont des *connaissances* »), et si elle est frappée de circularité, c'est parce qu'elle est formée à partir de croyances qui utilisent cette même source visuelle. (Bergmann 2004) donne un exemple très simple pour mettre en évidence que la circularité épistémique semble aussi problématique que la circularité logique :

Le Cas des Jurés

Juré n° 1 : Vous connaissez ce témoin qui s'appelle Hank ? J'ai des doutes à propos de sa fiabilité.

Juré n° 2 : Eh bien, peut-être que ceci peut vous aider. Hier, j'ai entendu Hank affirmer qu'il était un témoin fiable.

Juré n° 1 : Ainsi donc, Hank a affirmé qu'il était fiable, c'est bien ça ? Eh bien, voilà qui règle la question. Je suis maintenant convaincu que Hank est fiable.

²³² La distinction entre *circularité logique* et *circularité épistémique* est développée dans (Bergmann 2004, n. 3).

Il est évident que le raisonnement du juré n°1 est inacceptable. Et le problème semble bien résider *précisément* dans le fait qu'il est épistémiquement circulaire. De fait, de nombreux auteurs ont soutenu que toute forme de circularité épistémique, comme telle, était inacceptable, aussi bien celle qui consiste à étayer nos expériences perceptives par des informations empiriques que celle qui consiste à étayer le témoignage de Hank par son propre témoignage. Voici par exemple la condamnation prononcée par Richard Fumerton²³³ :

Vous ne pouvez pas utiliser la perception pour justifier la fiabilité de la perception !
 Vous ne pouvez pas utiliser la mémoire pour justifier la fiabilité de la mémoire ! Vous ne pouvez pas utiliser l'induction pour justifier la fiabilité de l'induction ! De telles tentatives pour répondre aux inquiétudes du sceptique reposent sur une circularité flagrante, et même pitoyable. (Fumerton 1995, 177)

Fumerton et les autres ont-ils raison de condamner toute forme de circularité épistémique ? La question, comme on va le voir, est disputée. Mais ce que révèle cette discussion, c'est que la tâche de satisfaire le principe de Méta-cohérence positive pour la perception est loin d'être une trivialité. Bien plus, la difficulté à satisfaire ce principe *autrement* que par une méta-justification circulaire (et donc problématique) est à l'origine d'un défi *sceptique* tout à fait traditionnel, qu'on a appelé souvent le « problème du critère » (et que j'appellerai plus loin « l'argument de la Méta-Régression ») : quelle justification non circulaire pourrions-nous avoir de croire que nos sens sont fiables ? Par conséquent, un philosophe qui voudrait rejeter l'introspectionnisme du libre arbitre pour des raisons non sceptiques peut difficilement s'appuyer sur la difficulté à satisfaire la méta-cohérence dans le cas de l'expérience du libre arbitre, comme si la méta-justification était une caractéristique évidemment absente pour le libre arbitre et évidemment présente pour toutes les autres croyances qu'un anti-sceptique veut préserver. En réalité, le principe de méta-cohérence représente une menace sceptique pour *toutes* les expériences doxastiques, et si l'on prend sérieusement en compte cette menace, on découvrira qu'il n'y a pas de différence épistémiquement pertinente entre la méta-justification du libre arbitre et la méta-justification de la perception, au sens où les solutions au scepticisme de la méta-cohérence peuvent fondamentalement être transposées au cas de l'expérience du libre arbitre. C'est en tout cas ce que je défendrai dans la section suivante.

Pour finir cette section, je vais présenter de manière plus rigoureuse les raisons pour lesquelles le principe de méta-cohérence (positive) constitue une source traditionnelle d'argument sceptique. Dans cette présentation, je m'efforcerai de mettre en évidence l'analogie très forte entre ce nouveau type d'argument sceptique et l'argument de la régression que j'ai présenté au chapitre 2 (p. 31). L'argument de la régression reposait sur l'idée que si l'on veut justifier inférentiellement toutes nos croyances, tout en évitant la circularité logique, on est conduit à une régression infinie des croyances justifiantes ; le nouvel argument fondé sur la méta-cohérence reposera sur l'idée que si l'on veut étayer (ou méta-justifier) toute croyance, tout en évitant la circularité épistémique, on est conduit à une régression infinie des croyances (ou des sources) « étayantes » (ou méta-justifiantes).

Il n'est pas besoin de présenter les choses de manière très technique pour faire apparaître le problème de la méta-régression. Repartons de la situation de dialogue très simple envisagée dans l'argument de la régression, situation dans laquelle un interlocuteur mal intentionné décide de reposer indéfiniment la question pourquoi :

- Pierre est à la piscine (p1).
- Pourquoi p1 ?

²³³ D'autres condamnations importantes de la circularité épistémique sont celles de (Vogel 2000), (Wright 2004), et (Huemer 2011, 10–15). J'entrerais dans le détail de leur discussion en abordant les *solutions* à l'argument.

- Parce qu'il est tous les jeudis à la piscine (*p2*) et qu'on est jeudi (*q2*).
- Pourquoi *p2* ?
- Parce que c'est sur son emploi du temps (*p3*) et qu'il n'a pas de raisons de ne pas le suivre (*q3*).
- Pourquoi *p3* ?
- Parce que je l'ai vu sur son emploi du temps (*p4*).

Dans ce dialogue, il faut remarquer une différence fondamentale entre les réponses *p2* et *p3* d'une part, et la réponse *p4* d'autre part. Au début du dialogue, l'interrogé répond en avançant des propositions qu'il croit vraies et qui fondent inférentiellement la proposition à justifier. On peut dire qu'en réponse à la demande de justification, il *exhibe* la proposition qui constitue sa raison ou sa justification.²³⁴ Dans la dernière réponse en revanche, l'interrogé n'*exhibe* pas une raison ; il ne *peut pas* l'exhiber car il s'agit cette fois d'une perception et non d'une proposition ou d'une croyance, et donc sa réponse consiste à *mentionner* sa raison, ou plus précisément sa source de justification (la vue). En faisant cela, il rend manifeste que sa croyance *p3* n'est *pas* inférentielle et donc qu'il est fondé à *ne pas* produire une proposition ou une croyance qui la justifie. On peut dire que cette dernière réponse le fait « décrocher » du jeu de la régression dans lequel l'interlocuteur essaye de l'emmener. Et on voit bien que l'interlocuteur manquerait d'à propos en posant pour *p4* la simple question « pourquoi ? » : en un sens, l'interrogé a *déjà* cessé de jouer ce jeu. Pourtant, l'interrogateur sceptique a une autre ressource pour relancer le jeu. Il peut encore poser une question pertinente à propos de *p4*. Non pas certes « qu'est-ce qui te permet de croire que tu as vu son emploi du temps », mais la question plus subtile : « qu'est-ce qui te permet de croire que cette source de justification que tu viens de mentionner (la vue) est une source fiable ? » Cette nouvelle question est proprement la question méta-épistémique. On peut imaginer (au moins de manière théorique) que l'interrogé trouve des réponses et on pourra donc continuer ainsi le dialogue :

- Qu'est-ce qui te permet de croire que *S1* est fiable ?
- C'est que je le sais par *S2*.
- Qu'est-ce qui te permet de croire que *S2* est fiable ?
- C'est que je le sais par *S3*.

Si l'on pense à des sources de connaissance par *témoignage*, ce genre de suite au dialogue est une possibilité tout à fait concevable :

- Qu'est-ce qui te permet de croire que Hank est fiable ?
- C'est que Jean m'a assuré de sa fiabilité.
- Mais qu'est ce qui te permet de croire que Jean est fiable.
- C'est que Robert m'a assuré de sa fiabilité.

Et il est évident que ce type de dialogue conduit à une régression des sources étayantes, que j'appelle « Méta-régression » pour la distinguer de la régression des raisons (ou régression de premier ordre).²³⁵ Pour éviter la régression, il ne semble y avoir que deux alternatives : soit s'arrêter sur une croyance dont la source de justification ne sera pas elle-même étayée (refus

²³⁴ Il est important de noter que le jeu consiste bien, pour l'interrogé, à produire ses raisons ou ses justifications en réponse à la question « pourquoi ». En effet, ce jeu a pour but de mettre en évidence la structure justificationnelle de la croyance en question *pour ce sujet*. Si, par exemple, l'interrogé se mettait à donner à son interlocuteur des arguments *ad hominem* dont lui-même n'accepte pas les prémisses, la structure de raisons ainsi produite ne représenterait plus la structure justificationnelle du sujet.

²³⁵ Le terme même de méta-régression a été utilisé par différents auteurs pour caractériser l'argument du critère comme je le fais : la présentation du problème qui se rapproche le plus de la mienne est celle de (McGrew and McGrew 2007, chap. 3). Voir aussi (Sanger 1999). La plupart des auteurs, cependant, adoptent une terminologie différente, soit en insistant sur la tentative d'éviter la régression à l'infini par des chaînes de méta-justification circulaire (on parle alors de « problème de la circularité épistémique »), soit en insistant sur la structure régressive, mais sans chercher à distinguer la régression de premier ordre de la régression de second ordre (on parle alors de « problème de la régression »).

de la méta-cohérence positive), soit réutiliser à un moment ou à un autre des réponses une source étayante déjà mentionnée auparavant. Pour formuler le problème de Méta-régression au plus proche du problème de la régression de premier ordre, on peut dire la chose suivante : accepter le principe de méta-cohérence revient à dire que toute croyance justifiée doit apparaître au sein d'une *chaîne de méta-justification* (ou d'une chaîne de sources étayantes), et qu'il n'y a que trois sortes de chaînes possibles : chaînes infinies, chaînes terminées, et chaînes circulaires. Le problème sceptique consiste alors à faire observer, comme pour la régression de premier ordre, qu'aucun de ces trois types de chaînes ne semble pouvoir satisfaire réellement le principe de méta-cohérence. L'argument peut se résumer ainsi :

Argument sceptique de la Méta-Régression épistémique

- (MC+) Une croyance que *p* ne peut être *ultima facie* justifiée (donc non réfutée fondationnellement) qu'en présence d'une méta-justification à croire *Kp* (et donc au sein d'une chaîne de méta-justification).
- (2) Il n'existe que trois sortes de chaînes de méta-justification: chaîne terminée, chaîne circulaire, et chaîne infinie.
- (3) Une chaîne terminée ne peut pas étayer une croyance car sa terminaison n'est pas elle-même étayée (d'après MC+).
- (4) Une chaîne épistémiquement circulaire ne peut pas étayer une croyance car la circularité épistémique est viciieuse.
- (5) Une chaîne infinie ne peut pas justifier nos croyances humaines car nous n'avons pas les moyens d'avoir effectivement de telles chaînes.²³⁶
- (C) Aucune croyance (au moins humaine) ne peut être étayée, et donc toute croyance humaine est victime d'une réfutation fondationnelle de justification.

On peut facilement constater la très forte analogie structurelle avec l'argument de la régression de premier ordre présenté p. 31. Mais une fois qu'on a constaté cette analogie structurelle, il faut garder à l'esprit qu'il s'agit bel et bien de deux arguments distincts. L'argument de la méta-régression peut se poser *même* si l'on refuse le principe inférentialiste qui génère la régression de premier ordre (et donc si l'on accepte une théorie *fondationnaliste* de la justification) : il repose quant à lui sur un autre principe (MC+) dont la formulation la plus claire accepte même explicitement le fondationnalisme. Corrélativement, un auteur peut très bien accepter une certaine structure de solution pour répondre à un argument et une *autre* structure de solution pour répondre à l'autre. Il *pourrait* y avoir une certaine affinité entre solutions de même structure, mais ce n'est pas une nécessité. En particulier, un auteur pourrait adopter une position globale qui accepte des chaînes de raisons terminées, et des chaînes de méta-justification circulaires.²³⁷ De fait, dans la section suivante, je montrerai pourquoi c'est certainement une version de cette solution qui est la plus plausible.

Mais quelle que soit la solution adoptée pour répondre à ce nouvel argument sceptique, mon argument principal suit toujours la démarche anti-sceptique : dans la mesure où le problème du défaut de méta-justification touche aussi bien les expériences perceptives que l'expérience du libre arbitre, on a toutes les raisons de penser qu'une bonne solution anti-sceptique sauvera du même coup l'expérience du libre arbitre. Nous allons voir à présent en détail comment les diverses solutions pourraient s'appliquer au libre arbitre, et laquelle est la plus plausible et la plus utile pour continuer la discussion.

²³⁶ Ou parce qu'une régression infinie, même possible, serait viciieuse.

²³⁷ C'est le cas notamment de (Bergmann 2006b).

II.3) Solution : Méta-cohérence disjonctive et circularités épistémiques vertueuses

Les trois solutions principales qu'on peut proposer pour éviter la méta-régression consistent à : (i) refuser le principe (MC+) qui génère la régression (solution méta-incohérentiste), (ii) refuser le caractère problématique d'une régression circulaire (solution circulariste), ou (iii) refuser le caractère problématique d'une régression infinie (solution « infinitiste »). Une quatrième solution consiste à dire que l'argument utilise une notion contextuellement équivoque de « justification » (solution « contextualiste »).²³⁸ Les solutions contextualiste et infinitiste ne présentent aucune particularité significative dans cette discussion de deuxième ordre par rapport aux discussions de premier ordre : les raisons que nous avons de les rejeter au chapitre 2 vaudront donc aussi bien pour ce nouvel argument.²³⁹ Je n'aborderai donc ici que les solutions méta-incohérentiste et circulariste. C'est-à-dire que je m'intéresserai d'abord aux deux questions suivantes : (i) Est-il vraiment plausible de rejeter le principe de méta-cohérence ? (ii) Peut-on accepter certaines formes de méta-circularité ? A la première question, je répondrai qu'un certain principe de méta-cohérence (positive) *doit* être satisfait. Et donc s'il faut satisfaire le principe de méta-cohérence, et si l'on ne peut le satisfaire que par des chaînes de méta-justification circulaires (car les chaînes terminées et infinies sont rejetées), alors pour répondre au sceptique il faudra voir quel *type* de chaînes circulaires sont acceptables, car les anti-circularistes (sceptiques ou non) ont certainement raison de souligner que *certaines* formes de circularité épistémique sont vicieuses.

Une solution qui refuserait absolument le principe de méta-cohérence positive n'est pas plausible, car elle reviendrait à accepter que l'on puisse (de manière justifiée) à la fois affirmer certaines croyances – par exemple la croyance perceptive « il y a une table en face de moi » – et en même temps concevoir que cette croyance pourrait très bien être purement hasardeuse. Notez que si l'on acceptait cette solution, elle pourrait sans problème être transposée à la défense de l'introspectionnisme du libre arbitre : l'introspectionnisme pourrait accepter de dire qu'il n'a aucune raison particulière de se fier à son expérience du libre arbitre, et continuer d'affirmer que cette expérience le justifie pourtant à croire qu'il est libre. Mais il est clair que cette réponse est insatisfaisante, et si elle est insatisfaisante, c'est qu'au moins une certaine version du principe de méta-cohérence (positive) doit être vraie. Pour déterminer la bonne solution au problème de la méta-régression, il faut donc bien comprendre quelle version du principe de méta-cohérence est quasiment impossible à refuser.

²³⁸ Si l'on compare avec la liste des solutions proposée pour l'argument de la régression (p. 110), il semble manquer une solution : la solution *externaliste*. J'aborderai la question de l'externalisme un peu plus loin en envisageant les différents types de *circularisme*. En effet, si du point de vue de la régression de premier ordre, l'externalisme consiste à refuser totalement le principe de la régression (le principe inférentialiste), du point de vue de la méta-régression en revanche, l'externalisme consiste à accepter le principe de régression (MC+) et à le satisfaire par une certaine forme de régression circulaire, ainsi que le montre la position de Michael Bergmann.

²³⁹ Je renvoie donc aux p. 117 *sqq.* pour la réponse à l'infinitisme, et p. 156 *sqq.* pour la réponse au contextualisme. Il faut noter que la solution infinitiste de deuxième ordre n'est pas *exactement* semblable à la solution de premier ordre : en effet, elle consiste à accepter une régression infinie des *sources* de connaissances, au lieu d'une régression infinie des croyances elles-mêmes. Mais cette différence rend l'infinitisme de deuxième ordre encore plus irréaliste que l'infinitisme de premier ordre. Quant au contextualisme, ma réponse consiste à dire que je reste invariablement intéressé à la notion « d'impeccabilité épistémique interne », qu'il s'agisse de la justification JKA ou de la justification JA. On notera que la solution contextualiste avait été abordée en réponse à l'argument de *clôture* (et non pas à l'argument de la régression), et concernait le rapport entre justification JA et J~HS ; c'est que l'argument de méta-régression présente des analogies structurelles à la fois avec l'argument de régression et avec l'argument de clôture, ainsi qu'il est déjà apparu par la comparaison du principe de clôture et du principe de méta-cohérence.

La meilleure manière de mettre en évidence la force intuitive du principe de méta-cohérence, et d'en donner la formulation la plus convaincante, c'est de repartir de l'analogie que j'ai soulignée au début de ce chapitre (p.186) avec le principe de clôture. L'un et l'autre principes, ai-je dit, peuvent être conçus comme des principes de « transfert » d'état épistémique entre deux positions de l'espace doxastique d'un agent S : transfert entre $\langle p \rangle^S$ et $\langle q \rangle^S$ pour la clôture,²⁴⁰ transfert entre $\langle p \rangle^S$ et $\langle Kp \rangle^S$ pour la méta-cohérence. Mais on peut envisager des liens plus ou moins forts entre deux positions épistémiques. Un lien beaucoup trop fort, dans le cas de la clôture, serait d'exiger la croyance consciente de toutes les conséquences q pour pouvoir attribuer le statut de « croyance justifiée » à la croyance que p . Mais inversement, il serait inacceptable de refuser absolument toute contrainte, tout lien de dépendance entre le statut épistémique de la position $\langle p \rangle^S$ et celui de $\langle q \rangle^S$. Ce qu'il faut trouver, c'est le bon type de connexion, et pour cela, il faut considérer ce qui rend la connexion vraiment indispensable.

Pour la méta-cohérence comme pour la clôture, le cœur de l'intuition réside dans le fait qu'un agent épistémique S serait manifestement irrationnel s'il adoptait une certaine attitude (ou conjonction d'attitudes). Pour la clôture, le scénario imaginé était celui d'un agent qui accepterait une certaine croyance que p , qui reconnaîtrait explicitement pouvoir déduire q de p , et pourtant continuerait après réflexion à considérer qu'il n'a aucune bonne raison de croire que q . C'est ce que j'ai appelé le « problème de l'inconséquent ».²⁴¹ Pour la méta-cohérence, le scénario clairement problématique pourrait être formulé de la manière suivante :

Problème de l'incohérent

Supposez que je demande à S si elle est d'accord avec p . Sa réponse est affirmative : « Oui », dit-elle. Puis je demande à S si elle est d'accord avec Kp . « Je ne suis pas d'accord avec ça », dit-elle. Je lui demande si elle souhaite à présent revenir sur sa déclaration que p . « Oh non », dit-elle. « Je maintiens ma déclaration que p . Simplement, je me refuse à soutenir que c'est quelque chose que je sais. Ce que je dis, c'est que p est vrai, mais que je ne le sais pas, ou qu'il est bien possible que je ne le sache pas ; et il n'y a là aucune contradiction logique. »

S a raison de faire remarquer qu'il n'y a aucune contradiction logique dans sa conjonction d'affirmations. Mais cette conjonction présente manifestement un autre type d'inacceptabilité, connu aujourd'hui sous le nom de « paradoxe de Moore ». Le paradoxe de Moore le plus célèbre concerne les assertions de la forme suivante :

(PM1) p mais je crois pas que $\sim p$. ($p \wedge B\sim p$)

Pour reprendre l'exemple de Moore lui-même :

Je crois qu'il pleut et il ne pleut pas. (Moore 1993a, 208)

Ce qui caractérise les conjonctions moore-paradoxaes, c'est que leurs conjoints *pourraient* être vrais en même temps, et que pourtant leur assertion conjointe (ou leur *croyance conjointe*) sont pourtant manifestement inacceptables, à la limite de la contradiction. Cette catégorie de propositions n'est pas restreinte aux propositions de la forme (PM1). Moore lui-même a remarqué qu'elle pouvait s'étendre à d'autres formes, telles que :

²⁴⁰ Où q est une conséquence logique de p (connue, ou justifiée, ou effectivement déductible).

²⁴¹ Je renvoie à la citation de Hawthorne (plus haut p.156) pour les détails du scénario de l'inconséquent.

(PM2) p mais je ne crois pas que p . ($p \wedge \sim Bp$)

Par exemple :

Je suis allé au cinéma mardi dernier, mais je ne crois pas que j'y sois allé.
(Moore 1942, 543)

Enfin, le caractère inacceptable demeure si l'on passe de la notion de croyance à d'autres notions épistémiques telles que la connaissance :

(PM3) p mais je ne sais pas que p . ($p \wedge \sim Kp$)²⁴²

Moore donne l'exemple suivant :

Les chiens aboient, mais je ne sais pas qu'ils aboient. (Moore 1962, 277)

Enfin, comme on le voit dans le scénario du « problème de l'incohérent », le caractère paradoxal est préservé si l'agent épistémique n'a pas la croyance de la forme $\sim Kp$ mais *doute* (explicitement) que Kp soit le cas. Autrement dit, la forme suivante reste moore-paradoxale :

(PM4) p mais il est tout à fait possible que je ne sache pas que p . ($p \wedge ?\sim Kp$)²⁴³

Ce sont ces deux dernières formes, (PM3) et (PM4) qui sont à l'origine du « problème de l'incohérent », et qui amènent à reconnaître une certaine connexion épistémique minimale entre la croyance que p et la croyance que Kp , c'est-à-dire un certain principe de méta-cohérence, qui exclut qu'on puisse parfois se retrouver *de manière justifiée* dans la situation de l'incohérent. Le refus de (PM3) correspond à la version *négative* du principe de méta-cohérence (un déni explicite de connaissance que p n'est pas cohérent avec le fait de continuer à affirmer que p) ; tandis que le refus de (PM4) correspond à la version *positive* du principe de méta-cohérence (un simple *doute* sur la connaissance que p , s'il est explicite, n'est pas cohérent non plus avec le fait de continuer à affirmer que p). Dans la mesure où les conjonctions (PM4) sont clairement aussi inacceptables que les conjonctions de type (PM3), on peut déjà pressentir ici (j'y reviens un peu plus loin) qu'on ne pourra pas plausiblement se restreindre à un principe simplement négatif.

L'intuition de méta-cohérence n'est donc, fondamentalement, rien d'autre qu'une explicitation de l'inacceptabilité des croyances moore-paradoxales qui utilisent la notion de connaissance.²⁴⁴ Voici, plus explicitement, comment on passe de l'une à l'autre. Une

²⁴² Le passage à cette troisième forme suggère évidemment d'envisager les propositions de la forme ($p \wedge K\sim p$) qui est à (PM1) ce que (PM3) est à (PM2). Seulement, cette forme de propositions est *logiquement contradictoire*, et non plus « moore-paradoxale », puisque la connaissance est factive ($K\sim p \Rightarrow \sim p$).

²⁴³ N'est-il pas possible, pourtant, de croire certaines choses sans prétendre les savoir ? Autrement dit, n'est-il pas possible de dire quelque chose de la forme suivante :

(PM5) Je crois que p mais je ne prétends pas le savoir.

Y aurait-il là vraiment un paradoxe ? N'est-ce pas au contraire le genre de chose que l'on dit ? Il me semble clair que lorsqu'on parle ainsi, on utilise le mot « savoir » dans un sens fort : « je n'en suis pas certain, je n'en ai pas de preuve formelle, etc. » Mais dans le sens que nous utilisons ici, le sens minimal externaliste, une telle affirmation serait manifestement inacceptable :

(PM5*) Je crois que p mais ne prétends pas qu'il y ait le *moindre* rapport entre ma croyance et le fait que p .

C'est cette inacceptabilité qui nous intéresse ici, et qui génère le principe de méta-cohérence.

²⁴⁴ Il faut remarquer qu'il existe en fait *deux* grandes traditions d'interprétation du paradoxe de Moore, l'une qui mène au principe de méta-cohérence, l'autre qui n'y mène pas. La première, que je vais développer, est défendue

croyance de la forme $(p \wedge \sim Kp)$ serait irrationnelle parce que, en présence de la croyance justifiée $\sim Kp$ (le deuxième conjoint), la croyance que p serait forcément *réfutée*. Par conséquent, il est impossible que les deux conjoints soient justifiés en même temps : soit on n'a pas de justification pour $\sim Kp$ et dans ce cas la croyance que p peut être justifiée, mais pas la conjonction ; soit on a une justification pour $\sim Kp$, mais alors la croyance que p (quelle que soit sa justification *prima facie*) se trouve réfutée fondationnellement. De même une conjonction d'attitudes épistémiques du type $(p \wedge ?\sim Kp)$ – consistant à croire que p mais à soupçonner positivement que $\sim Kp$ – rendra forcément irrationnel l'un de ses deux conjoints : soit le soupçon ($?\sim Kp$) sera injustifié, soit il sera justifié mais alors il constituera un réfuteur fondationnel contre la croyance que p (quelle que soit la justification *prima facie* pour p). De cette interprétation, on peut tirer immédiatement le principe de réfutation fondationnelle suivant :

(RF) Si S a (à t) une justification *prima facie* à croire que p , mais que S envisage (à t) la question de savoir si Kp est vraie, et si S face à cette question reste de manière justifiée dans une attitude de doute ($?\sim Kp$) ou s'il adopte de manière justifiée la croyance $\sim Kp$, alors la justification *prima facie* à croire que p est *ultima facie* réfutée fondationnellement.

Or ce principe de réfutation fondationnelle, si l'on le lit bien, n'est rien d'autre qu'un principe de méta-cohérence. En effet, il n'y a que trois manières de « réagir » à la question de connaissance : soit croire (Kp), soit croire ($\sim Kp$), soit rester dans le doute ($?\sim Kp$). Comme deux de ces attitudes constituent des réfuteurs, le principe (RF) revient donc à dire que (face à la question de connaissance Kp) une croyance que p ne peut être (*ultima facie*) justifiée que si S est en mesure d'adopter Kp de manière justifiée.

Une fois qu'on a conscience de cette origine du principe de méta-cohérence, on comprend à la fois pourquoi il est quasi impossible de le refuser absolument, et pourquoi une formulation trop simple, du type

(MC+) $Jp \Rightarrow JKp$

notamment dans (Shoemaker 1995), (Almeida 2001), et surtout (Huemer 2011, sec. 3). La seconde est la plus ancienne puisqu'elle est celle de Moore lui-même, reprise depuis par (Williamson 2000, sec. 11.3). L'interprétation de Moore-Williamson consiste à dire qu'une *assertion* de la forme (PM3) est inacceptable « because by asserting p positively you *imply*, though you don't assert, that you know that p » (Moore 1962, 277). Une contradiction apparaît donc, pour des raisons conversationnelles, entre le contenu asserté ($\sim Kp$) et une certaine implicature conversationnelle (Kp comme implicature de l'assertion ' p '). Pourquoi une assertion que ' p ' entraînerait-elle l'implicature que Kp ? Si l'on suit la théorie de (Grice 1975), les implicatures apparaissent en fonction des *normes pragmatiques* qui régissent les assertions. L'idée, en l'occurrence, serait donc que « la connaissance est la norme de l'assertion » (je ne dois asserter que p que si je sais que p , et donc si j'asserte que p , mes interlocuteurs sont en droit de supposer que j'ai respecté la norme, c'est-à-dire que je sais que p). Cette interprétation de Moore-Williamson, qui fait dériver le problème de (PM3) d'une *norme de l'assertion*, a un inconvénient majeur : elle explique seulement le caractère paradoxal d'une *assertion* de la forme (PM3)... et en tant que telle elle ne dit rien contre un agent qui aurait des *croyances* de la forme (PM3) ! Or comme le dit (Almeida 2001, 33) : « an adequate explanation of the nature of Moorean absurdity is, first and foremost, an explanation of the oddity of (the objects of) certain *beliefs* ». Il est en effet évident que des croyances (PM3) seraient absolument irrationnelles, et si l'on rend compte de *cela*, on pourra aussi expliquer (dériver) pourquoi les *assertions* sont inacceptables, puisqu'il est clairement inacceptable d'asserter une croyance irrationnelle – ou selon la formulation de (Shoemaker 1995, 76) : « what can be (coherently) believed constrains what can be (coherently) asserted ». Mais une explication qui s'en tient à l'inacceptabilité des assertions, en autorisant des croyances rationnelles de la forme « p et je ne sais pas que p » est clairement insuffisante. L'irrationalité d'une telle croyance est, à mon sens, une donnée brute ; et la *seule* explication de *ce* fait précis (par opposition aux faits concernant l'assertion), c'est le principe de méta-cohérence.

est trop forte ou au moins ambiguë. Ici encore, il suffit de suivre les mêmes étapes que pour le principe de clôture : nous avons vu que la formulation de base

$$(J\text{-Clot.}) [Jp \wedge p \Rightarrow q] \Rightarrow Jq$$

était très forte, compte tenu de l'origine de l'intuition de clôture, à savoir le rejet des attitudes inconséquentes. A la suite de (Hawthorne 2005), j'ai fait observer que (J-Clot.) était trop fort pour trois raisons. Premièrement, il exige une *croyance* justifiée pour toutes les conséquences q d'une croyance p ; mais comme ces conséquences sont infinies, il supposerait que nous soyons capables d'avoir une infinité de croyances. Deuxièmement, il exige une justification pour *toutes* les conséquences logiques q de p , y compris celles dont le sujet n'aurait aucunement conscience, et l'on voit mal pourquoi ce serait le cas. Ces deux premiers problèmes amènent à affaiblir la clôture (J-Clot.) par une contrainte de « conscience » : c'est uniquement pour les conséquences *dont il prend conscience* qu'un sujet S doit être en mesure *d'acquérir à ce moment-là* une croyance justifiée. Troisième problème : même avec cette contrainte de « prise de conscience », on pourrait croire que le principe rend toujours possible d'acquérir la croyance justifiée que q lorsqu'on était (auparavant) justifié à croire que p . Mais ce n'est évidemment pas le cas : si j'ai deux croyances justifiées (non réfutées) à t , dont je prends conscience à $t+1$ qu'elles impliquent logiquement une contradiction, cela ne me met pas en situation de croire une contradiction de manière justifiée. En l'occurrence, l'impossibilité de croire la conséquence de manière justifiée constitue (après réflexion) un réfutateur de la justification des prémisses. C'est ce dernier problème qui amenait à une formulation disjonctive du principe de clôture :

- (Clôture Disj.) Si S croit que p de manière *prima facie* justifiée à t et si S reconnaît à t qu'il peut déduire q de p , alors
- (a) soit S est en mesure d'avoir à t la croyance justifiée que q
 - (b) soit la justification *prima facie* pour p se trouve par là même réfutée (pour S à t).

On peut appliquer exactement les mêmes étapes de raisonnement pour le principe de méta-cohérence. Le principe fort (MC+) a trois défauts. Premièrement, il exige pour toute croyance que p une *croyance* justifiée que $Kp...$ et donc une croyance justifiée que KKp , etc., c'est-à-dire qu'il exige d'avoir une infinité de croyances actuelles. Deuxièmement, il exige une justification Kp y compris pour des agents épistémiques qui n'auraient aucunement envisagé la question méta-épistémique, et la possibilité d'un réfutateur fondationnel. Or on voit que pour éviter les scénarios du type « problème de l'incohérent », il n'est nullement besoin de poser des exigences aussi contraignantes : c'est uniquement *lorsqu'il prend conscience* de la question méta-épistémique que l'agent épistémique S doit être en mesure *d'acquérir à ce moment-là* une croyance justifiée Kp . Troisièmement, cette nouvelle formulation pourrait encore donner l'impression que le principe rend toujours possible d'acquérir (après réflexion) la croyance justifiée Kp lorsqu'on était (auparavant) justifié à croire que p . Mais ce n'est évidemment pas le cas : dans certains cas, le fait de considérer la question méta-épistémique apportera un *réfutateur* (fondationnel) de ma justification *prima facie*. Ce dernier problème amène à une formulation disjonctive du principe de méta-cohérence :

- (MC Disj.) Si la croyance de S que p (à t) est *prima facie* justifiée par un fondement épistémique F , et si S soulève à t la question de la vérité de $K_F p$, alors
- (a) soit S est en mesure d'avoir à t la croyance justifiée que $K_F p$
 - (b) soit la justification *prima facie* pour p se trouve par là même réfutée (pour S à t).

Michael Huemer donne une formulation équivalente, qui a l'avantage d'être plus simple (et l'inconvénient d'être moins explicite) :

(MR) Le fait de croire catégoriquement que *p* nous engage, après réflexion, à accepter que nous savons que *p*. (Huemer 2011, sec. 1)

La condition de « conscience » est clairement posée (« après réflexion »), et la condition disjonctive est exprimée par l'idée « d'engagement [rationnel] » : de même qu'un agent épistémique *S* est « engagé » à accepter les conséquences de ce qu'il croit (lorsqu'il *reconnaît* ces conséquences, et *à moins d'abandonner* ses croyances antérieures), de même, il est « engagé » à accepter que ses croyances constituent des connaissances (lorsqu'il *envisage* la question, et *à moins d'abandonner* ses croyances antérieures).

Dans sa version (MC Disj.), le principe de méta-cohérence est donc extrêmement difficile à refuser, et la plupart des auteurs qui ont essayé d'éviter la méta-régression en refusant la méta-cohérence se sont en fait attachés à critiquer des formulations plus fortes (et effectivement critiquables).²⁴⁵ Il ne faut donc pas trop espérer pouvoir répondre à l'argument de la méta-régression par un refus catégorique de tout principe de méta-cohérence. Mais on pourrait aussi espérer que la version (MC Disj.), puisqu'elle est plus faible que (MC+), fournisse ainsi en elle-même le principe d'une réponse à l'argument sceptique : il s'agirait alors de dire qu'on *accepte* le cœur d'intuition de la méta-cohérence, soit (MC Disj.), mais que l'argument de méta-régression *a besoin* d'un principe plus fort, comme (MC+) pour fonctionner. C'est-à-dire que le principe (MC Disj.) n'entraînerait pas de méta-régression sceptique. Malheureusement, ce n'est pas le cas. Bien sûr (MC Disj.) ne pose pas de problème de justification pour un agent épistémique qui accepte ses impressions perceptives au premier degré, *sans se poser la question* de leur valeur cognitive. Mais *dès que la question est posée*, c'est-à-dire pour tout philosophe (ou tout étudiant) qui réfléchit à la question du scepticisme, la méta-régression réapparaît.

Que faudrait-il pour éviter la régression ? Pour éviter la régression, il faudrait affaiblir (MC+) non pas dans le sens de (MC Disj.), mais dans le sens d'un principe de méta-cohérence *strictement négatif*. C'est-à-dire qu'il faudrait accepter que le philosophe, sous la pression des méta-questions de son interlocuteur sceptique, puisse être justifié à dire (à une certaine étape *n* de la chaîne des méta-justifications) :

« Ce que je viens de dire (à l'étape *n-1*) pourrait très bien n'être pas une connaissance (pourrait très bien être une croyance purement hasardeuse), mais ça ne m'empêche pas de continuer à le croire. »

²⁴⁵ La seule attaque que mentionne (Huemer 2011, n. 3) contre (MC Disj.) proprement dite, est celle de (Klein 2008, 117–120). Mais quand on y regarde de près, Klein n'attaque pas réellement le principe de méta-cohérence. Voici la conclusion défendue par Klein : « There is a "legitimate response" to Academic Skepticism that does not require us to have an argument powerful enough to fully justify us in believing that we have *cognitio-style* knowledge. » Par *cognitio*, Klein entend ici la notion de dépendance externaliste que Sosa appelle « connaissance animale » et que j'ai appelée « connaissance » tout court. Ce que dit Klein, c'est donc qu'il n'est pas nécessaire d'avoir une justification de type *JKp*... pour répondre au sceptique académique. Pourquoi ? Parce qu'un sceptique académique est quelqu'un qui *affirme* $\sim Kp$ et que pour lui répondre il suffit de lui montrer que son argument pour $\sim Kp$ n'est pas convaincant. On l'amène ainsi à *ne plus croire* $\sim Kp$, sans lui avoir apporté une justification pour *Kp*. Mais si l'on s'en tient là, on obtient une autre forme de scepticisme, à savoir le scepticisme *pyrrhonien* (qui reste dans une position d'enquête et de doute entre *Kp* et $\sim Kp$). Klein est très clair à ce sujet. Et pour éviter ce type de scepticisme, il faudrait positivement apporter une justification *JKp*. Evidemment, Klein ne se soucie pas de souligner cette condition nécessaire pour éviter le pyrrhonisme, pour la simple et bonne raison qu'il *accepte* le pyrrhonisme. Mais il n'empêche que si l'on veut l'éviter, il faudra bel et bien une méta-justification *JKp*.

Un principe de Méta-cohérence strictement négatif permet une telle affirmation, du moment que le philosophe en question n'a pas *positivement* la croyance $\sim Kp$. Mais cette affirmation est de la forme (PM4) et elle est clairement aussi inacceptable, ou moore-paradoxe, qu'une affirmation de la forme (PM3).²⁴⁶ Pour rendre compte de cette inacceptabilité, il faut bien un principe tel que (MC Disj.), c'est-à-dire un principe de méta-cohérence qui reste fondamentalement *positif*, bien qu'il soit plus faible que (MC+). Or comme on vient de le voir, un tel principe, du fait qu'il exige (après réflexion) une méta-justification *positive*, suffit à entraîner la méta-régression, au moins pour le philosophe qui se pose la question de méta-justification. Notez que, si l'on acceptait des affirmations de la forme (PM4) pour prémunir la perception de la méta-régression sceptique, *alors* on aurait aussi par la même occasion une défense possible de l'introspectionnisme, car l'introspectionniste aurait le droit de se déclarer libre sur la base de l'introspection tout en acceptant (NEP) $\sim JKA$: « je crois que je suis libre du fait de mon expérience d'être libre, mais je n'ai aucune raison particulière de penser que cette expérience fournit une connaissance (*i.e.* elle pourrait fort bien être purement hasardeuse) ». Mais il est évident qu'une telle réponse serait totalement inacceptable. Et encore une fois, rejeter une telle réponse, c'est accepter (MC Disj.) qui génère une régression des méta-justifications.

Il n'est donc pas possible, pour le philosophe au moins, d'éviter la méta-régression. Et donc si l'on acceptait les autres prémisses de l'argument de la méta-régression, cela signifierait que *tout philosophe* qui envisage la question méta-épistémique a un réfutateur de justification pour toutes ses croyances perceptives. Le scepticisme ne gagnerait pas contre le *vulgum pecus* qui ne se pose pas de questions, mais il gagnerait aussitôt qu'il est envisagé. Il est donc évident que tout philosophe anti-sceptique devra refuser cette conclusion. Et comme le principe (MC Disj.) est à peu près irréfutable, un philosophe anti-sceptique devra donc refuser une autre prémisse de l'argument de méta-régression : il devra considérer comme acceptables soit des chaînes de méta-justification circulaires, soit des chaînes de méta-justification infinies.²⁴⁷ Les chaînes infinies ayant déjà été écartées, cela veut dire qu'un philosophe qui veut éviter le scepticisme de la méta-régression est *obligé*, d'une manière ou d'une autre, d'accepter des chaînes de méta-justification circulaires.²⁴⁸

D'après ce que je viens de dire, la meilleure solution au problème de la méta-régression consisterait donc à étayer nos expériences perceptives par une chaîne de méta-justification circulaire. Cette réponse pose deux types de problèmes. Premièrement, il semble évident que la méta-circularité est viciieuse dans le « cas des jurés », donc il semble que le sceptique ait un très bon argument pour rejeter (en plus des chaînes infinies) les chaînes circulaires. J'ai argumenté pour l'instant que les chaînes méta-circulaires étaient notre « meilleure chance » contre le sceptique, mais le sceptique n'aurait-il pas raison, après tout ? Deuxièmement,

²⁴⁶ Si l'on compare avec le cas analogue de l'engagement de clôture, l'idée de se restreindre à un principe « strictement négatif » reviendrait à accepter les attitudes d'inconséquence (affirmer p tout en affirmant qu'il est possible que q ou que $\sim q$) *du moment* que l'agent S n'exprime pas *positivement* la croyance que $\sim q$. Cela revient donc à accepter le principe de non-contradiction, mais à ne voir aucun problème particulier dans la « situation de l'inconséquent ». Cette stratégie n'est clairement pas acceptable pour l'engagement de clôture (ou de « conséquence »), et elle ne me semble pas plus acceptable pour l'engagement de méta-cohérence.

²⁴⁷ Les chaînes terminées étant exclues par (MC Disj.) lui-même.

²⁴⁸ Il y a une manière très rapide de montrer que l'alternative entre scepticisme et circularisme est vraiment exhaustive. Si l'on accepte de poser le problème en décrivant les « sources épistémiques » avec une précision de « grain » très gros (de niveau très général), alors on peut s'intéresser à la méta-justification non plus de « la vision », ni de « la perception » (qui est déjà plus générale), mais à la méta-justification de « notre raison » entendue comme l'ensemble de nos facultés naturelles. Or à ce niveau de description, il est trivial que la circularité ne peut être évitée : « Every kind of reasoning for the veracity of our faculties, amounts to no more than taking their own testimony for their veracity » (Reid 1785 VI, ch.5, n.7). (Bergmann 2006b, 203) soulève, d'une manière différente, ce problème de généralisation qui *oblige* à choisir entre circularité et scepticisme.

quand bien même on arriverait à montrer que les chaînes circulaires (ou que certaines chaînes circulaires) sont épistémiquement acceptables, le problème est que nous ne semblons pas avoir une telle chaîne circulaire à notre disposition, si l'on essaie de décrire notre structure noétique de manière réaliste. Bien sûr, on *pourrait* dire que, pour la perception, nous avons la méta-justification scientifique que j'ai évoquée plus haut (nous savons à partir de la vue qu'il y a des rayons lumineux et une surface rétinienne, un lien causal entre les deux, etc.). Mais si la réponse au scepticisme *dépendait* de ce type de méta-justification, cela voudrait dire qu'avant ces découvertes scientifiques il n'y avait pas de méta-justification, et donc que les croyances perceptives étaient *ultima facie* injustifiées (réfutées) pour tout philosophe qui considérerait à ce moment-là la question méta. Autrement dit : Aristote ou Epicure n'étaient en fait *pas justifiés* à croire qu'ils avaient deux mains.²⁴⁹ Il est clair que cette conclusion reste une conclusion sceptique radicale, et je ne connais pas de philosophe anti-sceptique qui se soit satisfait d'une telle « réponse » à l'argument sceptique. Un philosophe anti-sceptique devra donc trouver un type de méta-justification qui était accessible à Aristote ou Epicure. Or, pour Aristote et Epicure en tout cas, il n'est pas absolument évident qu'on puisse *en fait* mettre en évidence une indication (fût-elle circulaire) que nos sens sont fiables ; et en tout cas, il est clair qu'Aristote et Epicure n'avaient pas d'indication sur la *manière* dont nos sens sont fiables. Dans les paragraphes qui suivent, je commencerai donc par rejeter une version forte du principe de méta-cohérence qui exige une théorie sur la *manière* dont nos sources épistémiques sont fiables ; la seule contrainte que nous (ou Aristote et Epicure) devons satisfaire est d'avoir une justification (nécessairement circulaire) à croire *que* nos sens sont fiables. Je répondrai ensuite aux deux problèmes que je viens de présenter. Au premier problème (l'idée que toute circularité épistémique est inacceptable), je suivrai la stratégie de (Bergmann 2004) et de (Schmitt 2004) qui consiste à distinguer différents types de circularité épistémique, certaines vicieuses, d'autres vertueuses. Et pour répondre au second problème (la méta-justification introuvable), je présenterai deux types de méta-justifications circulaires, toutes deux inspirées de la tradition dite « du Sens Commun », qui permettent de répondre à l'argument de la méta-régression. Je montrerai enfin que chacune de ces deux réponses permet du même coup d'étayer l'expérience du libre arbitre.

Un type de méta-justification qu'Aristote et Epicure n'avaient clairement pas, pas même sous la forme d'une chaîne épistémiquement circulaire, c'est une justification à croire que nos sens sont fiables *de telle et telle manière*. Il ne s'agit pas de dire, bien sûr, qu'il n'existait pas de théories de la vision, par exemple, dans l'antiquité : au contraire, il y en avait en grand nombre, l'alternative principale étant entre les « théories de l'émission » et les « théories de l'intromission ». ²⁵⁰ Les théories de l'émission, défendues notamment par Euclide et Ptolémée, supposaient que la lumière se déplace de l'œil vers la chose regardée ; les théories de l'intromission au contraire, défendues par les atomistes et par Aristote, supposaient le déplacement de certaines entités de l'objet vu vers l'intérieur de l'œil. Platon, quant à lui, adopta une solution mixte entre ces deux modèles. Ce qui est clair, dans ces discussions, c'est que chacun supposait l'existence d'une certaine dépendance entre l'esprit et l'objet (*via* l'œil) mais que, faute d'indication directe sur son fonctionnement, chacun construisait des hypothèses sur ce qu'il *pouvait* être. C'est sur la base de la croyance *que* la vision est fiable qu'étaient élaborées des hypothèses sur la *manière* dont elle est fiable. Le rapport entre la croyance « que » et la croyance « comment » (pour les questions de fiabilité) ne va évidemment pas toujours dans ce sens-là. Imaginez qu'un inventeur vous présente une nouvelle machine très complexe, et vous explique en détail le fonctionnement de chaque pièce, de telle sorte que vous comprenez finalement que les données produites par la machine correspondront

²⁴⁹ Dans une telle théorie, le scepticisme aurait été, par chance, réfuté à un certain moment de l'histoire humaine, lorsque nous avons acquis une bonne théorie scientifique de la perception.

²⁵⁰ Voir notamment (Lindberg 1976, chap. 1).

nécessairement aux dates des futures éclipses de lune. Dans ce cas, c'est sur la base de vos croyances « comment » que vous pourrez produire la croyance « que » (la croyance *que* cette machine est une source fiable d'informations sur les éclipses). Inversement, supposez que vous assistiez au spectacle d'un illusionniste qui, dix fois de suite, est parvenu à donner le numéro de carte bancaire de personnes désignées au hasard dans l'assistance. Naturellement, vous vous demanderez comment il peut faire cela ; c'est-à-dire que sur la base de votre croyance « que » (que les prédictions de l'illusionniste sont fiables), vous soulèverez (et parviendrez peut-être à résoudre) la question « comment ». Cette distinction permet de formuler un nouveau principe de méta-cohérence (positive) :

(MC-Comment) Si la croyance de S que *p* est *prima facie* justifiée par un fondement épistémique F, et si S soulève à *t* la question de la vérité de $K_F p$, alors

(a) soit S est en mesure d'avoir à *t* une croyance justifiée sur la manière dont F procure une connaissance (est fiable)

(b) soit la justification *prima facie* pour *p* se trouve par là même réfutée (pour S à *t*).

Pour les raisons qu'on vient de voir, il est clair qu'un tel principe de méta-cohérence amène directement au scepticisme, tout du moins pour les auteurs antiques. Mais il est clair aussi qu'un tel principe n'a aucune plausibilité : une fois que je me suis convaincu de la fiabilité de l'illusionniste en matière de cartes bancaires, je peux légitimement utiliser ses déclarations comme source d'information fiable même si je n'ai aucune idée de la manière dont fonctionne cette source d'information. Lorsque je me fie à une source F, je suis certainement engagé à croire que F est fiable, mais on ne voit pas pourquoi il faudrait accepter un réquisit tel que :

(MR-Comment) Le fait de croire catégoriquement que *p* nous engage, après réflexion, à donner une théorie du processus qui rend cette croyance fiable.

Dans notre recherche d'une méta-justification circulaire réaliste pour la perception, il ne s'agira donc pas de rechercher une indication (circulaire) du « comment » de la fiabilité perceptive ; tout ce dont nous aurons besoin, c'est d'une justification à croire *que* notre perception est fiable. Et évidemment, cette précision vaudra également pour l'introspection du libre arbitre : pour répondre à l'objection de l'impression non-étayée, il ne faut pas partir avec l'idée que l'expérience du libre arbitre ne pourrait être étayée justificationnellement que par des indications sur la manière exacte dont notre expérience est fiable. La seule contrainte légitime est l'exigence modeste d'étayer nos expériences par une justification à croire *que* nos expériences sont fiables, une méta-justification dont on a vu qu'elle ne pouvait être que circulaire. Reste à savoir si cette exigence modeste peut être satisfaite. Pour le sceptique, elle ne le peut pas, car une méta-justification circulaire ne peut *jamaïs* être une bonne façon d'étayer une croyance. Il faut donc écarter cette objection avant de montrer le type de méta-justification circulaire dont Aristote et Epicure pouvaient disposer.

La circularité épistémique est-elle intrinsèquement vicieuse ? La grande force des philosophes (infinistes ou sceptiques) qui refusent la circularité épistémique, c'est que *certaines formes* de circularité épistémique sont *incontestablement* vicieuses, et que leur vice *semble* bien résider dans leur circularité. J'ai déjà donné comme exemple le « cas des jurés » ; un autre type d'exemple qui tient une place considérable dans la littérature actuelle, est le cas des arguments du type « bons antécédents » (*trackrecord*) ou argument « à la Münchhausen » (*bootstrapping*)²⁵¹ dont l'exemple le plus célèbre est celui de la jauge d'essence utilisé par Jonathan Vogel :

²⁵¹ La terminologie est variable pour ce type d'arguments : (Cohen 2002) parle de « connaissance facile » (*easy knowledge*) tandis que l'expression de *bootstrapping* est due à (Vogel 2000). L'expression de *bootstrapping*

Cas de la Jauge d'Essence

Roxane conduit une voiture dont la jauge d'essence est fiable et en état de marche. Elle n'a jamais regardé le statut de la jauge, ou d'autres jauges du même genre ; elle n'a absolument aucune information à ce sujet. Par ailleurs, Roxane forme automatiquement des croyances à propos du niveau d'essence dans le réservoir de la voiture, simplement en consultant la jauge. Par exemple, si la jauge indique « P » elle forme de manière immédiate et directe la croyance que le réservoir est plein. Etant donné que la jauge est fiable, il semble clair que la croyance de Roxane que le réservoir de la voiture est plein, est formée par un processus fiable. Maintenant, Roxane peut également observer l'état de la jauge en lui-même, si elle le souhaite. Roxane note que l'aiguille de la jauge indique « P » au moment où elle croit, en lisant la jauge, que le réservoir est plein. Roxane réunit sa croyance que la jauge indique « P » avec sa croyance que le réservoir est plein, et elle déduit que l'indication de la jauge est correcte en cette occasion. Nous pouvons imaginer que Roxane répète cette étrange procédure un bon nombre de fois, accumulant les croyances que les indications de la jauge sont correctes à plusieurs instants t_i . Finalement, Roxane conclut, par induction, que les indications de la jauge sont généralement correctes, c'est-à-dire que la jauge est fiable. (Vogel 2009, 518–519)²⁵²

Il serait extrêmement contre-intuitif de défendre que le raisonnement de Roxane ne pose aucun problème épistémique, qu'il confère à la fois connaissance et justification. Ce qui nous intéresse ici, c'est uniquement la question de justification : il semble évident qu'un raisonnement à la Münchhausen n'apporte aucune justification au jugement de second ordre « ma source d'information (la jauge) est une source épistémique fiable ». Et le problème de ce raisonnement semble résider dans son caractère circulaire. Mais si *toute* chaîne de méta-justification est circulaire, et donc comparable à un raisonnement à la Münchhausen, alors cela veut dire qu'il nous est impossible d'avoir une méta-justification épistémiquement acceptable.²⁵³

semble être utilisée par la majorité des auteurs. Malheureusement, elle n'est pas facilement traduisible en français : elle renvoie à l'expression « lifting yourself up by your own bootstraps » (se débrouiller par soi-même) qui, littéralement, décrit une personne en train de se soulever du sol en tirant sur ses propres lacets. On retrouve ici la même image que celle du baron de Münchhausen qui parvient à sortir d'un marécage en se tirant par ses propres cheveux. C'est ainsi que Hans Albert (1975, 15) a choisi de désigner le trilemme sceptique sous le nom de « trilemme de Münchhausen ». La traduction française la plus compréhensible pour le terme « *bootstrapping argument* » est donc, me semble-t-il, « argument à la Münchhausen ».

²⁵² Le cas de la jauge d'essence a été formulé initialement par (Williams 1991, 547–549). Il est devenu l'exemple typique des arguments à la Münchhausen suite à l'article (Vogel 2000).

²⁵³ Il est important de noter qu'à l'origine le cas de la jauge d'essence a été utilisé contre les théories externalistes (fiabilistes) de la *connaissance*, et non pas comme un problème pour la *justification* internaliste. Pour (Vogel 2000), la leçon à tirer des raisonnements à la Münchhausen, c'est qu'ils ne peuvent en aucune manière conférer une *connaissance*, alors qu'un philosophe fiabiliste est (d'après lui) *obligé* de dire qu'ils confèrent une connaissance. Et Vogel pensait pouvoir en tirer non seulement une conclusion négative contre l'externalisme mais aussi une conclusion positive en faveur de l'internalisme. Mais la discussion des cas de *bootstrapping* a connu un infléchissement important lorsque plusieurs auteurs, en particulier (Cohen 2002; Van Cleve 2003; Bergmann 2004), ont fait observer que le même problème de circularité s'appliquait aussi bien aux épistémologies internalistes qu'externalistes, de telle sorte que l'argument de Vogel permettait au pire d'établir le scepticisme, mais pas d'établir l'internalisme. (Vogel 2009) tente de rétablir un argument internaliste de la manière suivante : s'il est probablement vrai que tout épistémologue (même internaliste) est engagé à accepter *certaines formes* de circularité, le problème de l'externaliste est qu'il est engagé à accepter *la pire forme* de circularité, celle qu'en aucune manière on ne pourra sauver même si l'on choisit (par anti-scepticisme) de tolérer parfois la circularité épistémique. Cette stratégie n'est pas absolument décisive, parce que les théories les plus plausibles pour distinguer entre formes acceptables et formes inacceptables de circularité peuvent être appliquées,

Pour éviter la conclusion sceptique sans pour autant légitimer les arguments évidemment inacceptables, du type « cas des jurés » ou du type « cas de la jauge d'essence », il faut donc montrer que ces arguments représentent une forme de circularité *très particulière* et que leur défaut vient précisément de leurs particularités, pas de la circularité en tant que telle. Autrement dit, il faut tracer une démarcation entre les formes inacceptables de circularité (cas des jurés, cas de la jauge) et les formes acceptables (celles qui permettront d'étayer ultimement les expériences perceptives). Si l'on parvient à diagnostiquer le défaut propre des arguments circulaires inacceptables, on pourra du même coup proposer pour la perception (et pour le libre arbitre) une méta-justification circulaire qui ne soit pas affectée de ce défaut. Il me semble qu'on peut repérer deux diagnostics principaux, et corrélativement deux types de solutions circularistes au problème de la méta-régression. Un premier diagnostic situe l'origine du problème dans le caractère *non naturel* des sources épistémiques que l'on tente d'étayer circulairement (le témoin Hank, la jauge de la voiture) ; la solution consistera alors à étayer de manière circulaire les sources épistémiques *naturelles* et seulement celles-ci. C'est ce que j'appellerai la solution « naturaliste ». ²⁵⁴ Un deuxième diagnostic situe le défaut des arguments à la Münchhausen dans le fait... qu'ils sont des *arguments*, c'est-à-dire dans leur nature *inférentielle* ; la solution, dans ce cas, consistera à n'accepter la circularité épistémique que pour une source épistémique qui s'auto-justifie de manière *basique*. Pour des raisons qui apparaîtront plus loin, j'appellerai cette solution « mooréenne ». En dépit de différences structurelles importantes, les solutions naturaliste et mooréenne ne sont pas toujours clairement distinguées ; cela est dû, probablement, au fait que l'une et l'autre solutions sont présentes dans les ouvrages de Thomas Reid, et chez les auteurs qui se revendiquent après lui de « la » tradition philosophique du Sens Commun. Je présenterai ces deux solutions « de Sens Commun » sans chercher à trancher décisivement entre elles, car elles peuvent indifféremment s'appliquer à la défense de l'introspectionnisme.

Commençons par la solution naturaliste. Le diagnostic dont part cette solution n'a rien à voir avec le *type de raisonnement* qui génère une méta-justification circulaire ; ainsi, (Bergmann 2006b, 199) imagine une personne, Becky, qui n'a jamais mis en doute la fiabilité de ses sens, et n'a même jamais considéré la question de leur fiabilité jusqu'au jour où elle s'aperçoit que telles et telles données des sens lui apportent une raison de croire que ses sens sont fiables. Becky aurait ainsi suivi un raisonnement structurellement semblable à celui de Roxane ou du juré 1. Et pourtant, Bergmann fait remarquer qu'il y a une différence importante entre ce scénario et les deux scénarios problématiques. Dans le cas des jurés, le juré 1 commence son raisonnement avec un *doute* positif contre la fiabilité de Hank ; il n'a pas confiance, *de prime abord*, dans la fiabilité de Hank. Et dans ces conditions, Bergmann reconnaît que le raisonnement circulaire ne lui permettra pas d'acquérir une méta-justification. De même, dit Bergmann, si l'anti-sceptique discute avec quelqu'un qui a *positivement* et *de prime abord* des doutes sur la fiabilité des sens (pour autant qu'une telle personne existe), aucun raisonnement à la Münchhausen ne permettra au sceptique de se sortir du marais de

au moins à première vue, à une conception externaliste de la connaissance. En utilisant une telle distinction, un externaliste peut probablement expliquer en termes proprement externalistes que le raisonnement *de Roxane* est inacceptable. C'est en tout cas la stratégie suivie par (Bergmann 2004; Zalabardo 2005). Je n'approfondirai pas plus ce débat complexe car c'est uniquement la circularité des *justifications* qui nous intéresse ici.

²⁵⁴ Le terme « naturaliste » n'est pas à prendre ici dans le sens de Dewey ou de Quine, courant aujourd'hui, c'est-à-dire dans le sens d'une philosophie qui prend exclusivement comme point de départ, comme méthode, et/ou comme modèle, les sciences de la nature (cf. (Rea 2002, chap. 2–3) pour une élucidation de cet usage du terme). J'utilise ici le terme dans le sens faible d'une épistémologie pour laquelle il est épistémologiquement décisif qu'une faculté appartienne à la *nature* de l'agent considéré. En ce sens, le *proper functionalism* de Plantinga et Bergmann est « naturaliste » (bien que ces auteurs ne soient pas naturalistes au sens de Quine).

défiante épistémique où il se trouve. Le problème est qu'en présence de ces doutes positifs, les *prémisses mêmes* du raisonnement sont injustifiées (fondationnellement réfutées). Mais Becky ne se trouve pas dans cette situation initiale : elle *n'a jamais eu* de questionnements ou de doutes au moment où elle acceptait les prémisses (perceptives). La différence pertinente entre le juré 1 et Becky ne tient donc pas au type de raisonnement utilisé mais au contexte dans lequel il est produit : l'un a (et l'autre n'a pas) des questionnements et des doutes initiaux sur la fiabilité de la source qui « cherche » à s'auto-justifier.²⁵⁵ C'est ce que Bergmann (2004) a appelé un « contexte de source mise en question ». Le cas de Roxane a obligé Bergmann à ajouter une précision : en effet, on pourrait supposer que Roxane ne mette à aucun moment en doute la fiabilité de la jauge ; cela ne semble pas suffire pour que son raisonnement soit impeccable. Quelle est alors la différence entre Roxane et Becky ? Intuitivement, la différence est que Roxane est dans une situation où elle *aurait dû* mettre en question la fiabilité de la jauge (ce qui nous aurait ramené au cas précédent d'inacceptabilité), tandis que, à première vue, il était parfaitement légitime pour Becky de ne pas mettre en question initialement la fiabilité de ses sens. Ainsi, Roxane et le juré 1 (mais pas Becky) se trouvent dans une situation « situation QD » :

Situations QD : Situations dans lesquelles, avant la formation de la croyance [épistémiquement circulaire de type « S est fiable »], le sujet *a* ou *devrait avoir* des questionnements et des doutes sérieux sur la crédibilité ou la fiabilité [des croyances formées par S]. (Bergmann 2006b, 198)

Pour Bergmann, c'est seulement dans ce genre de contextes que les raisonnements à la Münchhausen sont illégitimes. Becky a raison de se fier initialement à ses sens, elle peut donc les étayer circulairement. Roxane ne devrait pas se fier initialement à la jauge, elle ne peut donc l'étayer circulairement.²⁵⁶

Plusieurs auteurs ont remarqué que, en l'état, cette réponse ne nous permettait pas de faire beaucoup de progrès.²⁵⁷ En effet, elle suscite immédiatement la question suivante : *qu'est-ce qui fait* que Roxane « devrait » mettre en doute la jauge, tandis que Becky « ne devrait pas » mettre en doute ses sens ? Quelle est la différence pertinente qui fonde cette différence normative ? Une réponse très intuitive à cette question est la suivante : c'est que les sens sont une source épistémique *naturelle* tandis que la jauge ou le témoin Hank ne le sont pas.²⁵⁸ Ce type d'intuition se retrouve, par exemple, dans la distinction de (Goldman 1986, 93–95) entre les « processus natifs » qui confèrent la justification sans étayage préalable, et les « méthodes acquises » qui doivent être validées d'une manière ou d'une autre pour conférer de la justification aux croyances qu'elles produisent.²⁵⁹ L'origine de cette intuition,

²⁵⁵ On retrouve ici, cette fois au niveau méta-épistémologique, la distinction de (Pryor 2004, 368) entre échec *dialectique* d'un *argument* et échec *épistémique* du *raisonnement* correspondant (voir chapitre 2 p.165). De même qu'une déduction lourde de $\sim HS$ peut constituer un bon raisonnement sans être dialectiquement opérante sur le sceptique, de même une induction circulaire de KA peut constituer un bon raisonnement sans apporter dialectiquement une réponse au sceptique.

²⁵⁶ Frederick Schmitt donne les mêmes conditions que Bergmann pour la production d'une inférence circulaire légitime : « This requires that I have no legitimate doubt as to whether S is reliable sufficient to undermine the justifiedness of my S beliefs » (Schmitt 2004, 395).

²⁵⁷ Voir en particulier (Vogel 2009, 539) pour une objection différente de celle que je formule ici.

²⁵⁸ De façon étonnante, Bergmann lui-même ne pose pas cette seconde question, et *a fortiori*, n'y répond pas. De fait, la théorie circulariste qu'il développe finalement est une solution « mooréenne », c'est-à-dire une solution dans laquelle la seule circularité épistémique existante est une circularité *basique*, et *non pas* une circularité inférentielle dans des circonstances initiales favorables. On le retrouvera donc plus loin.

²⁵⁹ (Schmitt 2004, 396–397), à la différence de Bergmann, envisage cette possibilité de fonder la différence normative (entre sources qu'on devrait ou ne devrait pas mettre en doute) sur la différence factuelle entre sources natives et sources acquises. Il reconnaît entre les deux couples de notions une certaine coïncidence, mais (pour

dans la tradition analytique, remonte clairement à l'épistémologie de Reid. Pour Reid, la raison que nous avons de nous fier à nos sens est fondamentalement la même que la raison de nous fier à la « conscience » (cette capacité qui nous apprend l'existence de nos idées) :

Qui se porte garant de notre conscience ? Un homme peut-il prouver que sa conscience ne saurait le tromper ? Personne ne le peut ; et nous ne pouvons donner de meilleure raison de lui faire confiance, que le fait que chaque homme, lorsque son esprit est sain, est déterminé par la constitution de sa nature à y croire de manière implicite, et à se rire ou prendre pitié de l'homme qui doute de son témoignage. (Reid 1764, sec. I.3)

La meilleure justification dont nous disposons pour croire que « notre conscience et notre perception sont fiables » réside donc dans leur caractère *naturel*, dans le fait qu'elles font partie des facultés naturelles de l'esprit humain, ou pour le dire dans les termes de Reid des « principes du Sens Commun ». Autrement dit, Reid semble accepter un principe tel que :

Présomption de fiabilité

Lorsqu'une source épistémique *S* fait partie des facultés cognitives naturelles de l'esprit humain, nous avons une justification à croire que ses données constituent des connaissances.²⁶⁰

C'est cette conviction fondamentale qui détermine tout le projet de l'*Inquiry into the Human Mind* : rechercher consciencieusement et empiriquement *quelles facultés* et *quelles croyances* font réellement partie de la constitution naturelle de l'esprit humain, pour pouvoir en conclure que ces facultés et croyances ont une présomption de fiabilité. En ce sens, (Yaffe and Nichols 2000, sec. 1) ont raison de souligner que la philosophie du Sens Commun est bien éloignée d'un appel naïf au *consensus gentium*. Certaines croyances ou habitudes de pensée pourraient fort bien être « statistiquement » universelles sans pour autant faire partie de la structure profonde de l'esprit humain. Or c'est cette structure profonde que Reid cherche à établir : « Reid relève le défi de montrer, chaque fois qu'il prétend qu'une certaine thèse est un principe de sens commun, que cette croyance est effectivement exigée par la nature humaine. » (loc. cit.).²⁶¹ Comment fera-t-on précisément pour déterminer que la confiance

des raisons qui me semblent un peu rapides) rejette finalement qu'elles soient nécessairement coextensives. Et par conséquent, la distinction de Goldman ne permet pas, d'après Schmitt, d'*expliquer* quelles sources sont dignes ou pas de confiance initiale. Malheureusement, Schmitt ne semble pas voir qu'en l'absence de *cette* explication, le problème de départ reste irrésolu : *qu'est-ce donc* qui fait que telles sources sont dignes de confiance initiale et d'autres pas, si ce n'est pas leur caractère naturel ?

²⁶⁰ Ce principe est emprunté, *mutatis mutandis*, à (Plantinga 2000, 148), ainsi que l'expression « présomption de fiabilité ».

²⁶¹ Il y a une deuxième interprétation de ces textes de Reid, une interprétation qui insiste non pas sur le caractère « naturel » de la conscience ou de la perception (qui nous donnerait une justification présomptive à leur faire confiance), mais sur le fait que la nature humaine nous « détermine » de toutes les façons à leur faire confiance (que nous ayons ou non une justification à le faire). Cette interprétation « inévitabiliste » est tout à fait possible, et Reid a explicitement développé cette ligne d'argumentation. Par exemple : « For my part, I will never attempt to throw off [perception]; [...] First, because it is not in my power: why then should I make a vain attempt? » (Reid 1764, sec. VI.20). Quel que soit l'intérêt philosophique de cette stratégie, il est important de noter qu'elle n'est *pas* une stratégie de méta-justification. Ce que montre (éventuellement) l'argument inévitabiliste, c'est que *même si nous n'avions aucune justification* à croire à la fiabilité de nos sens, et même si nous étions à cause de cela dans un état épistémique profondément défectueux, il ne serait pas en notre pouvoir de remédier à ce défaut épistémique, et donc nous n'aurions pas de bonne raison pratique d'*essayer* d'y remédier (en abandonnant nos croyances perceptives). Autrement dit, même si nous étions *épistémiquement* défectueux dans nos croyances perceptives, nous serions cependant *déontologiquement* impeccables. Mais ce que je cherche ici à élucider, c'est seulement la question *épistémique* (interne) : avons-nous une raison épistémique (interne) de nous fier à nos sens ? A cette question, l'inévitabilisme en tant que tel n'apporte pas de réponse.

accordée aux sens fait partie de la nature humaine ? Une observation anthropologique minimale semble suffisante pour établir ce point. Peut-être objectera-t-on que *l'observation* anthropologique repose sur le témoignage des sens et qu'une telle méta-justification serait alors circulaire... mais la solution naturaliste consiste justement à dire qu'une méta-justification circulaire *peut* être épistémiquement impeccable lorsqu'elle repose ultimement sur une source naturelle. On pourrait résumer la chaîne circulaire de la manière suivante : grâce aux données délivrées par nos facultés perceptives (naturelles), nous sommes justifiés à croire que « nos facultés perceptives sont naturelles », ce qui nous donne une justification (au moins *prima facie*) à croire que « nous facultés perceptives délivrent des connaissances ». ²⁶² On peut noter, incidemment, que ce type de chaîne de méta-justification, si elle est épistémiquement circulaire, n'est pas à proprement parler un argument de type « bons antécédents » (*track-record*), ce qui peut sans doute être porté à son crédit. ²⁶³

Supposons qu'on adopte une solution « naturaliste » au problème sceptique de la méta-régression ; quelle serait alors la conclusion à tirer pour l'introspection du libre arbitre ? La réponse est assez évidente : si l'on admet (pour pouvoir étayer la perception) que nous avons une justification *prima facie* à considérer comme fiables les facultés qui appartiennent à la nature humaine, il est clair que l'on pourra adopter cette même défense pour l'expérience du libre arbitre telle que je l'ai définie au chapitre 1. On peut mettre en doute qu'il fasse partie de la nature humaine d'avoir une expérience *libertarienne* ou *incompatibiliste*. Mais on ne peut pas mettre en doute qu'il fasse partie de la nature humaine d'avoir une certaine expérience de nous-mêmes et de nos actions, exprimée en français par « libre », en anglais par « free », et en toute langue par un terme qui permet de traduire ceux-ci, et que cette expérience ait une certaine propension naturelle à provoquer, dans toute culture, un conflit intuitif (peut-être illégitime) avec la notion de déterminisme. Donc la solution « naturaliste » constitue une manière plausible d'apporter une méta-justification positive à notre expérience du libre arbitre : nous avons une justification *prima facie* à croire que notre expérience du libre arbitre (au sens indexical) constitue une source de *connaissance* ; c'est qu'il fait partie de la nature humaine, dans un bon état de fonctionnement, de produire une telle expérience, et que nous devons présumer la fiabilité de nos facultés naturelles en bon état de fonctionnement. ²⁶⁴

²⁶² Pour être tout à fait précis, il faudrait ici distinguer deux versions de la solution « naturaliste ». Dans une version proprement « reidienne », le caractère naturel de la source doit être *connu* par le sujet, et c'est précisément cette croyance (« ma source pour croire *p* est naturelle ») qui constitue sa justification (interne) pour croire « ma source pour croire *p* est fiable ». Dans une version « externaliste » du naturalisme, il suffit que la source soit *effectivement* naturelle (quoi qu'en pense le sujet) pour qu'elle puisse être étayée de manière circulaire (quelle que soit la chaîne circulaire utilisée, par exemple un argument de type *track-record*). A ce stade de la discussion, il est clair que la version externaliste ne peut pas nous satisfaire, puisque nous avons rejeté l'externalisme au chapitre précédent. Ce que nous cherchons ici, c'est un justifieur interne positif pour la croyance « la perception est fiable » ; le naturalisme *peut* apporter un tel justifieur, mais seulement dans sa version reidienne.

²⁶³ En effet, on pourrait avoir l'intuition persistante qu'un argument de type *track-record* a quelque chose d'intrinsèquement vicieux *même lorsqu'il est fondé sur une source épistémique naturelle*. Mais la solution naturaliste de type « reidien » (cf. note précédente) n'est pas menacée par une telle intuition, car elle pourrait être complétée par une théorie du caractère problématique des arguments spécifiquement *track-record*. Un diagnostic possible consisterait par exemple à dire que *l'induction* sur laquelle reposent les arguments *track-record* est extrêmement faible. En effet, cette induction semble être une *inférence à la meilleure explication* dont l'*explanandum* est un ensemble de corrélations : « S indique *p1* » or « *p1* », « S indique *p2* » or « *p2* », etc. Mais il serait absurde que l'agent épistémique A « s'étonne » de cette liste de corrélations si « S indique *p1* » est justement sa *raison* de croire « *p1* », etc. Il n'y a là rien à expliquer, ou si l'on veut dire qu'il y a un *explanandum*, la meilleure explication de la corrélation n'est pas que « S est fiable », mais tout simplement que « chacune de ces informations a été tirée justement de S » (une explication qui est *déjà connue* du sujet A, et qui *implique* la corrélation artificiellement présentée comme base d'induction).

²⁶⁴ La stratégie naturaliste pour le libre arbitre a été défendue, notamment, par C. A. Campbell (1931, 137) : l'affirmation du libre arbitre, dit-il, « nous arrive avec une autorité bien plus grande que celle des lois

Passons à présent au deuxième type de méta-justification circulaire « de Sens Commun » : la solution mooréenne. Cette solution part d'un diagnostic légèrement différent sur ce qui ne va pas dans un argument à la Münchhausen : le problème de ce type d'argument n'est pas un simple problème de *force dialectique* contre un sceptique qui entretient des doutes illégitimes ; le problème tient au *raisonnement* lui-même, c'est-à-dire au fait qu'il s'agit d'une méta-justification circulaire *et inférentielle*. Pour répondre au problème de la méta-régression, il suffira alors de faire observer que « il est possible qu'il y ait des [croyances épistémiquement circulaires] non-inférentielles » (Bergmann 2006b, 181).²⁶⁵

Pourquoi la circularité *inférentielle* serait-elle plus problématique que la circularité « basique » ? On peut le voir en réfléchissant au type de raisonnement « naturaliste » envisagé dans les paragraphes précédents. Imaginez le cas de Nathan (le frère de Becky) :

Cas de Nathan

Supposez que Nathan ne se soit jamais posé la question de la fiabilité de ses sens avant une certaine date *t1* : avant *t1*, il a toujours accepté leurs données (par ex. « il y a une table en face de moi ») de manière parfaitement justifiée (en l'absence de tout réfutateur fondationnel). En particulier, il a accepté les informations perceptives sur les activités cognitives des agents humains, et bien qu'il n'ait pas d'informations scientifiques sur le fonctionnement de la lumière et de la rétine, il a pu un jour (à *t2*) par simple observation anthropologique (accessible à Aristote et Epicure) conclure que les croyances perceptives faisaient partie de la nature même de l'esprit humain. Un peu plus tard (*t3*) Nathan réfléchit sur cette information, et il en induit que la perception sensible est fiable lorsqu'elle se trouve en conditions de bon fonctionnement.

Le problème d'un tel raisonnement est le suivant : ce raisonnement n'a d'intérêt que s'il apporte à Nathan une méta-justification pour la perception (si Nathan a *dès le départ*, à *t1*, une méta-justification pour la perception, alors le raisonnement est superflu). Donc, il faut supposer que, à *t1*, Nathan *n'a pas* de justifieur disponible pour croire que la perception est fiable, si la question se posait à ce moment-là. Bien sûr, un défenseur de la circularité inférentielle, tel que (Alston 1986),²⁶⁶ insistera sur le fait que Nathan *ne se pose pas* la question, justement, et qu'on n'a pas de raison de dire qu'il *devrait* se la poser.²⁶⁷ Mais tout cela est bien accidentel : il *pourrait* se produire que Nathan rencontre à *t1* un philosophe (sceptique ou non) qui lui pose la question. Que peut-il faire alors, étant donné qu'il n'a pas encore à ce stade la méta-justification inductive de *t3* ?²⁶⁸ Il peut se mettre à croire que sa

scientifiques bien authentifiées [...] parce que notre nature nous *oblige* à juger ainsi, semble-t-il. En tant que telle, cette affirmation ne peut pas être légitimement sommée de fournir des raisons positives d'y croire, mais seulement d'écarter de manière adéquate les raisons avancées pour ne pas y croire ». La naturalité est clairement ici le critère d'une justification *prima facie* ; ce que Campbell ne distingue pas très clairement, c'est qu'il s'agit d'une méta-justification *prima facie* et pas seulement d'une justification de premier ordre.

²⁶⁵ Bergmann ne considère pas ses réflexions comme une *condamnation* des solutions circularistes inférentialistes. La difficulté de sa position, c'est que certaines de ses réflexions permettent en même temps de souligner la force relative d'une solution circulariste proprement *basique*. Il me semble donc que, comme Reid, Bergmann offre les éléments des deux solutions, sans se sentir obligé de les distinguer clairement, ni *a fortiori* de prendre parti.

²⁶⁶ Voir aussi (Bergmann 2006b, 203–204).

²⁶⁷ Pourquoi en effet *devrait-il* se poser cette question à *t1* ? Il semble que la seule raison serait un principe universel tel que : « Un agent épistémique doit toujours envisager la proposition *Kp* pour toute croyance *p* qu'il entretient ». Mais une telle exigence est trivialement insatisfaisable pour un esprit fini, puisqu'elle est réursive à l'infini. Elle a donc très peu de plausibilité.

²⁶⁸ On peut supposer qu'à ce stade il n'a pas encore *les moyens* de faire cette induction, parce qu'il n'a pas encore réuni suffisamment d'informations sur la nature humaine ; on imagine Nathan au *début* de sa collecte d'informations perceptives qui vont former sa base inductive.

perception est fiable, mais alors cette croyance sera *injustifiée* puisqu'il n'a pas de méta-justifieur disponible à ce stade, et une croyance *Kp injustifiée* ne fournit pas une méta-justification. Et donc, dans ce cas, ses croyances perceptives seront réfutées fondationnellement. Il peut d'autre part laisser ouverte la question méta-épistémique, c'est-à-dire douter que sa perception soit fiable, mais là aussi, ses croyances perceptives se trouveront réfutées fondationnellement.²⁶⁹ Donc dans tous les cas, *si par hasard* la question méta était posée à Nathan un peu trop tôt, ses croyances perceptives seraient réfutées à *t1* et il n'aurait plus les moyens de poursuivre son raisonnement jusqu'à la méta-justification donnée en *t3*. Ces considérations sur le « cas malchanceux » où la question est posée à Nathan un peu trop tôt amènent à douter également de la légitimité du raisonnement *dans le cas chanceux*. Les prémisses du raisonnement ont une sorte de *fragilité* méta-épistémique qui ne semble pas pouvoir fonder une véritable méta-justification.

Tel est en tout cas le genre de considérations qui peut amener à préférer une circularité « basique » plutôt qu'inférentielle. Encore faut-il montrer pourquoi une circularité basique évite ce genre de problèmes. La solution apparaît immédiatement dès qu'on décrit ce qu'est une source circulairement et basiquement auto-justifiante : il s'agit d'une source épistémique *S* qui délivre à la fois des informations de premier ordre (*p*, *q*, etc.) et qui *par ailleurs* délivre basiquement une information du type « *S* est fiable » ou « *p* délivré par *S* constitue une connaissance ». Elle délivre ce deuxième type d'information *basiquement*, c'est-à-dire immédiatement, sans aucune étape ni chronologique ni logique. Il n'y a donc aucun moment, aucune étape chronologique ni logique, où ses informations de premier ordre sont méta-épistémiquement fragiles : à chaque instant *t1* où elle délivre une information de premier ordre *p*, elle pourrait dès cet instant et immédiatement produire l'information « *p* constitue une connaissance ». Une telle solution a donc deux caractéristiques importantes que je vais préciser à présent : d'une part, elle suppose l'existence de certaines sources épistémiques (ou d'une seule) capables de délivrer des expériences dont le contenu est immédiatement de deuxième ordre ; d'autre part, elle suppose que, parmi ces sources, l'une d'elles au moins apporte une information de deuxième ordre *à propos d'elle-même*. Comme on va le voir, l'une et l'autre caractéristiques trouvent leur origine historique dans la théorie reidienne des « premiers principes ».

Pour Reid, nos facultés intellectuelles délivrent parfois *de manière basique* des croyances (ou expériences doxastiques) *de deuxième ordre*. En effet, Reid met la croyance en la fiabilité de la perception sensible au rang des « premiers principes » de la connaissance humaine,²⁷⁰ et par « premiers principes », il entend précisément ces croyances basiquement justifiées sur lesquelles doit nécessairement reposer tout édifice de justification. La définition qu'il donne des premiers principes expose clairement l'idée d'une justification par expérience doxastique :

« des propositions que l'on croit aussitôt qu'on les comprend. [...] la proposition n'est pas déduite ou inférée d'une autre ; elle a l'éclat de la vérité par elle-même, et n'a pas lieu de l'emprunter à une autre. » (Reid 1785 Essay VI, ch.4, p. 555)

Ainsi, lorsque nous considérons la proposition « mes perceptions sont fiables », ou plus particulièrement une proposition du type « je *sais* (par ma vue) que voici une main », ces propositions ont pour moi l'éclat du vrai, elles me semblent vraies, et cette expérience doxastique me donne une justification basique (*prima facie*) à les croire. Est-ce que nous

²⁶⁹ La troisième possibilité (croire positivement que ses sens ne sont pas fiables) n'est pas pertinente puisqu'elle serait *a fortiori* réfutante (fondationnellement), mais que Nathan n'a aucune bonne raison d'aller si loin.

²⁷⁰ « Another first principle is, Those things do really exist which we distinctly perceive by our senses, and are what we perceive them to be. » (Reid 1785 Essay VI, ch.5, p. 587).

avons effectivement de telles expériences doxastiques ? Il semble bien que oui, au moins pour les expériences particulières du type « je *sais* que voici une main ». Est-ce que ces expériences peuvent recevoir un rôle de justification (*prima facie*) ? Il n'y a pas de bonne raison de leur refuser un tel rôle si l'on considère la variété des autres expériences doxastiques justifiantes (expériences doxastiques *a priori* vs expériences perceptives). On voit ici que la stratégie du phénoménalisme charitable, ou « sans discrimination » (que j'ai appliquée au chapitre précédent, p.172, pour défendre le caractère justifiant de l'expérience du libre arbitre) peut tout aussi bien s'appliquer pour les expériences de deuxième ordre. Michael Huemer a évidemment remarqué le bénéfice qu'il pouvait tirer de son phénoménalisme charitable pour les croyances de deuxième ordre, et les expériences doxastiques constituent donc pour lui (comme pour Reid) le cœur de la solution au problème de la méta-cohérence.²⁷¹ La solution reidienne de Bergmann utilise également un argument « anti-discrimination » fondé sur la comparaison avec les expériences doxastiques *a priori*.²⁷² Ces expériences doxastiques de deuxième ordre, je les appellerai des « expériences mooréennes », suivant l'expression de David Lewis, qui parle de « fait mooréen » pour la croyance de second ordre selon laquelle « nous savons beaucoup de choses ». ²⁷³ Ce type d'expérience joue en effet dans l'épistémologie de Moore un rôle aussi central que dans l'épistémologie de Thomas Reid : elles constituent des prémisses suffisamment certaines pour constituer la base d'un argument anti-sceptique (la fameuse « preuve de Moore » part de la prémisse de deuxième ordre « je *sais* que j'ai deux mains »), et surtout des prémisses en faveur desquelles je ne peux apporter aucune justification autre que mon expérience doxastique :

Est-ce que je *sais* vraiment que les propositions [telles que « mon corps existe »] sont vraies ? N'est-il pas possible que les *croie* seulement ? Ou que je ne connaisse que leur grande probabilité ? En réponse à cette question, je crois que je n'ai rien de mieux à dire qu'il me semble que je les connais vraiment, avec certitude. (Moore 1993b, 118)

Dans la suite de ce texte, Moore fait observer un point tout à fait fondamental à propos de nos croyances de deuxième ordre : la certitude *que nous savons* (que nous avons un corps, deux mains, etc.) ne requiert pas de savoir *comment nous savons* (ni, à vrai dire, d'avoir une croyance justifiée sur la manière dont nous savons) :

²⁷¹ « Now, when I consider the claim that I know that here is a hand, I might experience a second-order appearance: maybe it just seems to me that I know that here is a hand. This isn't guaranteed to occur, but in this case, for me at least, the second-order appearance does in fact occur. It does seem to me that I know that here is a hand, and I have no defeaters for the claim that I know this. Thus, I have justification for thinking that I know that here is a hand. » (Huemer 2011, 16)

²⁷² « One could argue that commonsense belief in contingent truths is very much like a priori belief insofar as they each have the same sort of experiential ground (*i.e.* a certain sort of seeming). Because it is plausible to take seriously both the existence of justified a priori beliefs as well as the account of them as experience-based, it is also plausible to take seriously both the existence of justified commonsense beliefs in contingent truths as well as an account of them as experience-based. Our starting point is the fact that we *do* seem to have justified noninferential beliefs of each kind. » (Bergmann 2006b, 210)

²⁷³ « We know a lot [...] We have all sorts of everyday knowledge, and we have it in abundance. To doubt that would be absurd [...] It is a Moorean fact that we know a lot. It is one of those things that we know better than we know the premises of any philosophical argument to the contrary. » (Lewis 1996, 549). Comme Lewis ne donne pas à proprement parler une « définition » de l'expression « fait mooréen », on ne sait pas très bien s'il entend par là « un fait de deuxième ordre nous attribuant positivement la connaissance », ou « un fait que nous connaissons mieux que nous ne connaissons les prémisses d'aucun argument philosophique allant en sens contraire ». Les deux caractéristiques ne sont qu'accidentellement réunies dans la proposition « nous savons beaucoup de choses ». Certains auteurs, notamment (Kelly 2005), retiennent la deuxième définition ; d'autres, notamment (Bergmann 2006b, 210), entendent par « vérités mooréennes » ou « faits mooréens » spécifiquement les vérités de deuxième ordre. C'est cet usage que je suis ici.

Nous sommes tous, je crois, dans cette étrange position, que nous *savons* un grand nombre de choses [...] et pourtant nous ne savons pas *comment* nous les savons. (*ibid.*)

On retrouve ici une idée présentée un peu plus haut : pour satisfaire la contrainte de méta-cohérence, il faut trouver une justification à croire *que* nous savons, mais il n'est pas besoin pour autant de trouver une indication justifiée sur la *manière* dont nous savons. Autrement dit, on peut (de manière justifiée) considérer une certaine croyance comme *connaissance* avant d'avoir la moindre théorie sur le *mode* de connaissance qui la caractérise (par exemple, la transmission d'information visuelle par les rayons lumineux) – et déterminer ensuite ce mode de connaissance *sur la base* de la certitude préalable qu'il s'agit d'une connaissance. C'est ce que Roderick Chisholm (1982) a appelé une solution « particulariste » au problème du critère. Mais pour résoudre le problème du critère (ou de la méta-régression), il ne suffit pas qu'il y ait une ou plusieurs sources d'expériences mooréennes pour étayer notre source perceptive, il faut en outre que ces sources elles-mêmes soient étayées ; et à moins de s'engager dans une régression infinie des sources mooréennes, il faudra bien faire apparaître à un moment ou à un autre une circularité épistémique, c'est-à-dire une source mooréenne qui s'étaye elle-même de manière basique.

La méta-justification de la perception supposera donc d'accepter, à un certain maillon de la chaîne, une source mooréenne circulaire. Mais à *quel* maillon faut-il situer la circularité ? Une solution consisterait à situer la circularité au tout premier maillon : c'est-à-dire que la faculté perceptive serait *elle-même* la source épistémique qui délivre de manière basique sa propre justification mooréenne. C'est la solution adoptée par Jim Pryor²⁷⁴, ou un peu plus tôt par le cardinal Mercier, qui d'après Roderick Chisholm (1982, 61–63) est le premier auteur à avoir bien posé le problème du critère ainsi que les conditions d'une solution satisfaisante.²⁷⁵ Il me semble que ce type de solution repose sur une confusion terminologique : quant à l'occasion d'une certaine expérience sensible (*Ep*) nous sommes justifiés à la fois à croire que *p* et à croire que *Ep* est une source de connaissance, il est clair qu'en un sens ces deux croyances « concernent une seule et même expérience », mais cela ne veut pas dire qu'elles sont *justifiées* par la même expérience. La seconde est justifiée par une expérience mooréenne à *propos* de *Ep*. Et il semble terminologiquement étrange de dire que *cette* expérience mooréenne est une « expérience sensible » au même titre que l'expérience *Ep*. Autrement dit, nos expériences mooréennes semblent relever d'une autre source épistémique que la

²⁷⁴ « One relevant question is whether your perceptual experiences only give you justification for believing simple propositions like *There is a hand*, or whether, if you are sufficiently reflective and conceptually well-equipped, they also give you justification for believing the more sophisticated proposition that *you see a hand*. [...] I believe your perceptual experiences give you justification for believing both sorts of proposition. » (Pryor 2000, n. 35)

²⁷⁵ Dans le vocabulaire de Mercier, le « premier problème fondamental de la certitude » (Mercier 1911, 42) est de trouver un moyen d'établir que nos « jugements immédiats » ou « adhésions » sont « motivés », c'est-à-dire qu'ils ne sont pas « le résultat d'une disposition *toute subjective* du sujet pensant » mais « des adhésions *causées* par un motif indépendant de l'acte intellectuel et antérieur à lui, par la *manifestation objective du rapport* dont le jugement est l'expression » (*ibid.*) On notera que dans cette formulation de la question *JKp* Mercier adopte une version forte (causale) de la contrainte anti-hasard. Sa solution est fondée sur la nature prédicative des jugements : pour lui, l'expérience doxastique d'adéquation du prédicat au sujet (« *a* me semble être *F* ») est *en même temps* une expérience d'adéquation de l'intellect à l'objet extérieur (« *a* se manifeste à moi comme *F* ») : « L'objection [du diallèle] a le tort de supposer que le critère de la vérité doit être contrôlé par un autre critère, un jugement par un autre jugement, celui-ci par un troisième et ainsi à l'infini. Or, le critère de la certitude n'est que la manifestation de la connexion objective entre les deux termes du jugement; *un même acte* de réflexion saisit à la fois cette connexion et le caractère nécessitant et exclusif qu'elle imprime à la pensée. » (1911, 246). Bien qu'il ne développe explicitement cette solution que pour les jugements *a priori* (p. 238-243), il soutient qu'elle est « également applicable aux jugements d'expériences ». Donc, pour les expériences sensibles également, ce sera un même acte, une même expérience, qui justifiera le jugement « *a* est *F* » et en même temps permettra de « motiver » ou d'étayer ce jugement.

perception proprement dite, une source épistémique très spéciale que Thomas Reid a appelée « le Sens Commun ». Le Sens Commun est, proprement, la faculté qui nous donne une justification basique en faveur des « Premiers Principes du Sens Commun » (tels que la fiabilité des sens). Et de même que les expériences doxastiques apportées par la perception ont une certaine qualité sensible qui les distingue d'autres expériences doxastiques, de même les expériences doxastiques apportées par le Sens Commun sont associées, d'après Reid, à un sentiment particulier : le sentiment « d'absurdité » (voire de « ridicule ») associé à la négation des premiers principes.²⁷⁶ Si la faculté « Perception » est étayée par la faculté « Sens Commun », qui va étayer cette dernière, et où situera-t-on la circularité recherchée ? La solution reidienne consiste à situer la circularité au niveau du Sens Commun lui-même, une faculté qui a justement pour attribution particulière de délivrer des justifications mooréennes, et qui peut donc en délivrer à propos d'elle-même aussi bien qu'à propos de la perception. Cette lecture de Reid, proposée par (Bergmann 2006b, 208–209) n'est pas sans soulever des questions d'interprétation historique,²⁷⁷ mais on peut voir dans l'extrait suivant que Reid a explicitement envisagé l'idée d'une circularité épistémique située dans une faculté proprement mooréenne :

Qu'est-ce donc qui nous assure de cette vérité fondamentale [à savoir, de la véracité de nos facultés] sur laquelle reposent toutes les autres ? Peut-être que l'évidence, qui ressemble à la lumière par bien d'autres aspects, lui ressemble également en ceci : de même que la lumière, qui est le révélateur de tous les objets visibles, se révèle elle-même en même temps, de même l'évidence, qui est le garant de toutes les vérités, se porte en même temps garante d'elle-même. (Reid 1785 Essay VI, ch.5, p. 593)

Telle est donc la solution circulariste basique (la solution que j'ai appelée « mooréenne ») dans sa version la plus plausible : nos facultés perceptives sont étayées par une source mooréenne (qu'on peut appeler le « Sens Commun »), et cette source mooréenne s'étaye elle-même, de manière circulaire mais non inférentielle (et donc non *fragile*).²⁷⁸

Si l'on adopte cette solution mooréenne au problème du critère (au lieu de la solution « naturaliste ») quelles conclusions doit-on tirer pour l'expérience du libre arbitre ? Il est évident que ce type de solution suggère une réponse très simple à « l'objection de l'impression non étayée » : pour étayer l'expérience du libre arbitre, il suffirait que nous ayons (de par notre faculté mooréenne) une expérience doxastique que « cette expérience

²⁷⁶ « We may observe, that opinions which contradict first principles are distinguished from other errors by this; that they are not only false, but absurd: and, to discountenance absurdity, nature has given us a particular emotion, to wit, that of ridicule, which seems intended for this very purpose of putting out of countenance what is absurd, either in opinion or practice. » (Reid 1785 Essay VI, ch.4, p.567).

²⁷⁷ Le problème d'interprétation principal, étudié par (Lemos 2004, sec. 4.2), est que Reid semble dans certains extraits *refuser* la légitimité de toute justification épistémiquement circulaire, en particulier dans les critiques qu'il adresse à Descartes.

²⁷⁸ Une solution très proche est défendue par Michael Huemer. Pour défendre la Méta-Justification de la perception, Huemer a bien recours à une intuition de type mooréen : “Phenomenal Conservatism holds that its seeming to one that P, in the absence of defeaters, provides one with some justification for believing that P. It seems to me now that there is a hand in front of me. Assuming I have no defeaters for the claim, this gives me justification for thinking that here is a hand. Now, when I consider the claim that I know that here is a hand, I might experience a second-order appearance: maybe it just seems to me that I know that here is a hand. This isn't guaranteed to occur, but in this case, for me at least, the second-order appearance does in fact occur. It does seem to me that I know that here is a hand, and I have no defeaters for the claim that I know this. Thus, I have justification for thinking that I know that here is a hand.” (Huemer 2007, 16). La différence avec la solution « reidienne » de Bergmann tient seulement au fait suivant : Huemer ne *réitère* pas ici la question de méta-cohérence à propos de l'intuition mooréenne elle-même. S'il réitérait la question (notre intuition mooréenne est-elle source de connaissance ?), il serait certainement amené à dire qu'elle l'est à son tour par une intuition mooréenne à propos de notre intuition mooréenne, ce qui revient à la solution de Bergmann.

constitue une source de connaissance ». Avons-nous une telle expérience ? La question ne semble pas avoir été souvent posée en ces termes dans la littérature sur le libre arbitre. Pourtant, comme l'a remarqué (Holton 2006, 1), la phrase la plus citée dans tout le débat analytique sur le libre arbitre semble y donner directement une réponse. Je veux parler de la fameuse déclaration de Samuel Johnson :

Sir we *know* our will is free, and *there's* an end on't.
 (Mon cher, nous *savons* que notre volonté est libre, un point c'est tout)
 (Boswell 1946 AD 1769, *Ætat*. 60)

Samuel Johnson exprime dans cette phrase une intuition mooréenne²⁷⁹ qui est, me semble-t-il, la position la plus naturelle pour un non philosophe qui se pose pour la première fois la question de second ordre à propos de l'expérience du libre arbitre.²⁸⁰ Si l'on n'a pas encore réfléchi à la question du déterminisme naturel et qu'on se concentre, à l'occasion d'une action précise, sur notre expérience (et notre croyance spontanée) de la liberté de cette action, si l'on examine cette conviction honnêtement, sérieusement, et attentivement, il me semble évident que nous ne serons pas amenés à tenir cette conviction pour un simple préjugé ou une simple illusion sans le moindre rapport de dépendance avec son objet. Il me semble évident que notre première intuition sera qu'il s'agit, au moins dans le sens faible que j'utilise dans cette étude, d'une *connaissance*. Dans la littérature philosophique, « l'intuition de Johnson » est très largement approuvée, sans que soit remarquée son caractère proprement mooréen ; Gary Watson (2003, 1) par exemple, considère que cette intuition « est partagée par la plupart d'entre nous », et qu'elle « constitue le point de départ de la discussion philosophique ». Si l'on trouve peu de remarques sur le statut mooréen de cette intuition, c'est seulement, à mon sens, parce que la question a rarement été posée en ces termes, mais on peut trouver chez les auteurs des formulations qui *révèlent* ce type d'intuitions. Le plus intéressant pour ma démarche est qu'on peut trouver ce genre d'indications chez des auteurs compatibilistes tels que John Locke²⁸¹ ou John Fischer.²⁸² Holton lui-même, qui est compatibiliste, donne explicitement raison à Johnson, et soutient que notre expérience d'être libre a une forte présomption d'être une connaissance.²⁸³ On peut avoir des doutes sur l'existence d'expériences doxastiques mooréennes *en général*, mais il me semble clair que *si* l'on accepte leur existence à propos de la perception (c'est-à-dire si l'on accepte de décrire comme « expérience doxastique » notre tendance naturelle à répondre « bien sûr que mes impressions visuelles sont une source de connaissance »), on n'a aucune raison particulière de nier leur

²⁷⁹ On peut s'en apercevoir notamment grâce à l'italique de « *know* », qui est dans l'original.

²⁸⁰ Il est clair que Johnson conçoit la connaissance en question comme une connaissance fondée sur *l'expérience* du libre arbitre ; on peut le voir notamment dans cette autre citation célèbre : « All theory is against the freedom of the will; all experience for it. » (Boswell 1946 AD 1778, *Ætat*. 69)

²⁸¹ « I cannot have a clearer perception of any thing than that I am free » (Locke 1979, 625). Si l'on prend « perception » au sens strict – distinct de « impression perceptive » – il s'agit d'un terme factif et plus précisément d'un terme d'espèce de connaissance (comme « voir » : je ne peux pas dire que « je vois que *p* » si *p* est faux ou si mon impression visuelle que *p* n'est vraie que par hasard). De plus, Locke semble s'attribuer une perception *la meilleure qui soit...* et donc assurément une *connaissance*.

²⁸² Dans son argument « de la prise en otage » (en faveur du compatibilisme), John Fischer s'appuie sur l'intuition que sa conviction d'être libre ne changerait pas en cas de découverte scientifique du déterminisme, mais surtout sur l'intuition qu'elle ne *devrait pas* changer : « I feel confident that this would not, nor should it, change my view of myself and others » (Fischer et al. 2007, 44). Ce « devrait » est manifestement *épistémique* : Fischer exprime ici l'idée qu'une découverte du déterminisme ne supprimerait pas le statut épistémique positif dont jouit l'expérience du libre arbitre... il exprime donc par là une intuition de second ordre à propos de son expérience du libre arbitre : cette expérience constitue une source épistémique solide.

²⁸³ « Our knowledge of our own free will is more certain than any thesis of philosophy [...] we would need very good grounds before rejecting it as illusory. » (Holton 2006, 1–2)

existence à propos du libre arbitre (c'est-à-dire qu'on n'a alors aucune raison de ne pas décrire les réactions à la Johnson comme expressions d'une expérience doxastique mooréenne). La solution « mooréenne » au problème sceptique de la méta-régression permet donc, tout comme la solution « naturaliste », d'offrir directement une réponse à l'objection anti-introspectionniste de l'impression non étayée.

Pour conclure, nous avons donc vu que la requête d'une méta-justification pour l'expérience du libre arbitre est une requête délicate, mais pas plus délicate pour l'expérience du libre arbitre que pour la perception ; si l'on accepte de sauver les expériences perceptives du scepticisme de la méta-régression, quelle que soit la stratégie de sauvetage, on sauvera du même coup l'expérience du libre arbitre. Quelle est la stratégie la plus plausible et la plus utile pour la suite de la discussion ? Ce n'est certainement pas de refuser absolument toute forme d'engagement de méta-cohérence, ni de satisfaire cet engagement par une méta-justification infinitiste. La seule solution plausible, pour le libre arbitre comme pour la perception, est une méta-justification épistémiquement circulaire. Qu'est-ce qui pourrait nous justifier (nous ou Aristote) à croire que nos expériences sensibles (ou nos expériences mathématiques) sont fiables ? Quel type de méta-justification circulaire pourrait étayer nos sens ou nos intuitions mathématiques à la différence d'autres sources épistémiques comme une boule de cristal ou n'importe quelle jauge d'essence ? J'ai soutenu que le bon type de circularité, le bon type « d'auto-justification pertinente », ²⁸⁴ ne pouvait correspondre qu'à deux modèles défendables : soit le caractère *naturel* de ces expériences doxastiques nous justifie à présumer leur fiabilité ; soit nous avons une expérience mooréenne qu'elles constituent effectivement des connaissances. Bien que je penche en faveur de la deuxième solution, je laisserai cette question totalement ouverte dans cette étude, car l'une et l'autre réponses me semblent tout à fait adaptées au cas de l'expérience commune du libre arbitre : cette expérience (à la différence d'une boule de cristal ou d'une jauge d'essence) est à la fois *naturelle* et typiquement accompagnée d'une *intuition mooréenne*. Ces deux types de réponse ont un point commun qui est fondamental pour mon argument général : elles permettent de dire que nous sommes justifiés à croire *que* nous avons des connaissances (perceptives, mathématiques, du libre arbitre) avant même de pouvoir déterminer *comment* nous savons.

La conclusion positive de cette section est donc la suivante : nous *avons* une justification (*prima facie*) à croire que notre expérience du libre arbitre constitue une source de connaissance (soit du fait de sa naturalité, soit du fait d'une expérience mooréenne). Soit $J_{pf}K_E LA_{index}$. Cette justification est antérieure à (et indépendante de) toute théorie sur la *manière* dont l'occurrence de cette expérience dépend du fait du libre arbitre.

Dans la mesure où notre justification est *prima facie*, il est tout à fait possible qu'elle soit réfutée si nous avons des raisons *positives* en sens contraire, c'est-à-dire si nous avons des raisons de croire que notre expérience *ne peut pas* constituer une source de connaissance. J'ai distingué plus haut (p.187) trois types de raisons de ce genre : des raisons empiriques sur le fonctionnement cérébral de notre expérience *E*, des raisons venant de l'analyse du contenu de *E* (trop lourd pour être *connu* de *quelque manière* que ce soit), enfin des raisons de penser que, *indépendamment* de toute analyse controversée de son contenu, *E* a des conditions de vérités trop lourdes pour être *connues*. Les premier et deuxième types de réfutations seront au cœur de la deuxième partie. Mais dans la mesure où cette première partie s'intéresse à la

²⁸⁴ C'est (Alston 1993, sec. 5.5) qui utilise l'expression de « significant self-support » pour désigner le type de support circulaire qui permet de ne pas ranger la perception exactement dans le même sac que les boules de cristal, en dépit de leur incapacité commune à être justifiées non-circulairement, et de leur capacité commune à être étayées par pur *bootstrapping*. Mais les deux réponses apportées par Alston dans cette section n'étaient pas satisfaisantes : son principal critère résidait dans les informations scientifiques sur la *manière* dont nous connaissons, et j'ai déjà développé pourquoi on ne pouvait pas en faire notre méta-justification de la perception.

justification « pré-scientifique » en faveur du libre arbitre, il me suffira de montrer qu'il n'y a pas de réfutation sérieuse du troisième type. C'est donc ce dernier type d'objection que je vais envisager à présent. L'ayant rejeté, nous pourrions conclure que, au moins en amont de toute analyse controversée du contenu de l'expérience *E*, il n'y a aucune bonne raison de rejeter la présomption de fiabilité en sa faveur.

III. L'objection de l'impression non étayable

L'objection de l'impression non étayable consiste à dire que l'expérience du Libre arbitre a des implications trop lourdes pour pouvoir être connues, et ce *en amont* de toute analyse controversée du contenu de cette expérience. Quelles sont ces implications lourdes et non controversées ? Une nouvelle fois, le cas emblématique sera le scénario de manipulation martienne que van Inwagen a imaginé, et que nous avons déjà rencontré à deux reprises. L'objection que nous allons considérer constitue la troisième et dernière utilisation possible de ce genre de scénarios. Mais il est important de bien distinguer cette nouvelle utilisation des deux premières. Dans l'argument de l'impression faillible, le scénario *MM* était présenté comme une possibilité logique qui empêchait la croyance *LA* d'être justifiée ($\sim JLA$). En réponse, j'ai soutenu qu'une croyance faillible (ou révisable) pouvait parfaitement être justifiée. Dans l'argument de l'impression lourde, le scénario *MM* était présenté comme une conséquence que le tenant du libre arbitre était engagé à croire et qu'il ne pouvait pas croire de manière justifiée ($\sim JMM$). En réponse, j'ai soutenu que nous pouvions tout à fait avoir une justification *prima facie* à croire que nous ne sommes pas manipulés par des martiens. Mais quand bien même la possibilité *MM* n'empêcherait pas la croyance (faillible) *LA* d'être justifiée, quand bien même nous pourrions avoir de manière justifiée la croyance $\sim MM$ elle-même, n'y a-t-il pas quelque chose d'étrange à dire que nous pourrions connaître que $\sim MM$? N'est-ce pas, à vrai dire, exclu *par hypothèse* pour un tel scénario ? Et si nous ne pouvons pas connaître $\sim MM$, cela ne constitue-t-il pas une réfutation de deuxième ordre pour la croyance $\sim MM$, au moins dans les cas où la question méta-épistémique est posée ? Enfin si la croyance $\sim MM$ peut être ainsi (fondationnellement) réfutée, n'avons-nous pas à présent (à cause de la clôture) une réfutation pour la croyance *LA* ?

Telle est l'objection de l'impression non étayable. Comme pour toutes les objections précédentes, je vais m'efforcer de montrer qu'elle correspond exactement à un argument sceptique – à mon sens, elle correspond même à l'argument sceptique le plus redoutable, que j'appellerai l'argument du double engagement. Mais la réponse sera plus rapide que pour les objections précédentes, car nous avons déjà présenté la quasi-totalité des outils conceptuels qui (s'ils sont bien employés) permettent d'y répondre.

III.1) L'objection et son principe de détectabilité externaliste

Pour bien comprendre cette objection, il faut apercevoir qu'elle procède fondamentalement en deux étapes, et utilise *à la fois* le principe de clôture *et* le principe de méta-cohérence. Et la très grande force de cette objection, c'est qu'elle peut tout à fait fonctionner en utilisant ces principes dans les versions « disjonctives » (les versions d'engagement de conséquence, ou d'engagement de méta-cohérence) dont j'ai montré plus haut qu'on ne pouvait pas raisonnablement les refuser. Pour plus de clarté, je vais cependant présenter d'abord l'objection en utilisant les versions directes ou non-disjonctives.

Outre ces deux principes, l'objection a fondamentalement une seule prémisse : la prémisse d'inconnaissabilité $\sim K \sim MM$. Sur quoi repose cette prémisse ? Tout simplement sur

l'intuition d'indétectabilité que nous avons déjà vue lors de l'objection de l'impression lourde (p.150). Cette intuition découle de la définition même des dispositifs martiens du scénario, en tant qu'ils sont « indétectables par toutes les techniques d'observation dont nous disposons » (van Inwagen 1983, 109). Nous avons vu au chapitre précédent que cette indétectabilité ne *pouvait* pas être tirée dans le sens d'une *injustifiabilité* ($\sim J \sim MM$) : bien que ces dispositifs soient indétectables, cela ne nous empêche pas d'avoir au moins une justification internaliste à rejeter leur existence. L'indétectabilité semble en effet être davantage du côté de la dépendance externaliste fait-croyance que du côté de la justification interne. Mais pour cette même raison, il semble tout à fait plausible de formuler l'intuition d'indétectabilité sous la forme d'un principe d'*inconnaissabilité* externaliste : $\sim K \sim MM$. Certes, ce principe d'inconnaissabilité reste en lui-même purement externaliste, mais dans la mesure où il peut être facilement *reconnu* par toute personne qui le considère sérieusement, il peut donner lieu à un problème externe-interne. Autrement dit, ce n'est pas le principe $\sim K \sim MM$ en lui-même qui sera à l'origine d'une réfutation fondationnelle, c'est plutôt le fait que l'on peut facilement être *justifié* à reconnaître sa vérité. C'est ce que j'appellerai la « prémisses d'inconnaissabilité justifiée » :

(IJ) $J \sim K \sim MM$

A partir de cette prémisses, l'argument utilise tout d'abord le principe de méta-cohérence : puisque la croyance $\sim MM$ est en présence d'un réfutateur fondationnel positif (puisque l'on *croit* de manière justifiée qu'elle ne peut constituer une connaissance), elle n'est pas justifiée. Pour cela, la version négative du principe de méta-cohérence est suffisante :

(MC-) $J \sim K p \Rightarrow \sim J p$

On déduit la conclusion intermédiaire :

(i) $\sim J \sim MM$

Jusque-là, on n'a rien dit concernant la justification de la croyance au libre arbitre. La méta-cohérence, à elle seule, remet seulement en cause la croyance $\sim MM$. Mais cette conclusion intermédiaire peut ensuite être utilisée comme point de départ d'un argument de clôture : puisque la croyance $\sim MM$ n'est pas justifiée, et qu'elle *doit* être justifiée pour qu'on puisse légitimement croire au libre arbitre (dont on *reconnaît* qu'il implique logiquement $\sim MM$), il s'ensuit que la croyance au libre arbitre elle-même n'est pas justifiée. C'est-à-dire qu'à partir de (i), on reprend exactement les étapes de l'argument de l'impression lourde de la page 152 :

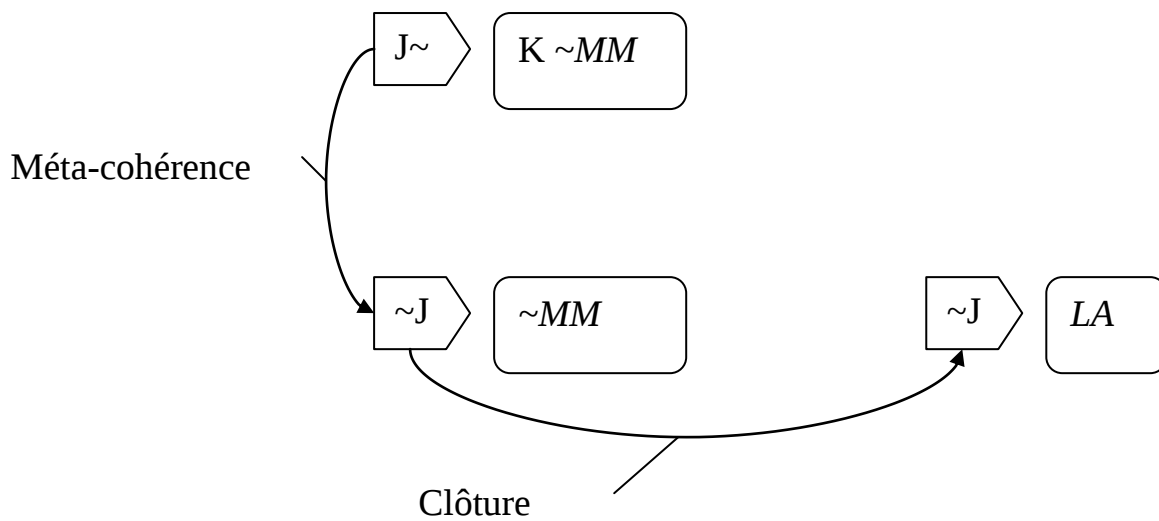
(i) $\sim J \sim MM$

(J-Clot.*) $[\sim J q \wedge J(p \Rightarrow q)] \Rightarrow \sim J p$

(2*) $J [LA \Rightarrow \sim MM]$

(C) $\sim J LA$

Si l'on se rappelle que les principes de méta-cohérence et de clôture correspondent tous deux à des principes de « transfert d'état épistémique », on peut représenter le fonctionnement de cet argument par un schéma de double transfert, allant de la position doxastique $\langle K \sim MM \rangle^S$ à la position $\langle \sim MM \rangle^S$, puis à la position $\langle LA \rangle^S$:



Je viens de présenter l'argument à partir des versions « simples » des principes de clôture et de méta-cohérence, celles qui ne posent pas de condition explicite de « conscience du problème » et ne font pas la distinction entre « absence de justification » et « réfutation *ultima facie* d'une justification *prima facie* ». Comme on l'a vu, ces versions « simples » sont trop fortes car il est probable qu'on puisse être justifié à croire que p sans croire explicitement toutes les conséquences q et la méta-croyance Kp , du moment qu'on n'a pas réfléchi explicitement à ces questions. Ainsi, si un agent irréflexif Irénée croit introspectivement au libre arbitre sans avoir jamais considéré le moindre scénario de type MM , ni *a fortiori* la possibilité de *connaître* $\sim MM$, il est probable qu'il sera parfaitement justifié dans sa croyance, contrairement à ce que semble conclure l'argument formel ci-dessus. Mais la défense d'Irénée est une bien maigre consolation, car les principes de clôture et de méta-cohérence restent certainement vrais dans des versions plus faibles, à savoir dans les versions « disjonctives » (ou les versions « d'engagement ») qui consistent à dire qu'*après réflexion*, un agent épistémique est *engagé* à assurer la conséquence et la méta-cohérence de ses croyances *sous peine de réfutation ultima facie*.

A partir des versions faibles de la clôture et de la méta-cohérence, l'argument permettrait donc de réfuter facilement la croyance introspectionniste de toute personne qui accepte de réfléchir :

Argument anti-introspection du double engagement

Supposons un agent épistémique S , sain d'esprit²⁸⁵ et qui possède, de par son expérience du libre arbitre E , une justification *prima facie* à croire au libre arbitre $J_{pf}LA$ et même une justification *prima facie* à croire qu'il le sait $J_{pf}KLA$.²⁸⁶ A t_0 , S croit LA sur la base de son expérience E , et comme il n'a aucune autre croyance occurrente sur la question, il est *ultima facie* justifié à t_0 . Mais à t_1 , S réfléchit au scénario MM :

²⁸⁵ La condition de « bonne santé mentale » de S est utile pour fonder les prémisses (IJ) et (2*) : en effet, si S était un agent épistémique totalement irrationnel, il pourrait *ne pas s'apercevoir* que $\sim MM$ est une conséquence logique de LA , ou ne pas s'apercevoir de l'indéfectibilité des dispositifs martiens. De telle sorte qu'il éviterait les réfuteurs doxastiques... mais simplement à cause d'une irrationalité caractérisée.

²⁸⁶ Que ce soit par expérience mooréenne, ou par naturalité.

- (2*) *S* reconnaîtra après réflexion qu'il peut déduire validement $\sim MM$ de *LA*, donc :
 (Clôture Engag.) Pour pouvoir continuer après réflexion à croire *LA* de manière justifiée (pour n'être pas réfuté par inconséquence), *S* est engagé à croire $\sim MM$ de manière justifiée.
 (MC Engag.) Pour pouvoir continuer après réflexion à croire $\sim MM$ de manière justifiée (pour n'être pas réfuté fondationnellement), *S* est engagé à croire $K\sim MM$ de manière justifiée.
 (IJ) *S* ne peut pas satisfaire l'engagement de croire $K\sim MM$ de manière justifiée.
 (C) Donc, *S* ne peut pas continuer après réflexion à croire *LA* de manière justifiée.

Dans cet argument, la prémisse (2*) est non controversée, et j'ai déjà expliqué les raisons pour lesquelles la clôture et la méta-cohérence ne pouvaient pas plausiblement être réfutées dans leurs versions faibles « d'engagement ». Donc la prémisse sur laquelle repose tout l'argument est la prémisse (IJ). Et cette prémisse est intuitivement très forte puisqu'il semble que les scénarios de type manipulation martienne excluent *ex hypothesi* que l'on puisse connaître $\sim MM$. Pour être précis et pouvoir vraiment bien évaluer l'argument, il faut se rappeler cependant que cette intuition n'est pas immédiate : ce qui est effectivement posé de manière stipulative dans le scénario, c'est que les dispositifs martiens (et donc leur absence) ne peuvent pas être *détectés*. Donc ce qui est irréfutable, c'est seulement :

(Indélect.) Nous ne pouvons pas détecter si $\sim MM$ est le cas.

Mais pour passer de l'indélectabilité à l'inconnaissabilité, il faut supposer un principe qui peut paraître très intuitif à première vue, mais qu'il convient d'explicitier :

(KDélect.) Si *S* ne peut pas détecter si *p* est le cas, alors *S* ne peut pas savoir que *p*.

Le « principe de détectabilité » représente déjà une *interprétation* (si intuitive soit-elle) de la condition anti-hasard (ou condition de dépendance fait-croyance) dont j'ai dit qu'elle était constitutive de la notion de « connaissance ». Compte tenu de la manière même dont je présente l'argument, le lecteur peut aisément deviner que je vais finalement mettre en question cette interprétation, et il faudra pour cela se mettre au clair sur la notion de « détection ».

Mais fondamentalement, ma démarche reste anti-sceptique, c'est-à-dire que mon intention principale n'est pas tant de repérer le point faible de cette nouvelle objection anti-introspection que de montrer qu'elle marche main dans la main avec un argument sceptique tout aussi redoutable.

III.2) L'argument sceptique du double engagement

Au point où nous en sommes, il devrait être évident que les scénarios sceptiques de type « malin génie » ou « cerveaux dans des cuves » se prêtent à un argument de double engagement tout aussi naturellement que le scénario de manipulation martienne. C'est-à-dire qu'on peut construire un argument *sceptique* du double engagement à partir de ces scénarios. A vrai dire, il me semble que, de tous les arguments sceptiques qui fonctionnent par scénario,²⁸⁷ l'argument du double engagement est la version la plus forte et sans doute même

²⁸⁷ Ce n'est pas le cas de tous les arguments sceptiques : les arguments de régression ou de méta-régression n'ont nullement besoin de faire appel à un scénario sceptique. Mais c'est tout de même le cas néanmoins de plusieurs arguments *différents* : nous avons déjà vu l'argument infaillibiliste et l'argument de clôture de la justification.

celle qui capture l'intention initiale des scénarios sceptiques. Quoi qu'il en soit, ce qui importe ici est seulement qu'il soit l'exact parallèle de l'argument anti-introspection.

Un argument sceptique du double engagement partira du fait que les hypothèses sceptiques sont des scénarios dans lesquels, *par stipulation*, nous ne pouvons pas détecter la manipulation épistémique dont nous sommes les victimes :

(Indélect.) Nous ne pouvons pas détecter si $\sim HS$ est le cas.

A partir de cette vérité incontestable, l'argument utilisera (explicitement ou implicitement) le « principe de détectabilité » (KDélect.) pour nous amener à accepter l'*inconnaissabilité* de la négation des hypothèses sceptiques :

(Inconnaiss.) Nous ne pouvons pas *savoir* que $\sim HS$.

Et dans la mesure où cette inconnaissabilité est *reconnue* par nous, nous sommes du même coup dans la situation « d'inconnaissabilité justifiée » qui peut générer un problème de méta-cohérence et nous empêcher de *croire* $\sim HS$ de manière justifiée :

(IJ) $J \sim K \sim HS$

(MC-) $J \sim Kp \Rightarrow \sim Jp$

(i) $\sim J \sim HS$

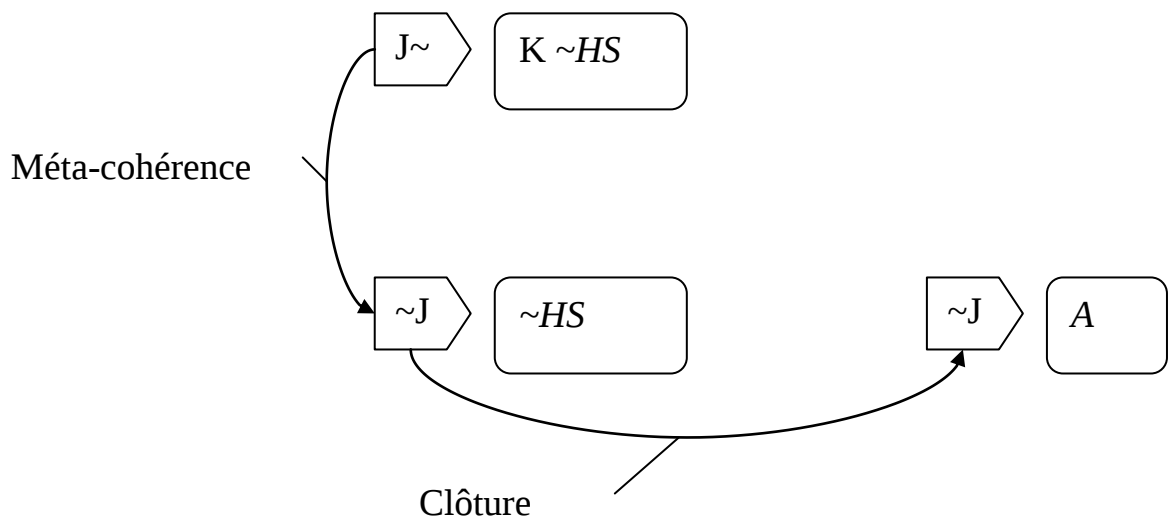
Et cette conclusion intermédiaire soulève ensuite un problème de clôture de la justification du fait que nos croyances perceptives anodines A impliquent logiquement $\sim HS$:

(J-Clot.*) $[\sim Jq \wedge J(p \Rightarrow q)] \Rightarrow \sim Jp$

(2*) $J [A \Rightarrow \sim HS]$

(C) $\sim J A$

On retrouve le schéma de réfutation en deux étapes :



Et pour éviter le problème des agents épistémiques qui « n'ont pas conscience du problème », il faut formuler l'argument sceptique comme une réfutation par *engagements* (de conséquence et de cohérence) d'un interlocuteur qui *prend conscience* du problème. Soit :

Argument sceptique du double engagement

Supposons un agent épistémique *S*, sain d'esprit et qui possède, de par son expérience perceptive, une justification *prima facie* à croire une proposition anodine *A* et même une justification *prima facie* à croire qu'il le sait $J_{pt}KA$. A *t*0, *S* croit *A* sur la base de son expérience perceptive, et comme il n'a aucune autre croyance occurrente sur la question, il est *ultima facie* justifié à *t*0. Mais à *t*1, *S* réfléchit au scénario *HS* :

(2*) *S* reconnaîtra après réflexion qu'il peut déduire validement $\sim HS$ de *A*, donc :

(Clôture Engag.) Pour pouvoir continuer après réflexion à croire *A* de manière justifiée, *S* est engagé à croire $\sim HS$ de manière justifiée.

(MC Engag.) Pour pouvoir continuer après réflexion à croire $\sim HS$ de manière justifiée, *S* est engagé à croire $K\sim HS$ de manière justifiée.

(IJ) *S* ne peut pas satisfaire l'engagement de croire $K\sim HS$ de manière justifiée.

(C) Donc, *S* ne peut pas continuer après réflexion à croire *A* de manière justifiée.

Il me semble que ce parallèle très exact suffit à montrer l'inefficacité dialectique de l'argument anti-introspection, qui pourra se satisfaire de toute réponse anti-sceptique à ce nouvel argument. Mais il sera utile de montrer plus explicitement la réponse la plus plausible pour pouvoir l'appliquer directement à l'introspectionnisme.

III.3) Solution : sensibilité des expériences et sûreté des conséquences

Comment peut-on répondre à ce nouveau type d'argument sceptique ? Les trois grands types de solutions sont évidemment : refus de la méta-cohérence, refus de la clôture, ou refus de l'inconnaissabilité (IJ).²⁸⁸ Pour les raisons que j'ai abondamment développées, il est extrêmement difficile de refuser la clôture et la méta-cohérence dans leurs versions faibles. La seule solution restante est donc de refuser (IJ) en affirmant $JK\sim HS$, c'est-à-dire : nous pouvons croire de manière *justifiée* que « notre croyance anti-sceptique $\sim HS$ constitue une *connaissance* (au sens externaliste) ». C'est donc ce type de réponse que je vais développer dans cette dernière section, et que j'appliquerai au cas du libre arbitre.

La grande difficulté de cette solution, c'est qu'elle semble attribuer un statut de *connaissance* (c'est-à-dire de dépendance externaliste fait-croyance) à la proposition la plus mal disposée à le recevoir ($\sim HS$). Si l'on pense aux croyances anodines *A*, il n'est pas très compliqué de les considérer comme dépendant externalistiquement de leur contenu (bien qu'elles n'en dépendent pas *infailliblement*) ; il n'est pas très compliqué non plus, pour la croyance $\sim HS$ elle-même, de lui reconnaître un statut de *justification internaliste* (puisque certaines de nos croyances peuvent constituer des raisons logiquement valides pour déduire $\sim HS$). Mais comment nos croyances $\sim HS$ seraient-elles *externalistiquement fiables* alors que, par hypothèse, elles ne peuvent pas dépendre du fait ainsi accepté ? Pour répondre à ce défi, il va falloir entrer dans une analyse plus précise de la condition de dépendance fait-croyance,

²⁸⁸ N'y a-t-il pas une quatrième solution fondée sur le refus de la prémisse (2*) ? Pas à ma connaissance, et je ne vois pas en quoi elle pourrait consister : si un agent épistémique n'est pas capable, après réflexion, de reconnaître qu'il peut déduire validement $\sim HS$ (je ne suis pas un cerveau dans une cuve) de *A* (j'ai deux mains), c'est qu'il a un grave dysfonctionnement interne, ce qui est exclu par hypothèse par l'idée d'un agent épistémique « sain ».

cette condition « anti-hasard » dont j'ai dit qu'elle était au cœur de tout concept de connaissance. Il existe en effet plusieurs interprétations possibles de cette condition anti-hasard, et du type de « fiabilité » qu'elle exige.²⁸⁹ La discussion portera sur deux grandes interprétations, l'une forte, l'autre faible. Si l'on adopte une lecture forte de la condition anti-hasard – la condition de « sensibilité » développée à l'origine par (Dretske 1971) et (Nozick 1981) – l'inconnaissabilité de $\sim HS$ est effectivement trivialement vraie. Pour éviter cette conclusion, j'adopterai donc finalement la stratégie de (Williamson 2000) et de (Pritchard 2005) qui défendent une interprétation faible de la condition anti-hasard (une interprétation en termes de « sûreté »). Cette analyse est, à mon sens, la seule manière d'éviter le scepticisme du double engagement. Pourtant, Dretske et Nozick ne sont pas de cet avis : bien qu'ils dénie à $\sim HS$ le statut de connaissance, ces auteurs pensent pouvoir éviter le scepticisme en refusant le principe de clôture de la *connaissance*. Y a-t-il donc, en fin de compte, une solution anti-sceptique défendable fondée sur le refus de la clôture ? La réponse est non : dans la mesure où l'argument du double engagement n'a pas recours au principe de clôture de la *connaissance*, mais seulement à (une certaine version de) la clôture de la *justification*, la réponse anti-clôture de Dretske et Nozick n'est tout simplement pas pertinente face à ce nouvel argument ; mais il faudra que j'expose ce point de manière plus développée, entre ma discussion de la condition de sensibilité (dont je maintiens qu'elle conduit au scepticisme du double engagement) et celle de la condition de sûreté (qui seule permet de l'éviter).

La grande force de l'argument du double engagement, c'est donc que les hypothèses sceptiques sont indétectables *ex hypothesi* et l'indétectabilité semble n'être qu'un autre nom du « hasard épistémique », cette absence de dépendance externaliste dont on a vu qu'elle est analytiquement exclue par toute notion de connaissance. Donc le raisonnement que j'ai explicité plus haut a toutes les apparences d'une analyse conceptuelle parfaitement innocente :

Sous-argument en faveur de l'inconnaissabilité de $\sim HS$

(Indélect.) Nous ne pouvons pas détecter si $\sim HS$ est le cas (*ex hypothesi*).

(KDélect.) Si *S* ne peut pas détecter si *p* est le cas, alors *S* ne peut pas savoir que *p*.

(Inconnaiss.) Donc nous ne pouvons pas savoir que $\sim HS$.

Si le concept de hasard épistémique n'est rien d'autre que l'indétectabilité, alors il est clair que les croyances $\sim HS$ sont *ex hypothesi* hasardeuses, et donc *ex hypothesi* inconnaissables. Ou pour le dire autrement, si les croyances $\sim HS$ sont « indépendantes des faits » dans le même sens qui empêche les cas de Gettier de compter comme connaissances, alors on sera bien obligé de considérer les premières comme aussi problématiques (épistémiquement) que les secondes. Encore faut-il donner une caractérisation plus précise de la condition anti-hasard pour s'assurer que c'est bien dans le même sens et pour les mêmes raisons que les croyances à la Gettier et les croyances $\sim HS$ sont « indépendantes » de leur contenu. Le sous-argument fonctionnera seulement si l'on parvient à montrer que le même sens « d'indépendance » (à savoir « l'indétectabilité ») est à la fois celui qui est incompatible avec la connaissance (le « hasard épistémique ») et celui qui est exclu *ex hypothesi* pour les scénarios sceptiques.

Or il faut reconnaître qu'une des premières intuitions (à la fois historiquement et naturellement) pour expliciter la condition anti-hasard offre justement un sens d'indétectabilité qui permettrait de faire fonctionner l'argument. D'où vient en effet l'intuition anti-hasard ? Des fameux « cas de Gettier », comme la croyance de Gwendolen que son fiancé s'appelle Ernest. Et qu'est-ce qui ne va pas au juste dans ce type de croyance ? Nous avons dit de manière générale qu'elle ne « dépendait en aucune manière » du *fait* qui est cru, et ainsi

²⁸⁹ On trouvera de très bonnes présentations, en français, des différentes théories fiabilistes dans (Engel 2007, chap. 1) et dans (Pouivet 2003).

n'était pas fiable. Mais peut-on être plus précis sur le type de dépendance ou de fiabilité qui est requis ? Une version historiquement importante du fiabilisme consiste à poser un réquisit de dépendance *causale* entre le fait et la croyance : c'est la « théorie causale de la connaissance » défendue par Alvin Goldman (1967) et David Armstrong (1973). Armstrong est célèbre, en particulier, pour l'analogie du thermomètre : une connaissance que *p* compte comme connaissance si et seulement si elle est produite par un mécanisme qui peut détecter le fait que *p* à la manière dont un thermomètre détecte la température. Si cette théorie de la condition anti-hasard était la bonne, il est clair que notre croyance $\sim HS$ serait *ex hypothesi* incapable de compter comme connaissance – car un tel type de « détection » est clairement exclu. Mais le fiabilisme causal ne représente pas, en fait, une grande menace pour la croyance $\sim HS$ car il est clair qu'il ne peut pas constituer une théorie *générale* de la connaissance : comme le rappelle (Nozick 1981, 170) les cas de connaissance mathématique ou de connaissance morale ne sauraient absolument pas rentrer dans le cadre d'une détection strictement causale des faits connus. Or si la contrainte de dépendance causale ne s'impose pas à *toute* forme de connaissance, elle ne peut pas être la condition anti-hasard que nous cherchons ici.

Cependant, Nozick a proposé une autre analyse de la condition anti-hasard, qui quant à elle peut à première vue se généraliser, mais n'en condamne pas moins les croyances $\sim HS$. Pour expliquer ce qui ne va pas dans les cas de Gettier, Nozick (1981, 172–173) – à la suite de (Dretske 1971) – a proposé une analyse « contrefactuelle » du hasard épistémique.²⁹⁰ Le problème de la croyance de Gwendolen (si l'on lui applique le diagnostic de Nozick), c'est que :

(G1) Si son fiancé ne s'appelait pas Ernest, Gwendolen croirait tout de même qu'il s'appelle Ernest.²⁹¹

Le conditionnel dont il s'agit ici n'est pas un conditionnel strict (ou une implication logique) : de « son fiancé ne s'appelle pas Ernest », on ne peut rien déduire *logiquement* quant aux croyances de Gwendolen. Il s'agit d'un conditionnel « subjonctif » ou « contrefactuel », qui d'après la théorie des contrefactuels de Stalnaker-Lewis, consiste à se demander ce qui se passerait « dans le monde possible le plus proche du monde actuel (ou les mondes possibles les plus proches) dans lequel l'antécédent est vrai ».²⁹² Ce que nous dit le conditionnel ($\sim$ Sens.), c'est donc que dans les situations les plus proches du monde actuel où son fiancé ne s'appelle pas Ernest – par exemple dans situations dans lesquelles le général Montcrieff aurait choisi de donner à son fils un autre prénom de baptême que le sien – la croyance de

²⁹⁰ Je suivrai dans mon exposé l'analyse de Nozick plutôt que celle de Dretske, parce qu'elle possède le grand avantage de présenter l'intuition de « l'analyse conditionnelle » dans une version plus pure et plus simple (au moins dans sa première version). J'aurai l'occasion de mentionner (en notes) les ajouts techniques de l'analyse de Dretske, mais ils ne me semblent pas pertinents pour les problèmes qui nous occupent.

²⁹¹ L'exemple pris par (Nozick 1981, 173) est un des exemples de l'article de Gettier lui-même, dans lequel une proposition générale vraie (« quelqu'un dans mon bureau possède une Ford ») est déduite d'une prémisse fausse (« Bob, qui est dans mon bureau, possède une Ford » – en fait c'est Bill qui est dans mon bureau et qui, à mon insu, possède une Ford).

²⁹² Il n'est pas possible de donner ici une idée fidèle du continent que représentent les recherches sur la théorie des conditionnels et spécialement sur les conditionnels contrefactuels. Je peux seulement renvoyer aux textes séminaux de (Stalnaker 1968) et (Lewis 1973), ainsi qu'à l'excellente synthèse de (Bennett 2003). Pour le reste, je suivrai l'usage en vigueur dans la littérature analytique, qui consiste à présupposer chez le lecteur une familiarité minimale avec ce genre de langage. Je noterai simplement que l'alternative entre « le monde possible le plus proche » ou « les mondes possibles les plus proches » correspond à un débat entre Lewis et Stalnaker : Stalnaker a initialement posé une condition d'*unicité* du monde le plus proche, dont Lewis a montré qu'elle pouvait ne pas être satisfaite, et qu'il valait donc mieux considérer l'ensemble *des* mondes maximalement proches du monde actuel. Mais Stalnaker a refusé cette révision lewisienne.

Gwendolen resterait inchangée, puisque sa croyance repose uniquement sur le *mensonge* de son fiancé qui lui a dit s'appeler Ernest. De ce diagnostic négatif, on peut tirer une leçon positive : on peut déduire la condition que la croyance de Gwendolen *aurait dû* satisfaire pour constituer une connaissance. Il aurait fallu que le conditionnel (contrefactuel) suivant fût vrai :

(G2) Si son fiancé ne s'appelait pas Ernest, dans ce cas Gwendolen ne croirait pas qu'il s'appelle Ernest.

On peut facilement généraliser ce diagnostic pour obtenir une analyse de la condition anti-hasard. La condition générale qu'on obtient alors (la « condition (3) » de Nozick) est aujourd'hui communément désignée comme « condition de sensibilité » (*sensitivity*).

(Condition de Sensibilité)

La croyance de *S* que *p* est une connaissance seulement si le conditionnel suivant est vrai :

(3) Si *p* n'était pas vraie, *S* ne croirait pas que *p*.

Pour être complet, il faut noter que cette condition de sensibilité n'est pas suffisante, d'après Nozick, pour qu'une croyance vraie devienne une connaissance. Elle ne constitue donc pas *à elle seule* la condition anti-hasard. Pour qu'une croyance vraie évite le hasard épistémique, il faut qu'elle satisfasse une deuxième condition, qu'on peut apercevoir en remarquant un deuxième défaut de la croyance de Gwendolen :

(G3) Si son fiancé s'appelait Ernest de naissance mais avait choisi un autre pseudonyme qu'Ernest pour se présenter dans la société londonienne, Gwendolen ne croirait pas qu'il s'appelle Ernest.

Ce conditionnel fait apparaître une nouvelle manière dont la croyance de Gwendolen est hasardeuse.²⁹³ Intuitivement, le conditionnel (G1) révèle que la croyance de Gwendolen est insensible à la *fausseté* de la proposition « il s'appelle Ernest », tandis que le conditionnel (G3) révèle qu'elle est insensible à la *vérité* de cette proposition. Ou pour le dire autrement, la croyance de Gwendolen ne « suit » pas les changements de valeur de vérité de cette proposition. Pour qu'une croyance vraie constitue une connaissance, il faudra donc que son occurrence « suivre à la trace » son contenu, que celui-ci aille vers le vrai ou vers le faux. Cette *double* condition, Nozick l'a baptisée « condition de traçage » (*tracking condition*) :

(Condition de Traçage)

La croyance vraie de *S* que *p* est une connaissance si et seulement si :

(3) Si *p* n'était pas vraie, *S* ne croirait pas que *p*.

(4) Si *p* était vraie (dans des conditions contrefactuelles proches), *S* croirait que *p*.

Ces deux conditions étant, d'après Nozick, *suffisantes* pour changer une croyance vraie en connaissance, elles constituent conjointement son analyse du hasard épistémique.

Le problème de ce diagnostic, c'est qu'il fournit exactement ce dont le sceptique a besoin pour établir l'inconnaissabilité de $\sim HS$. En effet, même si la « sensibilité » est une

²⁹³ Pour susciter l'intuition que (G3) révèle une forme de hasard épistémique, on peut préciser le scénario de la manière suivante : Jack (qui ne sait pas qu'il s'appelle Ernest) a choisi son pseudonyme au hasard, en ouvrant un journal et en prenant le premier prénom qui lui tomberait sous les yeux. Il est tombé sur le prénom « Ernest », mais il avait plus de chances de tomber, par exemple, sur le prénom « Robert ». Et dans ce cas, Gwendolen aurait cru qu'il s'appelle Robert.

condition plus faible que la stricte « détection causale » d'Armstrong et Goldman, il est clair qu'elle reste, en un certain sens, une condition de « détectabilité », une condition que la croyance $\sim HS$ ne peut pas satisfaire. Dans le sens qui vient d'être défini, la croyance $\sim HS$ n'est *pas* sensible puisque le conditionnel suivant est vrai :

(~ 3) Si $\sim HS$ n'était pas vraie (*i.e.* si le scénario sceptique *était* vrai), S croirait tout de même que $\sim HS$.

Si donc la condition de sensibilité (ou la double condition de traçage) était le bon diagnostic des problèmes de Gettier, cela voudrait dire que la croyance $\sim HS$ est forcément, par hypothèse, problématique *dans le sens même* où les cas de Gettier sont problématiques. Or comme ces cas constituent l'archétype ou l'étalon en matière de hasard épistémique, il n'y aurait alors que deux possibilités : soit rejeter totalement la condition anti-hasard (et appeler « connaissances » les cas de Gettier !), soit reconnaître que $\sim HS$ ne peut pas constituer une connaissance. Mais rejeter absolument la condition anti-hasard est inacceptable, donc cela voudrait dire que $\sim HS$ est effectivement analytiquement inconnaissable. Et donc, compte tenu du caractère difficilement refusable de l'engagement de conséquence et de l'engagement de cohérence, cela voudrait dire qu'on est bel est bien amené à rétracter nos affirmations (et nos croyances) anodines A lorsqu'on prend conscience qu'elles nous engagent à accepter $\sim HS$, ce qui nous engage à accepter $K\sim HS$, ce que nous ne pouvons accepter. Puisqu'elle oblige à dire que $\sim HS$ n'est pas une connaissance, l'interprétation de la condition anti-hasard en termes de sensibilité ou de traçage amène donc au scepticisme. Pour éviter le scepticisme, il faudra donc envisager une autre interprétation, plus faible. Mais avant cela, il faut répondre à une difficulté que peut soulever ma démarche : je présente ici la condition de sensibilité comme menant au scepticisme (parce qu'elle oblige à reconnaître $\sim K\sim HS$) alors que, dans la littérature, ses défenseurs la présentent au contraire comme une excellente *réponse* contre le scepticisme. Comment cela est-il possible ?

Pour Dretske et Nozick, l'analyse de la connaissance en termes de sensibilité (ou de traçage) constitue une bonne réponse au scepticisme parce que i) elle est intuitive en elle-même (pour les cas de Gettier qui n'ont rien à voir avec le scepticisme), et que ii) si elle est vraie, alors le principe de clôture est faux, et par conséquent, l'argument sceptique ne marche pas. Cette analyse présente donc l'avantage d'offrir non seulement une réponse au sceptique, mais une réponse *indépendamment motivée*.²⁹⁴ Dans notre discussion, cela semble vouloir dire qu'il y aurait, en fait, une manière convaincante de refuser le principe de clôture, contrairement aux arguments que j'ai présentés plus haut en faveur de la clôture. En réalité, quand on y regarde de plus près, on s'aperçoit que le principe de clôture que permet de rejeter la condition de sensibilité n'est *pas* celui sur lequel repose l'argument du double engagement : la sensibilité amène éventuellement à refuser la clôture de la *connaissance*, tandis que l'argument que nous discutons ici repose sur un principe de clôture (disjonctive) de la *justification*. On doit donc s'attendre à ce que la dénégation de « la clôture » (par Dretske et Nozick) ne résolve pas le problème soulevé par le double engagement. Mais voyons cela plus en détail.

Pour montrer que la condition de sensibilité (qui, intrinsèquement, est assez intuitive) remet en cause la clôture épistémique, on peut partir de l'exemple célèbre de Dretske :

²⁹⁴ Cet intérêt de l'analyse contrefactuelle dans la dialectique anti-sceptique est exposé dans (Dretske 2005, 19).

Cas du zèbre

Vous emmenez votre fils au zoo, vous voyez plusieurs zèbres, et comme votre fils vous interroge, vous répondez que ce sont des zèbres. Savez-vous que ce sont des zèbres ? [...] le fait que quelque chose soit un zèbre implique que ce n'est pas une mule et, en particulier, que ce n'est pas une mule habilement déguisée par les autorités du zoo pour ressembler à un zèbre. Savez-vous que ces animaux ne sont pas des mules habilement déguisées par les autorités de la ville pour ressembler à des zèbres ? (Dretske 1970, 1015–1016)

Dretske pose ici deux questions de « connaissance » : l'une pour la proposition Z (ceci est un zèbre), l'autre pour la proposition $\sim MD$ (ceci n'est pas une mule habilement déguisée). Essayons d'appliquer la condition de sensibilité. Il est clair qu'on a :

(Sens Z) Si ceci n'était pas un zèbre, je ne croirais pas que c'est un zèbre.

Dans les mondes possibles les plus proches, si ceci n'était pas un zèbre, je le verrais bien, soit parce que ça ne ressemblerait pas du tout à un zèbre, soit parce que le zoo aurait mis une autre affiche sur l'enclos, etc. Autrement dit, la croyance Z trace la vérité de Z dans les mondes possibles proches. Elle peut donc constituer une connaissance. Mais considérons à présent la croyance $\sim MD$. Satisfait-elle la condition de sensibilité ? Il semble clair que non :

($\sim$ Sens $\sim MD$) S'il n'était pas le cas que $\sim MD$ (si ceci *était* une mule habilement déguisée en zèbre), je croirais tout de même que ceci est un zèbre (et non pas que ceci est une mule habilement déguisée).

Par conséquent, si l'on considère que la sensibilité est une condition nécessaire de la connaissance, alors on doit accepter à la fois KZ et $\sim K\sim MD$ alors même que $\sim MD$ est une conséquence logique (reconnue) de Z . Autrement dit : si l'on prend la condition de sensibilité comme meilleure analyse de la condition anti-hasard, on a alors une réfutation de la clôture qui est motivée indépendamment de la question sceptique.

Mais quel principe de clôture a-t-on ainsi réfuté ? Clairement, le principe qui échoue est le principe de clôture *de la connaissance* :

(K-Clot.) $[Kp \wedge K(p \Rightarrow q)] \Rightarrow Kq$

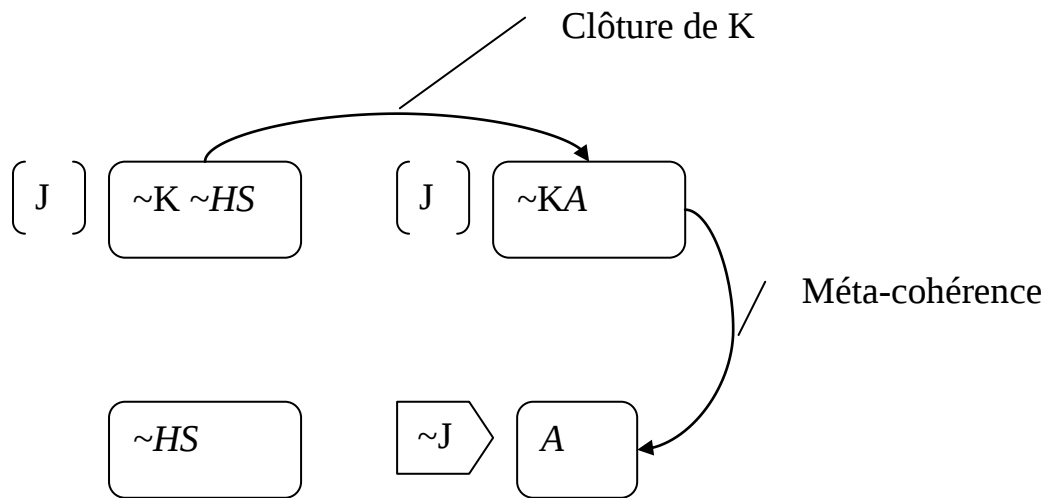
Mais à aucun moment ce principe n'a été utilisé ni présupposé dans l'argument du double engagement que j'ai présenté plus haut. Le réfuter ne devrait donc pas nous aider beaucoup. Pourquoi Dretske et Nozick présentent-ils alors cette réfutation de la clôture comme une réponse au scepticisme ? C'est qu'elle permet de répondre à deux *autres* types de scepticisme. Tout d'abord, elle permet de répondre au sceptique *académique*, qui chercherait à établir positivement que *nous n'avons pas de connaissance perceptive* : $\sim KA$. Un sceptique académique pourrait en effet appuyer sa conclusion sur l'indétectabilité des hypothèses sceptiques *plus* le principe de clôture épistémique. Notez que la conclusion recherchée par le sceptique académique est indépendante de la question de justification : il pourrait (s'il refuse le principe de méta-cohérence) continuer à dire que nous sommes parfaitement justifiés à croire à nos perceptions bien qu'elles ne constituent pas des connaissances. Mais un sceptique plus radical pourrait utiliser la conclusion du sceptique académique, *plus* le principe de méta-cohérence, pour établir que nous ne sommes pas justifiés à croire aux données des sens : $\sim JA$. Autrement dit, un tel sceptique opérerait un nouvel argument à deux étapes. Première étape :

(Inconn.) $\sim K \sim HS$
 (K-Clot.) $[Kp \wedge K(p \Rightarrow q)] \Rightarrow Kq$
 (ii) $\sim KA$

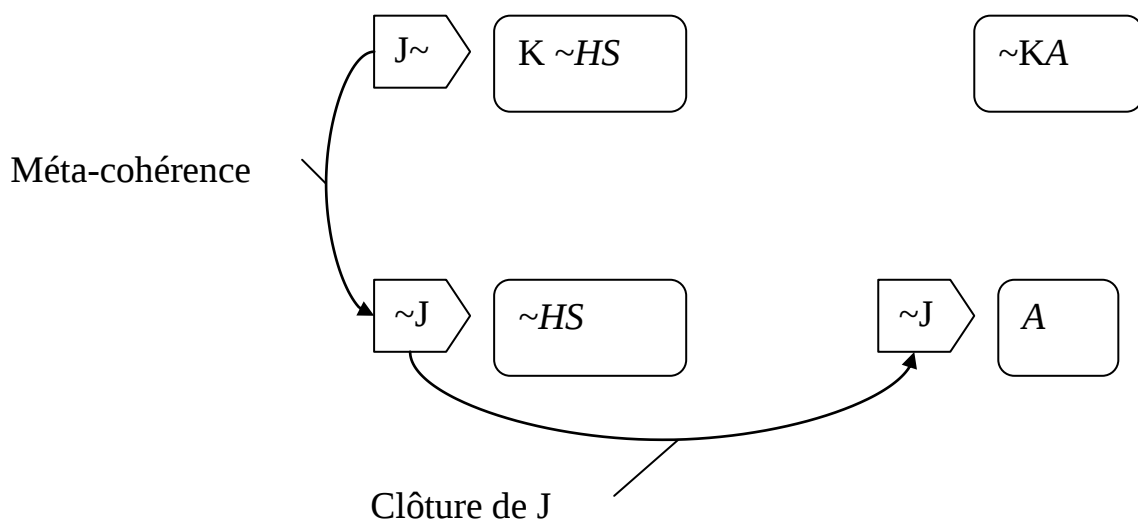
A l'issue de ce raisonnement, le sceptique amènerait à *prendre conscience* de la justification qu'il donne à sa conclusion pour passer à la deuxième étape :

(iii) $J \sim KA$
 (MC-) $J \sim Kp \Rightarrow \sim Jp$
 (C) $\sim JA$

Ce nouvel argument peut également être représenté par le schéma suivant.



Comparons à présent avec le schéma d'argument du double engagement :



En comparant ces deux schémas, on peut voir clairement l'intérêt et les limites d'un refus de la clôture de la connaissance. En refusant la clôture de la connaissance, on évite d'admettre la proposition (et la croyance justifiée) $\sim KA$, et on bloque ainsi *un* chemin possible vers la

conclusion absolument inacceptable que nos croyances perceptives sont tout bonnement injustifiées $\sim JA$. Mais cette stratégie a fort peu d'intérêt dialectique dans la mesure où le sceptique peut emprunter une *autre* voie (celle de $\sim J \sim HS$) pour arriver à la même conclusion au moyen d'un principe de clôture de la *justification* – un principe qui, quant à lui, est extrêmement difficile à refuser. Dans la mesure où ils acceptent explicitement $\sim K \sim HS$ (et donc, peut-on supposer, $J \sim K \sim HS$), Dretske et Nozick sont donc menacés directement par l'argument du double engagement, et le refus de la clôture de la connaissance n'aide aucunement à résoudre ce problème. D'une manière plus générale, toute analyse de la condition anti-hasard qui rend inconnaisables les propositions $\sim HS$ mènera au scepticisme du double engagement, quelles que soient par ailleurs ses prédictions sur la clôture de la connaissance. Pour éviter le scepticisme, il va donc falloir mettre en évidence une autre analyse du hasard épistémique qui permet de dire que les croyances $\sim HS$ ne sont pas hasardeuses ; nous verrons alors si une telle analyse valide ou non la clôture de la connaissance. Ce qu'on peut dire en tout cas avant toute analyse, c'est que la clôture de la connaissance semble au moins intuitive à première vue : si je sais que p et que j'en déduis (avec compétence) que q , n'est-il pas intuitif que ma croyance à la fois vraie, justifiée, et produite par un processus infaillible (la déduction) sera une connaissance ? Donc si une analyse permet de préserver la clôture de la connaissance, ce sera certainement un avantage et non inconvénient pour cette analyse. Dretske lui-même le reconnaît :

[L]a clôture [de la connaissance] a l'air d'un principe extrêmement plausible. Toutes choses égales par ailleurs, nous devrions donc la préserver. (Dretske 2005, 18)

Or, comme on va le voir, l'analyse de la connaissance en termes de « sûreté », outre qu'elle permet de répondre à l'argument sceptique en permettant une connaissance $K \sim HS$, a aussi l'avantage de préserver la clôture de la connaissance.

Voyons donc à présent en quoi consiste la notion de « sûreté » épistémique, cette analyse de la condition anti-hasard qui permet d'éviter le scepticisme du double engagement. Autrement dit, c'est une notion qui permet de dire que ce n'est pas un pur hasard si nos croyances anti-sceptiques $\sim HS$ correspondent aux faits. Comment cela est-il possible ? Informellement, l'intuition est la suivante : on ne peut pas dire que nos croyances $\sim HS$ sont vraies « par pur hasard » tout simplement parce qu'on n'aurait pas pu facilement se tromper à leur sujet. De là découle la formulation de Williamson :

Condition de Sûreté

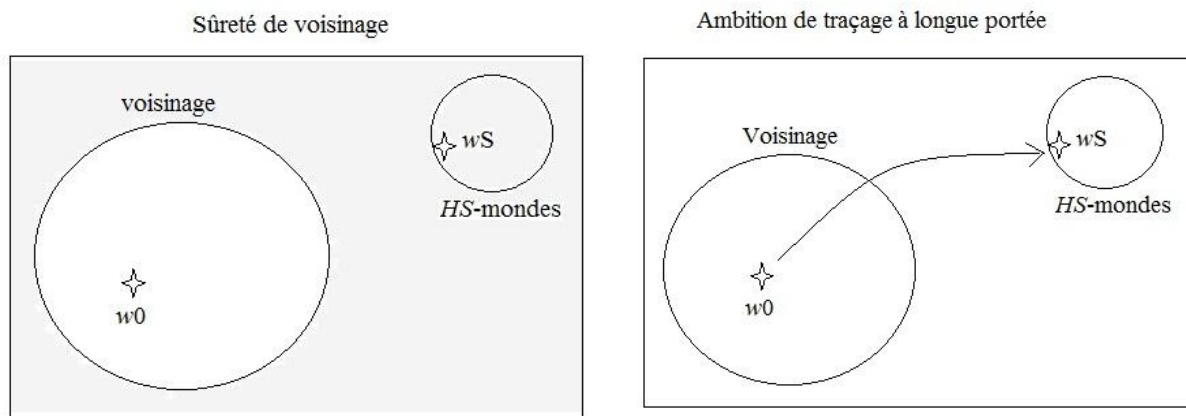
Si l'on sait, on ne pourrait pas facilement s'être trompé dans un cas semblable.

(Williamson 2000, 147)

Mais cette définition mérite quelques développements plus formels, pour voir en quoi elle se distingue exactement des autres analyses, et en quoi elle permet de résoudre le problème. En effet, même si l'on aperçoit que les croyances $\sim HS$ n'auraient pas pu facilement être des erreurs, on peut conserver l'intuition problématique qu'il n'y a aucune dépendance modale entre le fait et la croyance.

La différence fondamentale entre « sensibilité » et « sûreté », celle qui permet à la condition de « sûreté » de répondre au scepticisme, tient à la notion de « voisinage modal ». C'est cette notion qui amène (Vogel 2000) à distinguer entre « fiabilisme contrefactuel » (la sensibilité) et « fiabilisme de voisinage » (la sûreté). L'idée est la suivante. Si notre croyance $\sim HS$ n'est pas *sensible*, c'est parce que nous la conserverions « dans le monde possible le plus proche où $\sim HS$ est fausse ». En effet, le monde possible « le plus proche » où $\sim HS$ est fausse,

appelons-le wS , est un monde possible où nous sommes manipulés et où nous croyons ne pas l'être. Mais ce monde possible wS (cette possibilité logique d'erreur) permet-il vraiment de dire que, du coup, notre croyance $\sim HS$ n'est vraie que « par pur hasard » ? Cela semble tout à fait exagéré, tout simplement parce que wS est un monde extraordinairement lointain (une possibilité purement théorique) : wS appartient à l'univers des mondes possibles, mais il n'appartient pas pour ainsi dire à notre galaxie, ou à notre voisinage modal. Certes wS est le monde « le plus proche » où $\sim HS$ est fausse, mais cela n'empêche pas qu'il soit extrêmement lointain, beaucoup trop lointain pour être épistémiquement pertinent. Le cœur de l'intuition de « sûreté », c'est qu'il n'est pas légitime d'aller chercher si loin dans l'espace modal pour voir si notre croyance trace les faits : si elle les trace jusqu'aux confins de notre voisinage modal, c'est largement suffisant pour pouvoir dire qu'elle n'est pas un « pur hasard ». Or si l'on restreint notre attention à notre voisinage modal, notre croyance $\sim HS$ est vraie dans *tous* les mondes de ce voisinage. Dans les limites de notre voisinage, notre croyance est bien plus que sensible aux faits : elle est « infaillible ». On peut résumer cette différence fondamentale en disant que la condition de sensibilité est une exigence de dépendance fait-croyance « à longue portée », tandis que la condition de sûreté est seulement une exigence de dépendance « de voisinage ».



A partir de cette notion de voisinage, on peut donner une formulation plus explicite de la condition de sûreté, celle défendue par (Pritchard 2005, 163) :

Condition de Sûreté (bis)

La croyance de S que p est sûre si et seulement si dans tous (ou presque tous) les mondes possibles du voisinage dans lesquels S croit que p (sur la même base épistémique que dans le monde actuel), p est vraie.²⁹⁵

Il est intéressant de noter que la notion de « voisinage épistémique » n'est pas propre aux défenseurs de la condition de « sûreté » : en effet, Nozick lui-même en a besoin d'une manière ou d'une autre pour formuler sa « condition (4) », c'est-à-dire la condition de traçage du *vrai* (par opposition au traçage du *faux*) :

(4) Si p était vraie (dans des conditions contrefactuelles proches), S croirait que p .

La parenthèse « dans des conditions contrefactuelles proches » consiste justement à voir dans les mondes possibles *du voisinage* si la croyance vraie que p continue d'être vraie lorsque p

²⁹⁵ Pour les détails de cette analyse, notamment les restrictions « presque tous », ou « sur la même base épistémique », je dois renvoyer aux discussions de (Pritchard 2005, chap. 6).

reste vraie.²⁹⁶ Mais Nozick n'introduit cette « restriction au voisinage » *que* pour la condition de traçage du vrai (4) ; les analyses en termes de sûreté, quant à elles, sont plus cohérentes : elles restreignent notre attention au voisinage *à la fois* pour la dépendance au vrai et pour la dépendance au faux.²⁹⁷ La notion de voisinage modal permet donc de distinguer entre théories de la sûreté et théories de la sensibilité.

Pour bien comprendre la condition de sûreté, il faut enfin la comparer avec la théorie (ou plutôt *les* théories) des « alternatives pertinentes ». Comme on vient de le voir, la notion de sûreté repose sur l'intuition que certains mondes possibles – et en particulier les mondes correspondant aux scénarios sceptiques – sont trop éloignés pour être « épistémiquement pertinents ». En ce sens, elle entre dans la catégorie des réponses anti-sceptiques « des alternatives pertinentes », telle que les caractérise Alvin Goldman :

Une réponse que l'adversaire du sceptique peut faire, c'est que ces hypothèses sont des hypothèses « purement gratuites » [*idle*], et qu'une personne peut savoir une proposition même s'il y a des alternatives « purement gratuites » qu'il ne peut éliminer. (Goldman 1976, 775)

En ce sens *toute* épistémologie qui refuse l'infailibilisme est une théorie des alternatives pertinentes puisqu'en refusant l'infailibilisme on accepte que *certaines* possibilités d'erreur (du type *HS*) n'empêchent pas (ne sont pas « pertinentes » pour empêcher) de connaître les propositions anodines *A*. Mais il y a ensuite de grandes divergences sur deux questions fondamentales : qu'est-ce qui *fait* que les alternatives *HS* sont non pertinentes ? leur non-pertinence permet-elle de connaître la proposition $\sim HS$ ou seulement de connaître *A* ? Ainsi que l'a clairement mis en évidence Jim Pryor (2001, 97–98), il y a eu deux périodes différentes dans l'utilisation du vocabulaire des alternatives pertinentes. Les auteurs qui ont initialement introduit ce vocabulaire – et notamment (Dretske 1970, 1021–1022) – considéraient que la non-pertinence de *HS* permettait de connaître *A sans avoir à connaître $\sim HS$* (sans avoir à « éliminer » l'alternative non-pertinente *HS*). C'est donc une solution qui refuse la clôture. Quant à savoir ce qui *fait* que *HS* est une alternative non-pertinente à *A*, c'est que « elle ne se serait pas produite si $\sim A$ était le cas ». Autrement dit, la classification en alternatives pertinentes ou non « sera forcément développée en termes de contrefactuels » (Dretske 1970, n. 6). Mais depuis, le vocabulaire des alternatives pertinentes a été repris et majoritairement utilisé par des auteurs *contextualistes* qui acceptent la clôture épistémique.

²⁹⁶ Cette parenthèse n'est pas une erreur malencontreuse de Nozick : elle est indispensable pour pouvoir formuler une condition de traçage du vrai. En effet, si on l'enlève, on obtient le conditionnel suivant : « Si *p* était vraie, *S* croirait que *p* ». Mais ce conditionnel ne peut pas constituer une contrainte intéressante pour qu'une croyance vraie constitue une connaissance : ce conditionnel est *trivialement vrai* pour *toute* croyance vraie que *p* (d'après la logique des conditionnels contrefactuels : pour toute croyance vraie (de *S*) que *p*, le *p*-monde le plus proche est le monde actuel lui-même, et donc on a bien « dans le *p*-monde le plus proche, *S* croit que *p* »). Si l'on veut une condition modale non triviale, il faut donc « aller voir dans d'autres mondes »... mais on ne peut pas non plus exiger que *S* croie que *p* dans *tous* les *p*-mondes. D'où la nécessité d'avoir recours au voisinage de mondes pour formuler une condition de traçage du vrai à la Nozick.

²⁹⁷ On peut noter que, même pour la dépendance au vrai, la condition (4) de Nozick n'est pas *exactement* équivalente à la condition de sûreté, ainsi que le souligne (Engel 2007, 105–106) en prenant l'exemple de Sacha et Olga : Sacha croit (véridiquement) qu'Olga est mariée parce qu'il a remarqué son alliance ; mais Olga ne porte que très rarement son alliance et ne l'a mise aujourd'hui que par pur hasard ; donc dans des mondes très proches où Olga reste mariée, elle ne porte pas son alliance et Sacha ne produit pas la croyance qu'elle est mariée. A cause de ce conditionnel, la croyance de Sacha ne satisfait pas la condition (4) de Nozick : $p \rightarrow Bp$. Mais si l'on complète le scénario en disant qu'en l'absence d'alliance au doigt d'Olga, Sacha n'aurait même pas considéré la question, ou qu'il se serait abstenu de tout jugement, on garde tout de même l'idée qu'il ne se serait pas *trompé* dans ces éventualités proches, et donc sa croyance reste *sûre*. On conserve le conditionnel : $p \rightarrow \sim B \sim p$. On pourrait décrire les choses ainsi : la condition (4) de Nozick exige de tracer *positivement* le vrai, alors que la sûreté exige seulement de le tracer *négativement*.

Pour (DeRose 1992) et (Lewis 1996), ce qui rend une alternative pertinente ou non tient essentiellement au *contexte* dans lequel on se trouve (les hypothèses sceptiques ne sont certainement pas pertinentes dans les contextes de la vie courante où nous nous attribuons des connaissances anodines). Et leur non-pertinence permet non seulement de dire que nous pouvons connaître *A*, mais aussi que nous pouvons connaître leurs négations $\sim HS$, tout du moins dans ces contextes où elles sont non-pertinentes. Compte tenu des discussions qui précèdent, il est clair qu'aucune de ces deux « théories des alternatives pertinentes » ne peut nous satisfaire ici : la théorie de Drestke parce que la notion de dépendance contrefactuelle mène au scepticisme du double engagement, la théorie de DeRose ou Lewis parce qu'elle repose sur le contextualisme. La notion de sûreté épistémique représente donc un troisième type de théorie des alternatives pertinentes, qui empruntent des aspects aux deux premières. Comme dans la théorie de Lewis, la non-pertinence d'une alternative permet de connaître sa négation ($\sim HS$) et pas seulement de connaître certaines propositions (*A*) qui impliquent sa négation. En ce sens, l'analyse de la connaissance en termes de sûreté permet de préserver la clôture *y compris pour la connaissance*. Mais comme dans la théorie de Drestke, ce qui rendra une alternative pertinente ne tient pas à un contexte variable, mais à la « situation modale » de cette possibilité – une situation modale qu'on peut dire « objective ». Ainsi, la possibilité *HS* restera toujours une possibilité objectivement éloignée de nous, éloignée d'un point de vue externaliste ; et donc il nous sera toujours possible de *connaître* $\sim HS$ (au sens externaliste) quand bien même nous aurions des croyances ou des doutes sceptiques qui nous empêcheraient d'être internalistiquement justifiés à le croire.

Pour résumer, la notion de sûreté épistémique consiste à dire qu'une certaine croyance n'est pas épistémiquement hasardeuse lorsque son profil modal objectif nous met à l'abri des possibilités d'erreur. Et n'étant pas épistémiquement hasardeuse, elle peut compter comme connaissance, même si elle ne peut pas « détecter » (causalement ou contrefactuellement) son contenu. La détectabilité n'est donc pas une condition nécessaire à la connaissance. Mais est-il *suffisant* qu'une croyance soit sûre (vraie dans son voisinage modal) pour qu'elle compte comme connaissance ? En montrant que $\sim HS$ *pouvait* compter comme connaissance (parce qu'elle est sûre et qu'elle n'a pas à être sensible), on n'a pas encore montré qu'elle comptait effectivement comme connaissance. Une dernière remarque s'impose donc pour expliquer comment $\sim HS$ peut compter comme connaissance.

Ce qui fait que nos croyances $\sim HS$ comptent comme connaissances, ce n'est sans doute pas *uniquement* qu'elles sont des croyances sûres. En effet, certaines croyances sûres ne pourront sans doute pas compter comme connaissances à cause de la manière dont elles sont formées. Par exemple, imaginons un fou dont les facultés épistémiques basiques sont profondément perturbées mais *pas* sa faculté de déduction compétente. Ce fou croit être le roi de pique, et il en déduit (validement) qu'il n'est pas un cerveau dans une cuve de formol. La croyance $\sim HS$ qu'il forme ainsi est sûre, puisqu'elle est vraie dans tout le voisinage modal objectif (pour le fou comme pour nous). Mais on sera sans doute tenté de ne pas la considérer comme connaissance, parce qu'elle est déduite d'une croyance qui, quant à elle, n'est pas du tout une connaissance. Autrement dit, certaines méthodes de formation de croyance sont telles que, même si elles délivrent des croyances sûres (et vraies), ces croyances ne constitueront pas des connaissances. Et la méthode « déduction compétente à partir d'une croyance *non connue* » entre manifestement dans cette catégorie. La question est de savoir quelles méthodes *permettent* d'obtenir une connaissance lorsqu'elles produisent une croyance sûre (et vraie), et si nous avons à notre disposition une telle méthode pour croire $\sim HS$. La réponse vient assez naturellement à l'esprit : à la différence du fou, nous avons la possibilité de déduire $\sim HS$ à partir de croyances anodines qui sont elles-mêmes *connues*, et même sensibles. Or lorsqu'on déduit *validement* et *avec compétence* une proposition *sûre* à partir d'une croyance *sensible*,

on voit mal ce qui pourrait l'empêcher de constituer une connaissance. L'intuition qui est ici sollicitée est la suivante : la déduction compétente ne suffit peut-être pas à *produire* de la connaissance (même lorsqu'elle conclut à une croyance sûre), mais elle semble à première vue suffisante pour *transmettre* la connaissance lorsqu'elle conclut à une croyance sûre. Après avoir défendu (au chapitre précédent) la transmission de la justification *prima facie* de A à $\sim HS$, c'est donc un modèle de transmission de la *connaissance* qui est ici proposé. Bien sûr, il existe des réticences contre la notion de transmission de la connaissance sous la déduction compétente. Mais ces réticences viennent de deux types de considérations que nous sommes maintenant en mesure de rejeter : premièrement du problème de la circularité, et deuxièmement de l'analyse contrefactuelle. Le problème de circularité consiste à observer que l'hypothèse sceptique HS met en cause la fiabilité même du fondement épistémique de A (c'est-à-dire qu'elle implique $\sim KA$), et qu'il y aurait donc circularité épistémique à déduire $\sim HS$ de A . Mais nous avons vu que de toutes les façons la circularité ne *pouvait pas* être totalement éliminée pour les croyances méta-épistémique de type KA .²⁹⁸ La deuxième raison de refuser la transmission de la connaissance venait de l'analyse contrefactuelle de la connaissance : on peut montrer par de nombreux exemples (comme le cas du zèbre) que la sensibilité des croyances ne se transmet pas à travers la déduction compétente. Mais John Hawthorne (2005) a bien montré que ce phénomène de non-transmission de la sensibilité, loin de montrer l'échec de la transmission de connaissance, montrait bien plutôt le caractère inacceptable de la condition de sensibilité : en effet, on peut produire des exemples dans lesquels il est *évident* que la déduction compétente confère de la connaissance, alors qu'elle fait perdre la sensibilité. L'exemple le plus simple est celui du « vrai chien, faux chat ».

Cas du vrai chien – faux chat

Imaginez qu'un vrai chien et un faux chat²⁹⁹ sont dans une chambre, le premier me cache la vue du second. Je regarde le chien et forme la croyance qu'il y a un chien dans la chambre. De là, je déduis qu'il y a un animal dans la chambre. Imaginez, en outre, que s'il n'y avait pas eu de chien dans la salle, j'aurais vu le faux chat, et j'aurais formé la croyance que c'est un (vrai) animal. Dans ce cas, ma croyance qu'il y a un chien réussit le test de Nozick [la sensibilité contrefactuelle] mais ma croyance qu'il y a un animal dans la chambre échoue. [...] le défenseur de cette théorie [de la sensibilité] ne sera guère ravi par ce résultat. (Hawthorne 2005, 34)

Il semble évident, en effet, que la déduction compétente qu'il y a un animal me permet, en l'occurrence, de connaître qu'il y a un animal, bien que cette croyance ne soit pas contrefactuellement sensible. En réfléchissant sur ce type de cas, on peut voir que la condition de sensibilité est *en elle-même* peu plausible parce qu'elle fait échouer la clôture dans des cas où *personne* ne voudrait la faire échouer.³⁰⁰ Et par ailleurs, on peut apercevoir aussi *comment*

²⁹⁸ Par ailleurs, il faut noter qu'au sens strict la déduction de $\sim HS$ à partir de A n'engage *pas* de circularité épistémique puisque la proposition $\sim HS$ n'est *pas* en elle-même une proposition méta-épistémique : elle *implique* certes une proposition de type KA , mais son contenu propre est bien différent du contenu de KA (des cerveaux, des cuves, des génies...). Il n'y aurait circularité épistémique (acceptable ou non) que si l'on déduisait KA à partir de $\sim HS$ après avoir déduit $\sim HS$ à partir de A , mais l'anti-sceptique n'a nullement besoin de postuler une telle déduction : il pourrait défendre par exemple que, au moment où nous déduisons $\sim HS$ à partir de A , nous connaissons déjà le contenu KA sur la base d'une autre source épistémique (par exemple sur la base d'une expérience mooréenne).

²⁹⁹ Le faux chat en question est aussi un faux animal, par exemple un chat en plastique, et pas un chien déguisé en chat.

³⁰⁰ Il faut noter que Nozick et Dretske ont aperçu ce type de problème et ont essayé de réparer cette faiblesse de la condition de sensibilité en la relativisant aux modes de formation de la croyance en question – (Nozick 1981) parle de « méthodes » épistémiques, (Dretske 1971) d'une relativisation aux « raisons » (qui doivent être « conclusives », c'est-à-dire sensibles). L'idée est la suivante : s'il n'y avait pas eu de chien, j'aurais cru tout de

marque la transmission de connaissance : si la croyance qu'il y a un animal constitue une connaissance, c'est manifestement parce qu'en déduisant valablement cette proposition à partir d'une croyance elle-même connue (et même sensible), on n'aurait pas pu facilement se tromper. De même, en déduisant valablement $\sim HS$ à partir d'une croyance anodine A qui constitue une connaissance (et même une connaissance sensible), on produit une croyance sur laquelle on ne pourrait pas facilement se tromper, une croyance sûre. On peut donc avancer que ces conditions sont *suffisantes* d'une manière générale pour qu'il y ait connaissance :

Condition de déduction sûre à partir de connaissance sensible :

La croyance de S que q est une connaissance toutes les fois que

- (i) S déduit valablement (et avec compétence) que q sur la base de p , et
- (ii) la croyance de S que p est une connaissance (et même une croyance sensible)
- (iii) la croyance de S que q est sûre.

En somme, il y a donc une présomption initiale que la déduction à partir de croyances sensibles transmet le statut de connaissance aux conséquences lorsque ces conséquences sont sûres. Et cette présomption n'est nullement contrecarrée par les considérations de circularité et de sensibilité. La déduction de $\sim HS$ (qui est épistémiquement sûre) à partir de A (qui est épistémiquement sensible) constitue donc la meilleure réponse à l'argument sceptique du double engagement, car cette déduction a toutes les chances de nous conférer positivement la *connaissance* de $\sim HS$, et qu'on ne voit pas de quelle autre manière la croyance $\sim HS$ pourrait compter comme connaissance. Or c'est bien le statut de connaissance qu'il faut pouvoir lui conférer pour éviter le scepticisme du double engagement.

Appliquons enfin ce résultat à l'argument anti-introspection de l'expérience non étayable : il est clair que nous pouvons déduire de notre croyance au libre arbitre (elle-même fondée sur l'expérience E) que nous ne sommes pas manipulés par un dispositif martien. A vrai dire, nous pouvons même déduire cette proposition $\sim MM$ d'autres croyances anodines (ou en tout cas de croyances incontestablement *connues*, et même très certainement *sensibles* à leur contenu) et n'ayant rien à voir avec le libre arbitre, par exemple, que personne n'a jamais rien mis dans mon cerveau, ou qu'il n'y a pas de martiens. Le fait que les dispositifs martiens soient indétectables rend donc « insensible » la croyance $\sim MM$, mais elle ne la rend pas épistémiquement hasardeuse puisqu'il s'agit d'une croyance *sûre* que l'on peut déduire valablement de croyances elles-mêmes connues (et même sensibles). Si l'on lui pose successivement les questions « acceptes-tu la conséquence $\sim MM$ de ta croyance LA ? » et « crois-tu *savoir* que $\sim MM$? », le défenseur de l'introspection ne sera donc aucunement réfuté dans sa croyance au libre arbitre, puisqu'il est en position de répondre légitimement à chaque fois : « oui, je suis tout à fait prêt à soutenir que mes actions ne sont pas manipulées par les martiens, et oui, je suis tout à fait prêt à dire que cette dernière croyance est une *connaissance* et non pas 'vraie par pur hasard' ».

Pour conclure cette section, on peut donc dire que l'argument de l'expérience non étayable ne permet pas de réfuter la justification introspective en faveur du libre arbitre.

même qu'il y a un animal, mais je ne l'aurais pas cru *de la même manière*, ou par la même méthode, ou pour les mêmes raisons. Si ma *croyance* est insensible, mon fondement épistémique, lui, est sensible. Et donc une condition de sensibilité relativisée aux méthodes ou aux raisons permettrait de sauver le verdict intuitif que je *sais* qu'il y a un animal. Malheureusement, John Hawthorne a montré qu'il ne s'agissait là que d'une rustine locale à un problème beaucoup plus grave : en effet, il a mis en avant d'autres exemples pour lesquels les raisons et méthodes *elles-mêmes* sont insensibles, et où l'idée d'une rupture de clôture est tout aussi inacceptable que pour le cas du vrai chien – faux chat. Je laisse au lecteur le plaisir d'aller découvrir le cas des 6 kg de saumon (Hawthorne 2005, 36–37).

Certes, le contenu de l'expérience du libre arbitre, même en amont de toute analyse controversée, a des conséquences lourdes telles que $\sim MM$ qui ont la particularité d'être indétectables *ex hypothesi*. Mais les expériences perceptives anodines ont, elles aussi, des conséquences lourdes indétectables *ex hypothesi* (les hypothèses sceptiques), et cela ne les empêche pas de conférer une justification *ultima facie* en faveur de leur contenu. Par simple analogie, on peut donc voir la faiblesse de l'argument.

Si l'on veut aller plus loin et comprendre *pourquoi* l'indétectabilité des conséquences lourdes de *A* ou de *LA* n'empêchent pas ces croyances de constituer des connaissances, il semble qu'il n'y ait que trois réponses possibles : soit l'indétectabilité de ces conséquences lourdes ne nous empêche pas de *connaître* leur fausseté (du moment qu'elles sont *sûres*), soit une absence de connaissance de leur fausseté ne nous empêche pas de *croire* légitimement à leur fausseté (refus de la méta-cohérence), soit l'impossibilité de croire légitimement à leur fausseté ne nous empêche pas de croire légitimement des propositions qui les impliquent logiquement (refus de la clôture de justification). Ayant montré que la méta-cohérence et la clôture de justification ne pouvaient pas être rejetées dans certaines versions faibles, j'ai conclu que la seule théorie plausible consistait à reconnaître positivement le statut de *connaissances* aux croyances $\sim HS$ ou $\sim MM$ du fait qu'elles sont épistémiquement *sûres* et déduites valablement (et avec compétence) de propositions elles-mêmes connues (et même sensibles).

Cette « condition de sûreté » pourra resservir dans la suite de cette étude, lorsque l'analyse conceptuelle du contenu de *E* révélera d'autres conséquences lourdes (typiquement : l'indéterminisme), et qu'on se demandera si ces conséquences ne pourraient pas constituer, par leur inconnaissabilité, des réfutations de la croyance au libre arbitre. Mais dans le cadre de cette première partie, je m'en suis tenu à ce que l'on peut dire en amont de toute analyse controversée du contenu de l'expérience ; et à ce niveau, le verdict me semble être clairement le suivant : non seulement nous avons une raison *prima facie* de considérer notre expérience du libre arbitre comme une source de *connaissance* de son contenu (cf. section précédente de ce chapitre), mais en outre, nous n'avons aucune raison en sens inverse, c'est-à-dire aucune raison de penser que notre expérience du libre arbitre aurait un contenu trop lourd pour être connu. En amont de toute analyse du contenu de *E*, nous sommes donc entièrement justifiés (justifiés *ultima facie*) à tenir notre expérience du libre arbitre comme une source de *connaissance* de son contenu.

IV. Conclusion : la justification indexicale de second ordre et le « point de départ épistémologique » du problème du libre arbitre

Au début de ce chapitre, je me suis efforcé de distinguer deux types d'objections ou de réfutations qu'on pourrait faire contre notre croyance introspective au libre arbitre : d'une part les objections « pré-scientifiques », et d'autre part les objections qui pourraient provenir de l'analyse conceptuelle du contenu de cette croyance, ou d'informations empiriques sur le fonctionnement de notre expérience *E*. Le but de cette distinction est de pouvoir faire apparaître un « état initial de la question du libre arbitre », un « point de départ épistémologique » où l'on se trouve avant toute enquête spécialisée (philosophique et/ou empirique). Et toute la première partie de cette étude a consisté à montrer que, si l'on se place de ce point de vue, la croyance introspective au libre arbitre est parfaitement justifiée. C'est-à-dire qu'il n'y a *aucune* objection solide à cette croyance en amont de l'enquête spécialisée. La deuxième partie consistera à voir s'il y a ou non des réfutations qui résultent de l'enquête

spécialisée, et en particulier s'il y a des réfutations qui résultent de l'analyse conceptuelle du contenu de l'expérience.

Dans cette dernière section, je vais faire le point rapidement en présentant de manière plus synthétique les différentes objections, celles que nous avons déjà considérées et celles qui sont à venir dans la deuxième partie. Le but principal de cette section sera d'expliquer le rôle *épistémologique* des discussions de la deuxième partie. En effet, comme cette deuxième partie va entrer (enfin) dans l'analyse métaphysique de ce qu'est le libre arbitre, les discussions épistémologiques seront moins nombreuses, mais elles resteront le cadre et l'enjeu de toutes les discussions métaphysiques. Aussi faudra-t-il être préalablement très au clair sur ces enjeux : à quelles conditions ces analyses métaphysiques constitueront-elles des *réfutations* ? Comment pourraient-elles faire apparaître un « Problème d'Intégration » ? Quelle serait la réaction la mieux justifiée face à un tel problème ?

Les deux premières objections que nous avons considérées dans le chapitre 2 ont un statut un peu à part par rapport à toutes les autres : elles tentent de montrer que la croyance introspective au libre arbitre n'aurait *aucune* sorte de justification (objection de la pure impression et objection de l'impression faillible). A cela j'ai répondu que l'expérience commune du libre arbitre donnait au moins une certaine justification *prima facie* en faveur de son contenu, au même titre que les expériences perceptives (non moins faillibles) ou toute autre expérience doxastique. A partir de là, toutes les autres objections consisteront à savoir si cette justification *prima facie* est *réfutée* d'une manière ou d'une autre. Et pour envisager cette question de manière systématique, j'ai proposé deux distinctions fondamentales qui permettent de classer les réfuteurs possibles : la distinction entre réfuteurs directs (ou de premier ordre) et réfuteurs fondationnels (ou de deuxième ordre), et la distinction entre réfuteurs pré-scientifiques et réfuteurs scientifiques (ou spécialisés). Ces deux distinctions sont indépendantes l'une de l'autre et permettent de dresser le tableau suivant :

réfuteurs possibles de la justification <i>prima facie</i>	directs	fondationnels
pré-scientifiques	chapitre 2, II.3	chapitre 3
post-empiriques	chapitre 5	chapitre 6

Comme le rappelle ce tableau, ma démarche a consisté à examiner d'abord (chapitres 2 et 3) les éventuelles réfutations pré-scientifiques, c'est-à-dire celles qui se situent en amont de toute information philosophique ou empirique spécialisée.

Dans la fin du chapitre 2, j'ai montré que la justification *prima facie* fournie par l'expérience E n'était en présence d'aucun réfuteur *direct* de son contenu. En effet, tant qu'on reste en amont de toute analyse philosophiquement controversée, les seules implications du contenu de E sont du type $\sim MM$, et nous n'avons assurément aucune raison positive de croire que nos actes sont manipulés par des dispositifs martiens – et nous ne manquons pas non plus de raison positive pour *refuser* légitimement ces scénarios, dans la mesure où nous pouvons déduire leur négation d'autres croyances justifiées.

Dans le chapitre 3, je suis passé des réfutations pré-scientifiques directes aux réfutations pré-scientifiques fondationnelles, qui contestent la capacité de l'expérience E de constituer une source de *connaissance*. Là aussi, j'ai montré que ces objections ne pouvaient pas constituer des réfutations de la justification par expérience E, si du moins on veut éviter de dire que les justifications par expériences perceptives sont elles-mêmes réfutées. Ces objections fondationnelles étaient de deux types. D'abord une objection *négative* disant que nous *manquons* d'une justification « méta-épistémique », c'est-à-dire d'une raison de tenir l'expérience E pour une expérience fiable (objection de l'impression non étayée) ; et j'ai

montré que nous n'étions pas dans une situation différente de celle des expériences perceptives de ce point de vue, c'est-à-dire que si l'on veut éviter le scepticisme général, on est bien obligé d'accepter une conception réaliste des méta-justifications, c'est-à-dire une conception ultimement circulaire et fondée sur le sens commun. Or une telle conception permet tout à fait de dire que *nous avons* une raison valable (au moins *prima facie*) de tenir notre expérience E pour une source de *connaissance*. A partir de là, la question fondationnelle (dans cette partie *et dans la suivante*) pourra simplement consister à se demander si cette méta-justification *prima facie* est réfutée ou non – c'est-à-dire si nous avons une raison *positive* (et plus forte) en sens inverse, une raison *positive* de croire que notre expérience ne constitue *pas* une connaissance. Dans la deuxième moitié du chapitre 3, j'ai montré qu'il n'y avait aucune raison solide dans ce sens si l'on s'en tient aux données pré-scientifiques. La seule raison qu'on pourrait avancer à ce stade, c'est que les conséquences de type $\sim MM$ ne semblent pas pouvoir être *connues* puisqu'elles sont indétectables (objection de l'impression non étayable) ; mais j'ai montré que cette objection reposait sur un standard de « sensibilité » qui est trop élevé, et mènerait au scepticisme même pour la perception. Pour éviter le scepticisme, il faut accepter que des croyances « sûres » puissent compter comme connaissance, au moins lorsqu'elles sont déduites validement de croyances elles-mêmes connues (et même sensibles). Or les croyances de type $\sim MM$ peuvent aisément satisfaire cette contrainte.

La conclusion générale de la première partie est donc la suivante : lorsque nous croyons au libre arbitre sur la base de notre expérience commune E, nous avons une justification *prima facie* à accepter son contenu, et en outre une justification *prima facie* à tenir cette croyance pour une *connaissance*, et ces justifications *prima facie* ne sont nullement réfutées avant qu'on envisage des données philosophiques ou empiriques spécialisées. C'est-à-dire qu'en amont de telles enquêtes spécialisées, nous sommes *ultima facie* justifiés à croire (sur la base de E) que nous sommes libres et que cette croyance est une connaissance en bonne et due forme. C'est ce que j'appelle le « point de départ épistémologique » de la question du libre arbitre. Il me semble absolument crucial de prendre clairement conscience de ce point de départ pour pouvoir ensuite évaluer les éventuelles modifications que pourraient apporter des informations philosophiques ou empiriques spécialisées. Il ne s'agit ni de surestimer ni de sous-estimer ce point de départ. En un sens, prétendre à une justification pré-scientifique en faveur de telle ou telle proposition (LA ou KLA) est une thèse extrêmement *faible*, puisqu'elle est compatible avec toutes sortes de réfutations scientifiques. Mais c'est aussi une thèse dont l'aspect positif a trop souvent été négligé : à savoir qu'en *l'absence* de réfutations scientifiques, il n'y a nul besoin de rechercher une *confirmation* scientifique, puisque la justification pré-scientifique peut continuer de fonctionner.

Dans la deuxième partie, j'envisagerai ces éventuelles réfutations scientifiques, en utilisant la même classification entre réfutations directes et réfutations fondationnelles. Pour pouvoir déterminer si l'analyse philosophique du contenu de notre expérience E constitue une source de réfutation pour la justification indexicale, il faudra tout d'abord procéder à cette analyse elle-même : c'est ce que je ferai dans le chapitre 4. Et je défendrai que l'expérience du libre arbitre a toutes les chances d'avoir un contenu qui implique l'indéterminisme causal. C'est-à-dire que je défendrai une analyse *libertarienne* de notre expérience commune du libre arbitre. Disposant d'une telle analyse, je me demanderai ensuite si ce contenu libertarien est réfuté directement ou fondationnellement.

Dans le chapitre 5, j'envisagerai les éventuelles réfutations directes (ou de premier ordre) de la croyance libertarienne. Ayant défendu (au chapitre 4) que le contenu de notre croyance au libre arbitre implique logiquement l'indéterminisme, et (au chapitre 2) que nous avons une justification *prima facie* à accepter les conséquences logiques reconnues de nos croyances

justifiées, je partirai du fait que nous avons, au moins, une justification *prima facie* à croire à l'indéterminisme. La question sera donc de savoir si nous avons des raisons positives (et plus fortes) en sens contraire. Et j'envisagerai deux types de considérations qui pourraient contrecarrer directement notre justification *prima facie*. D'une part, des considérations *a priori* tendant à montrer qu'un monde contenant le libre arbitre (au sens libertarien) n'est tout simplement pas *possible*, logiquement ou métaphysiquement. D'autre part, des considérations *empiriques* consistant à dire que le libre arbitre, quoi que logiquement ou métaphysiquement possible, n'est pas compatible avec certaines informations scientifiques sur le monde dans lequel nous vivons effectivement. La science nous aurait appris que nous vivons, en fait, dans un monde déterministe, ou dans un monde où les indéterminismes ne se situent pas au bon endroit pour permettre le libre arbitre. Dans la discussion traditionnelle, ces deux objections correspondent à ce que Robert Kane (1996, 13) a appelé la « Question d'Intelligibilité » du libertarianisme : peut-on donner une description intelligible de ce que pourrait être un libre arbitre libertarien dans le monde tel que nous le connaissons aujourd'hui ? Et cette question se subdivise en une question d'intelligibilité *a priori* (le principal problème, ici, étant « l'argument de la chance ») et une question d'intelligibilité *scientifique* (où trouver concrètement un indéterminisme *pertinent pour la liberté* chez des agents humains composés de neurones et d'atomes ?). L'objection scientifique principale consiste à dire que le monde tel que la science contemporaine le décrit, s'il n'est pas déterministe à strictement parler, semble *quasi* déterministe *pour ce qui est des actions humaines*, ou déterministe « *for all practical purposes* » (FAPP). Si la science a aujourd'hui établi le déterminisme FAPP, alors elle constituera un réfuteur direct de la justification que nous avons initialement pour une conception indéterministe de nos actions. Pour chacune de ces deux objections directes, ma réponse fera fond sur les efforts considérables qui ont été déployés depuis une vingtaine d'années pour construire des modèles du libertarianisme qui soient à la fois logiquement intelligibles et compatibles avec ce que nous savons aujourd'hui scientifiquement des agents humains. Je pense en particulier aux travaux de Robert Kane (1996), de Tim O'Connor (2000) et de Mark Balaguer (2010). Et je conclurai que les prétendues réfutations directes du libertarianisme ne sont pas convaincantes : nous n'avons aucune raison positive de penser que *nous ne sommes pas* libres dans le sens libertarien. Dans cette situation, notre justification introspective à croire que *nous sommes* libres dans le sens libertarien restera donc irréfutée à moins qu'il y ait une réfutation de deuxième ordre.

Dans le chapitre 6, j'envisagerai enfin les objections de deuxième ordre, qui me semblent être les objections les plus importantes. Elles reviennent à dire que notre expérience du libre arbitre est réfutée lorsqu'on prend conscience qu'elle ne pourrait pas constituer une source de *connaissance*. Là encore, nous avons vu qu'il y avait deux possibilités. Soit on s'appuie sur des données empiriques (par exemple les expériences de Benjamin Libet) qui semblent indiquer une déficience de l'expérience du libre arbitre *indépendamment* de toute analyse compatibiliste ou incompatibiliste (ainsi, c'est l'introspectionnisme a-compatibiliste qui se trouverait réfuté scientifiquement). Soit on considère que c'est l'analyse incompatibiliste elle-même qui fait apparaître pour l'expérience du libre arbitre un contenu trop lourd pour être connu par expérience. Je présenterai rapidement les réponses aujourd'hui assez répandues aux objections de type Libet ; mais l'essentiel de mon propos consistera à répondre au deuxième type d'objection. C'est ce deuxième type d'objection qui constitue à mon sens le cœur du « Problème d'Intégration » du libertarianisme, et je crois aussi la raison la plus profonde des résistances contemporaines à adopter une thèse libertarienne. Je finirai donc cette partie en rappelant brièvement la structure épistémologique de ce type de problème.

réfutateurs possibles de la justification <i>prima facie</i>	directs	fondationnels
pré-scientifiques	objection de l'impression lourde (ch.2, II.3)	objection de l'impression non étayée (ch.3, II) objection de l'impression non étayable (ch.3, III)
post-empiriques	objection d'inintelligibilité <i>a priori</i> (ch. 5.I) objection scientifique du déterminisme FAPP (ch.5.II)	objection à la Libet (ch. 6.I) Problème d'Intégration de la croyance Indéterministe (ch. 6.II)

Un problème d'Intégration, comme on l'a vu en Introduction (p.13), est un conflit entre trois types d'intuitions :

- (IM) une intuition mooréenne (ou épistémologique particulière) sur un domaine d'objets. Par exemple : que « la croyance '2+2=4' constitue une connaissance », ou que « la croyance 'je suis libre' constitue une connaissance ».
- (IA) une intuition analytique sur ce domaine d'objets. Par exemple : que « les nombres (2 ou 4) sont des objets abstraits », ou que « être libre implique l'indéterminisme ».
- (IE) une intuition épistémologique générale sur les conditions de possibilité de la connaissance. Par exemple : que « connaître un objet *O* suppose d'avoir un rapport causal avec *O* », ou que « connaître l'indéterminisme supposerait de le connaître comme résultat de la méthode d'investigation scientifique ».

Si l'on se situe en amont de l'analyse philosophique du concept de libre arbitre, il me semble clair qu'il n'y a pas de problème d'intégration sérieux pour la croyance au libre arbitre : en effet, à ce niveau-là, les seules conséquences analytiques de notre expérience *E* qui pourraient sembler difficiles à *connaître* par expérience seraient les conséquences de type $\sim MM$, et je viens de montrer que tout auteur anti-sceptique serait prêt à accepter que ces conséquences ne constituent pas, en fait, une objection sérieuse.

C'est donc l'analyse incompatibiliste, associée avec une intuition « d'exclusivité scientifique » pour la connaissance de l'indéterminisme, qui font *apparaître* un problème d'Intégration (un problème de « type Benacerraf »)... et elles ne font apparaître un tel problème *que* parce que nous avons par ailleurs, et antécédemment, une intuition épistémologique *positive* à propos de cette croyance, c'est-à-dire l'intuition mooréenne que notre expérience du libre arbitre n'est pas une pure illusion mais nous permet de *savoir* que nous sommes libres pour telle ou telle action. Autrement dit, quand bien même l'investigation philosophique et scientifique ferait naître en nous des intuitions (IA) et (IE) susceptibles de provoquer un problème d'Intégration, on ne pourrait pas trivialement en conclure que l'intuition mooréenne est *ipso facto* réfutée : il faudrait en conclure qu'on a fait apparaître un trilemme, et que l'une de ces trois propositions doit être abandonnée. Reste à savoir *laquelle* devrait alors être abandonnée (et *pourquoi* abandonner celle-ci plutôt que celle-là). C'est ce que j'appellerai la *Question d'Arbitrage*.

Question d'Arbitrage (pour le Problème d'Intégration du Libertarianisme)

Etant donné que nous avons initialement une intuition mooréenne selon laquelle (IM) notre expérience *E* nous donne connaissance du fait que nous sommes libres si la réflexion philosophique et scientifique nous amène à avoir en outre des intuitions analytiques et épistémologiques qui génèrent un Problème d'Intégration, à savoir : (IA) être libre suppose l'indéterminisme causal

(IE) notre expérience commune ne peut pas délivrer une *connaissance* de l'indéterminisme causal
laquelle de ces trois intuitions devrait être abandonnée ?

La question d'arbitrage admet clairement trois réactions possibles.

Abandonner (IM). C'est la solution qu'on pourrait appeler « empirisme libertarien »³⁰¹ : l'analyse incompatibiliste du contenu de notre expérience constitue une réfutation fondationnelle de notre croyance commune au libre arbitre, parce que ce contenu incompatibiliste est trop lourd pour être connu par simple expérience commune. Il faut donc abandonner notre croyance commune au libre arbitre, en attendant que d'autres sources d'informations (physique, neurosciences, etc.) nous renseignent sur la question.

Abandonner (IA). C'est la solution que j'appellerai « compatibilisme d'ajustement »³⁰² : notre expérience E a l'air d'avoir un contenu incompatibiliste, mais comme cela impliquerait de la révoquer en doute, et qu'un tel doute représenterait un écart incroyable ou injustifié par rapport au sens commun, c'est qu'elle doit en fait avoir un contenu moins lourd que cela, un contenu compatibiliste.

Abandonner (IE). J'appellerai cela la solution « d'Intégration forcée » : puisque notre expérience du libre arbitre a des conditions de satisfaction incompatibilistes et qu'on ne peut nier son statut de connaissance sans violer gravement le sens commun, il faut donc remanier notre épistémologie générale pour y faire entrer la connaissance introspective du libre arbitre.

Chacune de ces trois réactions, telles que je viens de les décrire, consiste à refuser une des intuitions *à cause* des deux autres, ou *en réaction* au Problème d'Intégration. Une autre possibilité, plus satisfaisante sans doute, serait de pouvoir éviter le problème en ayant des raisons *indépendantes* pour rejeter l'une ou l'autre de ces intuitions. Si l'on ne trouve pas de telles raisons, alors le problème d'Intégration se posera à plein, et il faudra bien adopter une des trois solutions « réactives ». Peut-on trouver de telles raisons ?

Des raisons indépendantes d'abandonner (IM) ? C'est-à-dire des raisons de n'accorder aucun crédit à notre expérience du libre arbitre, quel que soit son contenu. La thèse ainsi défendue est ce que j'appelle le « scepticisme du Libre Arbitre ».³⁰³ J'ai essayé de montrer dans cette première partie qu'il n'y avait pas de telles raisons³⁰⁴ : tant que le problème d'Intégration n'est pas soulevé par une analyse problématique de E, nous avons toutes les raisons de nous fier à cette expérience comme à une source de connaissance. Bien sûr, on *pourrait* être amené à abandonner (IM), mais ça ne peut être qu'en réaction à un problème d'intégration.

³⁰¹ Cette solution ne peut être appelée « libertarienne » que dans le sens faible où elle admet une *conception* libertarienne du libre arbitre (elle admet sa concevabilité, et même sa réalisabilité dans le monde tel qu'il est). Il est clair qu'elle n'est pas libertarienne dans le sens fort, qui affirme l'*existence* de ce type de libre arbitre.

³⁰² J'expliquerai un peu plus loin la différence entre ce type de compatibilisme et d'autres types de compatibilisme.

³⁰³ Cette thèse serait-elle encore « sceptique » si on lui ajoute l'idée d'une justification *non-expérientielle* à croire au libre arbitre (par exemple une justification par témoignage divin) ? Une telle position ne serait pas « sceptique » dans le sens de la suspension du jugement, mais elle serait sceptique dans un sens plus général, celui de la remise en cause de nos facultés naturelles. On peut comparer avec le cas suivant : un témoignage divin m'apprend que mon expérience visuelle d'avoir deux mains est le fruit d'un générateur aléatoire implanté dans mon cerveau, mais par ailleurs, le même témoignage divin m'informe que j'ai deux mains. Dans ce cas, je ne retiendrai pas mon jugement (je croirai que j'ai deux mains), mais je serai clairement dans une situation « sceptique » dans un sens important.

³⁰⁴ Pour être précis, cette conclusion dépend aussi de la réponse que j'apporterai aux objections de type « Libet » au chapitre 6 : en effet, ces objections, quoique scientifiques, ne sont pas dépendantes du problème d'Intégration puisqu'elles réfutent l'expérience du libre arbitre à cause de son *processus de formation* et non à cause de la lourdeur de son contenu.

Des raisons indépendantes d'abandonner (IA) ? C'est-à-dire des motivations pour devenir compatibiliste qui ne *dépendraient pas* de l'intention de préserver notre croyance de sens commun, mais simplement de l'étude de l'expérience elle-même. C'est ce que j'appellerai le « compatibilisme indépendamment motivé ». Dans le chapitre 4, je reviendrai sur cette distinction entre « compatibilisme d'ajustement » et « compatibilisme indépendamment motivé », et je soutiendrai que le *deuxième type* de compatibilisme n'est pas légitime. C'est-à-dire que, à mon sens, si l'on étudie le contenu de E *pour lui-même*, et sans se poser la question de la *vérité* ou de l'acceptabilité de ce contenu, on a toutes les raisons de considérer qu'il a des conditions de vérité incompatibilistes. Peut-être que le caractère « non connaissable » d'un tel contenu pourrait *ensuite* nous amener à réajuster notre analyse (pour sauver la connaissance de sens commun), mais en tout cas, il n'y pas de raisons *indépendantes* d'abandonner (IA).

Des raisons indépendantes d'abandonner (IE) ? C'est ce que j'appelle la thèse « d'Intégration directe », et c'est la thèse que je défends dans ce travail. Je m'efforcerai de montrer dans le chapitre 5 que (IE) repose fondamentalement sur des intuitions épistémologiques beaucoup trop exigeantes, des intuitions que l'on est obligé d'abandonner de toutes les façons pour éviter le scepticisme général. Les éléments d'épistémologie anti-sceptique, que j'ai établis dans cette première partie pour la perception, s'avèreront *suffisants* pour pouvoir dire que l'expérience du libre arbitre nous procure une véritable *connaissance* (même si elle a un contenu incompatibiliste).

solutions	réactives	indépendamment motivées
~ (IM)	empirisme libertarien	scepticisme du Libre Arbitre
~ (IA)	compatibilisme d'ajustement	compatibilisme indépendamment motivé
~ (IE)	Intégration forcée	Intégration directe

Dans la mesure où je défends une solution indépendamment motivée au problème d'intégration, je n'ai pas réellement besoin de donner des « conditions d'arbitrage » entre les trois intuitions. Je n'ai pas besoin de spécifier quelle intuition devrait être abandonnée *par rapport aux autres*, puisque l'intuition (IE) doit être abandonnée pour elle-même. Mais je vais tout de même exposer brièvement quel arbitrage je *ferais* si (IE) ne pouvait pas être indépendamment rejetée. Cela permettra sans doute de mieux comprendre ma démarche générale. Si l'intégration directe échouait, si les trois intuitions du problème d'intégration étaient justifiées en elles-mêmes et indépendamment des deux autres, alors il faudrait bien abandonner l'une d'elles « de manière réactive » – faute de quoi, on se trouverait dans un état épistémique soit inconséquent (on croirait au libre arbitre sans croire sa conséquence logique qu'est l'indéterminisme), soit incohérent (on croirait à l'indéterminisme tout en sachant que cette croyance est purement hasardeuse épistémiquement). Mais quel critère employer pour déterminer la bonne réaction ? Je vois trois critères possibles pour faire cet arbitrage : le droit du plus fort, la priorité du sens commun, et la priorité de la raison.

Le droit du plus fort consiste à dire que l'intuition qui doit (épistémiquement) être abandonnée, celle qui se trouve réfutée par le problème d'Intégration, est celle qui a le moins de « force intuitive ». Les expériences doxastiques se présentent avec un degré plus ou moins fort de persuasion. Par exemple, je peux avoir deux intuitions mathématiques qui s'avèrent incompatibles entre elles : si l'une d'elles est une faible inclination à croire tandis que l'autre est ressentie comme presque indubitable, alors il semble que la réaction la mieux justifiée est de « favoriser la proposition la plus intuitive » (Pust 2012, sec. 3.3.2). Dans ce cas, on observe donc une forme de « droit de l'intuition la plus forte ». En partant d'une analyse de la « méthode de l'équilibre réflexif » de Rawls, Joel Pust (2000, chap. 1) a montré que cette stratégie de pondération des intuitions représentait en réalité la méthodologie philosophique la

plus commune. On retrouve notamment cette méthodologie chez (van Inwagen 1983, 219–220) lorsqu’il établit une liste de quatre intuitions fondamentales pour le problème du libre arbitre, classées en fonction de leur force intuitive (pour lui, évidemment).³⁰⁵ Cette méthodologie philosophique est justifiée classiquement de la manière suivante. (i) *Toute philosophie doit d’une manière ou d’une autre reposer ultimement sur des intuitions.* Pour reprendre la formulation de Kripke :

Pour certains philosophes, le fait que quelque chose ait un contenu intuitif est une raison très insuffisante en sa faveur. Pour moi, je crois que c’est une raison très puissante pour y croire. Je ne sais vraiment pas, d’une certaine manière, comment on pourrait avoir une raison plus conclusive. (Kripke 1980, 42)

Cette première thèse n’est rien d’autre que la thèse du phénoménalisme charitable défendue par Huemer (cf. chapitre 2, p.172). Pour arriver au droit du plus fort, il faut ajouter : (ii) à moins d’avoir beaucoup de chance, ces intuitions (philosophiques) de base auront tendance à générer des conflits intuitifs. Dans ce genre de cas, que pourrait-on faire *de mieux* que de reposer sur les *meilleures* de nos intuitions, c’est-à-dire sur les plus fortes ? Cette règle du « droit du plus fort » est indispensable pour certains conflits d’intuitions, mais on peut se demander si elle est absolument universelle : est-ce que *n’importe quelle* intuition, si elle est suffisamment forte psychologiquement, peut réfuter n’importe quel autre type d’intuition ? N’y a-t-il pas d’autres critères que la simple intensité psychologique pour faire un arbitrage entre intuitions ? Et en particulier, certaines sortes d’intuitions n’ont-elles pas une forme de *priorité* par rapport aux autres ? Je vois principalement deux critères de priorité qui pourraient venir limiter le droit du plus fort, et éventuellement changer l’arbitrage qu’il faudra faire pour l’Intégration du Libre Arbitre.

La priorité de la « Raison », c’est-à-dire des intuitions purement analytiques ou *a priori* sur les intuitions qui concernent un fait contingent. Appliqué au problème d’intégration du libre arbitre, ce principe interdirait de réviser l’intuition incompatibiliste (IA) simplement parce qu’elle aurait pour conséquence de rendre le libre arbitre inconnaissable. La réaction de « compatibilisme d’ajustement » ne serait donc pas légitime. John Fischer a critiqué ce type de réaction en soulignant qu’il serait extrêmement « étrange d’abandonner des principes métaphysiques », des « vérités métaphysiques manifestement *a priori* concernant le libre arbitre », simplement pour accommoder les faits contingents à la conclusion que l’on veut obtenir (nous sommes libres, et nous le savons). Pour Fischer, cette attitude constitue une « manière étrange et peu attrayante de retourner sa veste métaphysique (*metaphysical flipflopping*) ». (Fischer et al. 2007, 47) Le principe de priorité de la raison interdit donc de rejeter (IA) pour résoudre le problème d’intégration; il interdit aussi, très certainement, de rejeter les intuitions d’épistémologie générale (IE) qui fondent l’inconnaissabilité de l’indéterminisme par introspection. En effet, s’il y a effectivement de telles intuitions, elles sont très certainement des intuitions analytiques sur le sens du mot connaître. Par conséquent, il semble que la seule réaction légitime serait de rejeter l’intuition (IM), c’est-à-dire de

³⁰⁵ Les quatre prémisses fondamentales de van Inwagen sont les suivantes, par ordre décroissant de force intuitive :

- (1) Nous sommes parfois moralement responsables des conséquences de nos actes.
- (2) La validité de la règle (β) implique que le libre arbitre suppose l’indéterminisme.
- (3) Si nous sommes parfois moralement responsables (1), alors nous avons le libre arbitre.
- (4) La règle (β) est valide.

De ces prémisses, van Inwagen déduit :

- (5) L’indéterminisme est vrai.

Si on lui apportait une preuve de fausseté de cette conclusion, il abandonnerait la prémisse la plus faible intuitivement, à savoir la prémisse (4).

devenir empiriste libertarien : les analyses conceptuelles sur le libre arbitre et sur la connaissance nous *obligeraient* à cesser de croire que nous savons que nous sommes libres.

Un autre principe de priorité, la priorité du Sens Commun, entraîne un arbitrage très différent. C'est une des thèses les plus importantes de la Tradition du Sens Commun que les données du Sens Commun sont toujours plus assurées que toutes les autres, et en particulier que les conclusions philosophiques. Thomas Reid en a fait le cœur de sa critique des philosophies cartésienne et lockéenne :

Les défauts et imperfections de la philosophie actuelle concernant l'esprit, qui l'ont tout particulièrement exposée au mépris et au ridicule des hommes raisonnables, sont principalement liés à ceci: que les défenseurs de cette philosophie, partant d'un préjugé favorable en sa faveur, ont tenté d'étendre sa juridiction au-delà de ses justes limites, et d'appeler à sa barre les données du Sens Commun. Mais ces dernières refusent une telle juridiction; ils méprisent le tribunal du raisonnement, et désavouent son autorité [...] Dans ce combat inégal entre le Sens Commun et la Philosophie, cette dernière s'en tirera toujours avec déshonneur et à ses dépens.

(Reid 1764, sec. I.4)

Dans l'*Essai sur les pouvoirs actifs*, Reid a directement appliqué cette théorie de la prééminence du Sens Commun au cas de la liberté :

Cette conviction naturelle, que nous avons d'agir librement, et qui est concédée par la plupart de ceux qui soutiennent la doctrine de la nécessité, devrait jeter toute la charge de la preuve contre ce parti ; car, à cause d'elle, le parti de la liberté a ce que les juristes appellent un *jus quaesitum*, ou un droit d'ancienneté, qui devrait prévaloir tant qu'il n'est pas cassé.

(Reid 1788, sec. IV.vi)

A la suite de Reid, Chisholm³⁰⁶ et Taylor³⁰⁷ ont également défendu que l'intuition de Sens Commun en faveur du libre arbitre (l'intuition (IM)) a une forme de prééminence par rapport à toute théorie ou intuition philosophique ultérieure. Cette méthodologie n'est pas utilisée seulement par des auteurs incompatibilistes. Holton, un compatibiliste, exprime exactement le même genre d'intuition lorsqu'il soutient que l'intuition « à la Johnson » donne à notre expérience du libre arbitre une *prééminence* quasi indéracinable.³⁰⁸ De même, le compatibiliste Stephen Boulter, bien qu'il développe une théorie non-reidienne du Sens

³⁰⁶ « I assume that people do things when they could have done something else instead. This, I would say, is one of those things that we have a right to believe about ourselves [...]. In the case of a conflict between such a proposition and a philosophical theory, the burden of proof should be upon the philosophical theory. » (Chisholm 1976, 200)

³⁰⁷ « There are, however, two things about myself of which I feel quite certain. [...] the second is that [...] it is sometimes up to me what I do. This might all be an illusion, of course; but so also might any philosophical theory, such as the theory of determinism, be false. The point remains that it is far more difficult for me to doubt that I sometimes deliberate, and that it is sometimes up to me what to do, than to doubt any philosophical theory whatever, including the theory of determinism. We must, accordingly, if we ever hope to be wiser, adjust our theories to our data and not try to adjust our data to our theories. » (Taylor 1991, 39) Il est à noter que, dans cette formulation, Taylor ne tranche pas de manière très nette entre une solution de type « droit du plus fort » où l'intuition la plus forte serait justement celle de Sens Commun, et une solution de type « droit du Sens Commun par rapport à toute forme de théorisation ».

³⁰⁸ « Our knowledge of our own free will is more certain than any thesis of philosophy; so if it comes to a clash between the two, it is philosophy that should give way. » (Holton 2006, 2)

Commun,³⁰⁹ approuve l'idée générale de Reid selon laquelle (i) notre croyance au libre arbitre est de Sens Commun, et (ii) que ce statut impose une « charge de la preuve » considérable sur les épaules de ceux qui nient le libre arbitre, une charge qu'aucun argument philosophique ne pourra renverser.³¹⁰

Si l'on applique ce principe, l'intuition (IM) devrait donc être préservée au bénéfice des deux autres. Quelle intuition faut-il alors rejeter ? L'intuition (IE) ou l'intuition (IA) ? Il peut y avoir débat sur la question : certains auteurs libertariens considéreront notamment que l'analyse incompatibiliste est elle-même une intuition de Sens Commun (par opposition aux théories philosophiques compatibilistes). Dans ce cas, le seul arbitrage possible consisterait à réviser nos théories épistémologiques pour préserver le Sens Commun. Mais d'autres auteurs, notamment Taylor, défendent une méthodologie du Sens Commun qui *pourrait* (si besoin) s'accomoder du compatibilisme d'ajustement.

Quel principe d'arbitrage faudrait-il donc adopter si l'on ne pouvait vraiment pas éviter le problème d'Intégration ? Le droit du plus fort (général), la priorité de la raison, ou la priorité du Sens Commun ? Personnellement, je ne suis pas convaincu par les principes de priorité d'une manière générale. Si une intuition analytique très faible entre en conflit avec une intuition empirique scientifiquement très solide, je ne vois pas pourquoi on ne pourrait pas, par principe, abandonner la première. De même, une méthodologie du Sens Commun qui ferait systématiquement prévaloir nos jugements de Sens Commun sur toute donnée scientifique me semble extrêmement difficile à défendre – à moins de définir la notion de « principes du Sens Commun » en un sens si étroit qu'il soit nécessairement au fondement de toute théorie scientifique... mais alors il y a peu de chances qu'on retrouve notre conviction naturelle d'être libres au sein des principes de Sens Commun ainsi définis.

L'épistémologie que je défends ici n'est donc *pas* une épistémologie du Sens Commun dans le sens fort qui est ici en jeu. La tradition du Sens Commun est une tradition très vaste, qui inclut un grand nombre de thèses épistémologiques « généreuses ». Certaines thèses sont excessivement généreuses et tendent vers un simple quiétisme. Le risque, lorsqu'on défend une épistémologie de Sens Commun, est donc d'adopter une attitude si généreuse qu'elle n'a finalement plus grand-chose d'une défense proprement philosophique. Pour rejeter ce reproche, il peut donc être utile de préciser en quel sens ma démarche appartient à la tradition du Sens Commun, et en quel sens elle s'en démarque. Les deux éléments essentiels que je garde de cette tradition sont :

La méthodologie anti-sceptique : pour rechercher quels principes épistémologiques sont vrais, il faut partir de l'hypothèse qu'ils ne mènent pas au scepticisme.³¹¹

La Métajustification non-scientifique de la perception : nous avons une raison de tenir nos expériences sensibles comme sources de connaissance, et cette raison précède les informations scientifiques sur le fonctionnement de la perception.³¹²

³⁰⁹ Pour Boulter, une croyance est de Sens Commun en fonction de critères d'efficacité évolutionniste par la sélection naturelle. « A belief ought be counted as a common sense belief and treated as a default position if that belief has ecological and social fitness value. » (Boulter 2007, 159). Voir également (Boulter 2007, chap. 2).

³¹⁰ Voir (Boulter 2007, 159) pour le traitement de la question du libre arbitre. Pour l'affirmation de la prééminence du Sens Commun, voir notamment (Boulter 2007, 21–22) : « if a common sense belief clashes with a philosophical theory or argument, the common sense belief is always given the benefit of the doubt. That is, it is always the philosophical theory or argument which must give way, and the common sense belief preserved. »

³¹¹ Notez que cette méthodologie en tant que telle n'implique aucune thèse particulière sur le scepticisme. De fait, dans ce travail, mon propos est seulement conditionnel : si le scepticisme est faux pour la perception, alors il est faux également pour le libertarianisme.

Un troisième élément que je défends et qu'on pourrait être tenté de rattacher à la tradition du Sens Commun (quoique de manière très lâche) est le « phénoménalisme charitable » qui attribue une justification *prima facie* à toute expérience doxastique. Ces trois éléments ensemble permettent de dire que nous sommes *prima facie* justifiés à croire que nous sommes libres et que nous le savons... mais ils ne disent *rien* sur la résistance de ces justifications *prima facie* contre l'attaque éventuelle d'un problème d'intégration. Pour cela, il faudrait admettre le principe de priorité qu'on vient de voir :

Prééminence des intuitions du Sens Commun : les intuitions du Sens Commun ont un privilège spécial leur permettant d'avoir le dessus contre d'autres données de force intuitive égale voire supérieure.

Et c'est précisément sur cette thèse forte que ma démarche se sépare de la tradition du Sens Commun. La « défense de Sens Commun » que je développe ici en faveur du libre arbitre ne repose pas sur l'idée d'un privilège *qualitatif* des intuitions de Sens Commun, mais sur l'idée qu'une prise en compte réaliste (non-sceptique) de la *structure* de nos justifications donne à l'expérience du libre arbitre une position initiale assez solide épistémiquement, à savoir une méta-justification *prima facie*.

Ma solution d'arbitrage consisterait donc à n'accepter aucun principe de priorité, ni de la raison ni du sens commun, et donc d'appliquer un strict « droit du plus fort ». Et comme Peter van Inwagen, je ne pourrais faire mieux que de donner l'ordre des intuitions selon la force intuitive qu'elles ont *pour moi*. Or pour moi, l'intuition (IM) a clairement une très grande force intuitive, plus forte que l'intuition incompatibiliste (IA). Par ailleurs, je n'ai *pas* l'intuition d'une inconnaissabilité de l'indéterminisme (c'est pourquoi je n'ai pas besoin d'adopter une solution « réactive » au problème d'intégration), mais *si* j'étais amené à l'avoir, je suppose qu'elle reposerait sur deux prémisses :

(1) Une croyance indéterministe (dérivée de notre expérience du libre arbitre) ne peut pas satisfaire les conditions XYZ.

(2) Les conditions XYZ sont nécessaires pour éviter le hasard épistémique.

(AH) Une croyance épistémiquement hasardeuse ne peut pas constituer une connaissance.

Donc (IE) notre expérience commune ne peut pas délivrer une *connaissance* de l'indéterminisme.

Parmi ces prémisses, je suppose qu'un argument bien construit rendrait (1) à peu près indiscutable (trivialement analytique). De même, je tiens pour évidente et irréfutable l'idée qu'un certain type de hasard est incompatible avec la connaissance. En revanche, l'analyse philosophique de la bonne notion de hasard épistémique est clairement une question épineuse, et donc la prémisse (2) serait toujours sujette à caution ; elle serait certainement plus faible que (1) et (AH), et à mes yeux plus faible également que (IM). En revanche, je ne suis pas si catégorique sur la comparaison de (2) et de (IA) : je considère comme possible, au moins théoriquement, qu'on me présente une théorie du hasard épistémique qui soit pour moi bien plus convaincante que les arguments incompatibilistes (et amène à la conclusion que toute croyance indéterministe liée à l'expérience du libre arbitre serait hasardeuse). Dans ce cas, ma solution de repli serait le compatibilisme d'ajustement. Dans le cas contraire, si l'analyse du hasard se révélait moins intuitive que les arguments incompatibilistes, j'adopterais une

³¹² Restent deux possibilités, toutes deux présentes dans la tradition du Sens Commun, la métajustification naturaliste et la métajustification par expérience mooréenne.

solution « d'intégration forcée », c'est-à-dire que j'abandonnerais cette intuition *prima facie* justifiée pour préserver la connaissance du libre arbitre. Ma solution d'arbitrage pourrait donc se résumer dans le tableau suivant :

(AH) condition anti-hasard	
(1) non satisfaisabilité de certaines conditions XYZ bien définies par le libertarianisme	
(IM) nous savons que nous sommes libres	
(2) analyse du hasard épistémique par XYZ	(IA) incompatibilisme
/	\
intégration forcée	compatibilisme d'ajustement

Le fait que je sois ouvert, en principe, au compatibilisme d'ajustement est une précision qui sera utile pour bien appréhender l'argumentation incompatibiliste du chapitre suivant : je ne tiens pas ces arguments pour si évidents qu'ils puissent m'amener à rejeter le libre arbitre. A la différence de John Fischer, mais à l'instar de Peter van Inwagen, je suis prêt à « retourner ma veste métaphysique » au cas où un argument à venir me forcerait à réviser mes croyances. La seule différence ici avec la position de van Inwagen, c'est que ce dernier a envisagé la possibilité de devenir compatibiliste si le *déterminisme* était finalement prouvé scientifiquement, tandis que je considère ici la possibilité de devenir compatibiliste si l'on me prouvait *l'inconnaissabilité* d'une liberté indéterministe à partir de notre expérience commune d'être libre.

Mais ceci ne constituerait qu'une solution de repli *au cas où* l'inconnaissabilité du libre arbitre (libertarien) serait établie. En l'état actuel de nos connaissances, je soutiens que notre expérience E a toutes les raisons d'avoir des conditions de vérité incompatibilistes (chapitre 4), que ce contenu incompatibiliste n'est pas réfuté de manière directe (chapitre 5), et que ce contenu ne nous oblige pas non plus à abandonner l'idée que notre expérience du libre arbitre confère une *connaissance* (chapitre 6).

DEUXIÈME PARTIE :

ABSENCE DE RÉFUTATIONS

Samuel Johnson famously said, ‘We *know* our will is free, and there’s an end on’t.’ Most of us share this opinion, perhaps, but we are deeply divided about what it is that we claim to know. And that is where the philosophical conversation begins.

(Watson 2003, 1)

Chapitre 4 : Le contenu incompatibiliste de notre expérience du libre arbitre

Now, are these two data - the belief that I do sometimes deliberate, and the belief that it is sometimes up to me what I do - consistent with the metaphysical theory of determinism? We do not know yet. We intend to find out. It is fairly clear, however, that they are going to present difficulties to that theory. (Taylor 1991, 42)

It is indeed obvious that if we are allowed to give the word 'freedom' any meaning that we please, we can find a meaning that will reconcile it with determinism: but this is no more a solution of our present problem than the fact that the word 'horse' could be arbitrarily used to mean what is ordinarily meant by 'sparrow' is a proof that horses have wings. (Ayer 1954, 6)

I.	Le conflit conceptuel comme meilleure explication du conflit originaire	251
II.	Comment être (et comment ne pas être) compatibiliste	255
II.1.	L'attitude naturaliste : non pertinence de la question factuelle.....	257
II.2.	Frankfurt et le semi-compatibilisme : non pertinence de la question morale	262
II.3.	Dennett, ou l'Optimisme : non pertinence de la question de valeur	264
II.4.	P. F. Strawson, ou le Quiétisme : non pertinence de la question de résilience	268
II.5.	Vargas, ou le Révisionnisme : non pertinence des pratiques linguistiques...	273
II.6.	Le compatibilisme comme Théorie de l'Illusion	277
III.	Echec des théories compatibilistes du conflit illusoire	280
III.1.	La théorie classique : l'erreur de catégorie de la « volonté libre »	280
III.2.	La théorie modale de Robert Audi : confusion de la nécessité du conséquent	287
III.3.	Dennett et les « grands méchants loups » (Bugbears) : confusions sur la notion de déterminisme	294
III.4.	Bilan des théories de l'illusion.....	304
IV.	Analyse phénoménologique de notre expérience du libre arbitre	307
IV.1.	L'expérience phénoménale agentive comme origine du concept de libre arbitre	308
IV.1.a.	Caractéristiques de l'expérience phénoménale agentive	308
IV.1.b.	Arguments en faveur du modèle phénoménal	312
IV.2.	Le contenu de l'expérience phénoménale agentive	317
IV.2.a.	Préliminaires méthodologiques.....	318
IV.2.b.	Les différentes composantes de la phénoménologie agentive.....	320
IV.2.c.	L'expérience de la Source et des Possibilités Alternatives dans le Choix	325
IV.2.d.	La portée incompatibiliste	336
V.	Conclusion.....	344

Dans la première partie de cette étude, je me suis intéressé au rôle épistémologique que l'on peut accorder à notre expérience du libre arbitre *quel que soit le contenu* exact de cette expérience, et j'ai défendu que, en amont de toute analyse conceptuelle de ce contenu, notre expérience du libre arbitre conférerait certainement une justification à accepter ce contenu (c'est-à-dire à croire que « nous avons le libre arbitre ») et que nous étions par ailleurs justifiés à tenir cette expérience pour une source de connaissance. Mais ces justifications pourraient fort bien se trouver réfutées par l'analyse ultérieure du contenu conceptuel de notre expérience E. Si cette expérience s'avère avoir un contenu logiquement contradictoire ou scientifiquement réfuté, alors notre justification expérientielle à croire au libre arbitre se trouvera elle-même réfutée de manière directe. Et même si le contenu de E se révèle à l'analyse être logiquement possible et scientifiquement non réfuté, il se pourrait encore qu'il soit absolument inconnaissable, ou inconnaissable sur la base de notre expérience E. Auquel cas notre justification initiale à croire « que E est une source de connaissance » se trouverait réfutée, et notre justification à croire le contenu de E se trouverait également réfutée, de manière fondationnelle. La deuxième partie de cette étude sera consacrée à l'examen de ces éventuelles réfutations. Et pour voir si l'analyse conceptuelle de l'expérience E apporte une réfutation, la première étape est évidemment de voir ce que révèle cette analyse conceptuelle.

Dans ce chapitre 4, je soutiendrai que l'analyse conceptuelle la plus probable de notre expérience E est l'analyse incompatibiliste. Dans la discussion traditionnelle du libre arbitre, ce chapitre correspondra donc à cette étape obligée qu'est l'opposition « arguments compatibilistes vs arguments incompatibilistes », et on peut dire que je soutiendrai la « thèse incompatibiliste ». Mais l'approche adoptée dans cette étude, c'est-à-dire l'approche centrée sur *l'expérience* du libre arbitre, donnera à cette discussion un tour assez différent de la discussion classique. Les principaux arguments incompatibilistes réapparaîtront, mais dans un rôle dialectique différent de leur rôle habituel. Du côté compatibiliste, certaines grandes thèses parfois présentées comme compatibilistes seront rejetées comme simplement non pertinentes pour la discussion. Et en fin de compte, « l'argument » incompatibiliste qui sera au cœur de ce chapitre ne sera pas une déduction à partir de principes métaphysiques sur l'action, mais une inférence à la meilleure explication à partir de données expérimentales ou phénoménologiques sur notre expérience.

Mon plan sera le suivant. Je soutiendrai tout d'abord que le conflit intuitif originaire entre déterminisme et expérience du libre arbitre, s'il n'implique pas *logiquement* que le concept de libre arbitre soit incompatibiliste, donne néanmoins une justification *prima facie* à le croire, tout simplement parce que cela constituerait l'explication *la plus simple* de l'origine du conflit intuitif. Dans une deuxième section, j'examinerai ce qu'un compatibiliste doit faire pour renverser cette justification *prima facie* en faveur de l'incompatibilisme. C'est-à-dire que je dresserai le cahier des charges pour être un vrai compatibiliste, et j'éliminerai au passage un ensemble de stratégies « compatibilistes » qui ne remplissent pas ce cahier des charges. La grande tâche que le compatibilisme doit remplir, c'est d'offrir une théorie de l'origine du conflit intuitif qui soit, en elle-même, une explication plus simple que ne l'est l'explication par une véritable incompatibilité conceptuelle. Dans la troisième section, j'examinerai les stratégies compatibilistes qui ont essayé de remplir ce cahier des charges ; j'examinerai les différentes « théories de l'illusion incompatibiliste » proposées par des auteurs compatibilistes, et je conclurai qu'aucune de ces théories n'est en elle-même plus simple et plus plausible que l'explication du conflit par une véritable incompatibilité conceptuelle. La conclusion de ces trois premières sections sera donc que, par inférence à la meilleure explication du conflit intuitif originaire, nous avons toutes les raisons de considérer que notre expérience du libre arbitre contient un concept incompatible avec le déterminisme. Dans la quatrième et dernière section, j'envisagerai la possibilité de tester cette hypothèse incompatibiliste de manière plus directe : ne peut-on pas analyser *directement* notre concept de libre arbitre plutôt que de faire

des abductions à son sujet ? Pour cela, il faudra adopter une position sur l'origine de notre expérience doxastique d'être libre. A partir notamment de données de X-Phi, je défendrai le modèle phénoménal, c'est-à-dire l'idée que notre expérience doxastique d'être libres provient de notre expérience agentive en première personne. A partir de là, il sera possible d'adopter une démarche phénoménologique à propos de cette expérience agentive, et je présenterai certains résultats récents en phénoménologie agentive, qui renforcent l'idée que nous avons une expérience phénoménale réellement incompatibiliste.

Une remarque préliminaire est fondamentale pour bien lire ce chapitre. La thèse qui y sera défendue n'est *pas* une thèse inconditionnellement incompatibiliste. Ce que je soutiendrai, c'est seulement que *si on aborde la question de compatibilité en elle-même*, c'est-à-dire en particulier si on l'aborde indépendamment du problème d'Intégration du Libre Arbitre, alors la thèse la plus plausible est la thèse incompatibiliste, principalement parce qu'elle offre l'explication la plus simple du conflit intuitif (et d'autres observations phénoménologiques sur notre expérience agentive). Cette thèse conditionnelle pourrait donc être conciliée avec une certaine forme de compatibilisme, que j'ai appelée à la fin du chapitre précédent (p.240) « compatibilisme d'ajustement ». C'est-à-dire qu'on pourrait accepter de faire le raisonnement suivant : « certes, si on considère la question de compatibilité en elle-même, alors la conclusion la plus plausible est une conclusion incompatibiliste ; mais un libre arbitre incompatibiliste nous serait totalement inconnaissable, et il est absolument non négociable que nous sommes libres et que nous le savons ; donc cela veut dire que, toutes choses considérées, la bonne réponse à la question de compatibilité doit être l'une ou l'autre des théories compatibilistes, même si cette théorie, en elle-même, offre du conflit originaire une moins bonne explication (une explication plus complexe) que l'explication incompatibiliste ».

Dans ce chapitre, je ne m'attaquerai pas à ce type de compatibilisme. De fait, ce type de stratégie est à mes yeux une solution de repli tout à fait raisonnable si le problème d'Intégration s'avérait insoluble. (C'est donc au chapitre 6 principalement que j'aurai l'occasion d'exprimer mon désaccord avec le compatibilisme d'ajustement, puisque c'est dans ce chapitre que j'argumenterai que le problème d'Intégration n'est *pas* insoluble.) Ma cible dans le présent chapitre sera seulement le « compatibilisme indépendamment motivé », c'est-à-dire une théorie qui prétend que, indépendamment de toute motivation à préserver notre croyance au libre arbitre, et à préserver son statut épistémique, l'explication la plus simple et la plus plausible du conflit intuitif originaire est une solution qui en fait un conflit illusoire reposant sur un concept fondamentalement compatible avec le déterminisme.

Pourquoi séparer ainsi le compatibilisme d'ajustement et le compatibilisme indépendamment motivé ? Mon intuition est la suivante : il me semble que la plupart des auteurs compatibilistes seraient prêts à reconnaître la faiblesse du « compatibilisme indépendamment motivé ». C'est-à-dire qu'un bon nombre d'auteurs compatibilistes ont (à mon sens) pour motivation fondamentale de préserver notre croyance au libre arbitre, et de préserver son statut de connaissance. Si tel est le cas, alors il sera toujours vain d'essayer de leur apporter de nouveaux arguments ou de nouvelles justifications en faveur de l'incompatibilisme : dans leur structure épistémique, de tels auteurs auront habituellement tout autant de justifications *prima facie* incompatibilistes que le plus incompatibiliste des auteurs. Pour convaincre un compatibiliste d'ajustement, il faudra *seulement* lui montrer la possibilité d'une intégration épistémique du libertarianisme. L'intention de ma démarche est donc de suivre au plus près « l'ordre des raisons » qui est commun aux incompatibilistes et à la plupart des compatibilistes,³¹³ jusqu'au point qui fait leur véritable désaccord, à savoir le problème d'intégration.

³¹³ Si du moins la plupart des compatibilistes sont effectivement des compatibilistes d'ajustement.

I. Le conflit conceptuel comme meilleure explication du conflit originaire

Dans cette section, je vais soutenir que nous avons une forte justification *prima facie* en faveur de l'incompatibilisme. Dans un vocabulaire moins épistémologique, cette idée est souvent formulée en disant que la position initiale ou « par défaut » est l'incompatibilisme, ou que « la charge de la preuve » revient aux compatibilistes. La suite du chapitre consistera à défendre que les compatibilistes n'ont pas réussi à relever cette « charge de la preuve ». Mais un certain nombre de compatibilistes seront sans doute en désaccord dès le point de départ : ils refuseront de prendre sur eux la charge de la preuve. En l'absence d'un argument incompatibiliste absolument convaincant, pourquoi ne pourrait-on pas être compatibiliste ? Ne faut-il pas considérer au contraire que « c'est aux incompatibilistes que revient la charge argumentative » (Nahmias et al. 2006, 30) ? Ce genre de débat pour savoir « qui a la charge de la preuve » est fréquent en philosophie, et malheureusement bien difficile à arbitrer ; un des problèmes est sans doute le fait que la notion même de « charge de la preuve » n'est pas clairement définie, et constitue parfois un simple moyen rhétorique de renvoyer la balle à l'adversaire. Pour éviter ce genre de situation arbitraire, il faudra clarifier cette notion de « charge de la preuve », et revenir aux *arguments* précis qui sont apportés pour placer la charge de la preuve ici ou là. Je commencerai par l'argument des compatibilistes qui souhaitent placer la charge de la preuve sur les épaules des incompatibilistes.

Un certain nombre d'auteurs compatibilistes estiment que la charge de la preuve, dans ce débat, revient aux incompatibilistes. Le raisonnement qui sous-tend ce genre d'affirmations est, je crois, celui explicité notamment par (Nahmias et al. 2006, 30), et qu'on pourrait appeler l'argument de la « compatibilité par défaut ». Le principe de cet argument est très simple :

Principe de Compatibilité par défaut :

Pour deux concepts C1 et C2, quels qu'ils soient, on est fondé à présumer leur compatibilité si l'on n'a pas de raison positive de croire à leur incompatibilité.

Et ce principe de compatibilité conceptuelle est appuyé, par Nahmias et al., sur un principe modal :

Principe de Possibilité par défaut :

On est fondé à présumer que quelque chose est *possible* si l'on n'a pas de raison positive de croire qu'il est *impossible*.

Dans la mesure où une thèse de compatibilité est une thèse de possibilité, ce deuxième principe permet clairement de dériver le premier. Et par simple application du premier principe, on peut effectivement conclure que la position par défaut, à propos du concept de libre arbitre et du concept de déterminisme, est la thèse de leur compatibilité.³¹⁴

³¹⁴ Voici l'argument tel que le formulent (Nahmias et al. 2006, 30) : « incompatibilism about any two concepts is not the default view. As William Lycan explains, "A theorist who maintains of something that is not obviously impossible that nonetheless that thing is impossible owes us an argument" (2003: 109). Either determinism obviously precludes free will or those who maintain that it does should offer an explanation as to why it does. The *philosophical* conception of determinism—i.e., that the laws of nature and state of the universe at one time entail the state of the universe at later times—has no obvious conceptual or logical bearing on human freedom and responsibility. So, by claiming that determinism *necessarily* precludes the existence of free will, incompatibilists thereby assume the argumentative burden. ».

Que penser de cet argument ? Permet-il effectivement de placer la charge de la preuve sur les épaules des incompatibilistes, en dépit du conflit intuitif ? A mon sens, l'argument est bon, mais la conclusion qu'il permet d'établir, une fois qu'on la comprend bien, ne remet absolument pas en cause l'idée d'une « présomption d'incompatibilisme » telle qu'elle est présente chez de nombreux auteurs incompatibilistes. En effet, il est très probable qu'en l'absence de toute raison de croire à l'incompatibilité entre déterminisme et libre arbitre, nous serions fondés à présumer leur compatibilité. Donc, si on se situe au degré zéro de la discussion, on peut sans doute dire que la « charge de la preuve » revient à la thèse incompatibiliste, dans le sens où l'incompatibilisme doit reposer sur une certaine donnée qui penche en sa faveur (notez que cette donnée n'a pas à être une « preuve » si l'on entend par là une forme particulièrement explicite et contraignante de donnée). Mais *il y a* une telle donnée : il y a justement le conflit intuitif rapporté au chapitre 1 (p.63 sqq). Et ce que défendent la plupart des incompatibilistes, c'est que *cette* donnée renvoie la charge de la preuve sur les épaules des compatibilistes. Il n'y a donc aucune incompatibilité entre les deux idées. Les compatibilistes ont raison de dire qu'au degré zéro des données, il y a une présomption de compatibilité ; mais cela n'exclut aucunement qu'au « degré un » – lorsqu'on prend en considération des faits d'observation commune, mais pas encore les données de la discussion philosophique – la « charge de la preuve » se trouve renvoyée au compatibiliste. Il est d'ailleurs intéressant de noter que Nahmias et al. ne contestent nullement ce point : Nahmias et al. présentent leur argument de « présomption compatibiliste » comme une *raison* pour l'incompatibiliste de rechercher pour leur thèse un « renfort dans les intuitions ordinaires ». En l'absence d'un tel renfort, la position par défaut resterait le compatibilisme. Mais en *présence* d'un tel renfort, voici ce que conclurait Nahmias :

Si une théorie philosophique s'avère privilégiée par les croyances populaires, cela semble poser la charge de la preuve sur les épaules de ceux qui argumentent contre les intuitions populaires. S'il s'avère qu'une majorité significative des gens produit des jugements qui renforcent soit le compatibilisme soit l'incompatibilisme, cela donnerait au moins des « droits de premier occupant » (*squatters' rights*) à la position qui a un tel renfort. (Nahmias et al. 2005, 564)

Il n'y a donc aucun problème à accepter, avec Nahmias et al., l'intuition d'une « présomption de compatibilité » au degré zéro de la discussion philosophique. La vraie question est simplement de savoir si cette présomption est ensuite renversée par le conflit intuitif naturel. C'est-à-dire que la question est de savoir si le conflit intuitif naturel apporte une justification à croire à l'incompatibilisme.

Il me semble difficilement contestable que le conflit intuitif naturel apporte au moins un certain degré de justification *prima facie* en faveur de l'incompatibilisme. On peut discuter de la force de cette justification, et de son éventuelle réfutation *ultima facie*, mais si on reconnaît l'existence de ce conflit intuitif naturel, il est difficile de dire qu'il n'existe aucune sorte de donnée en faveur de l'incompatibilisme, et que le simple principe général de présomption de compatibilité conceptuelle continue de s'appliquer. Il y a deux manières distinctes dont le conflit intuitif naturel peut apporter une justification initiale en faveur de l'incompatibilisme, l'une par justification basique, l'autre par inférence à la meilleure explication. Pour voir l'une et l'autre, je vais rappeler brièvement en quoi consiste le conflit intuitif naturel.

Les deux données empiriques les plus importantes, que nous avons rapportées au chapitre 1, sont la réponse à la « question d'actualité » (Entre l'univers déterministe A et l'univers indéterministe B, lequel est, d'après vous, le plus semblable au nôtre ? Voir p.66) et la réponse à la « question abstraite » (Dans l'univers déterministe A, est-il possible que les

gens fassent des choix vraiment libres ? Voir p.66, et la note 91). Dans l'un et l'autre cas, on observe une tendance inter-culturelle à donner majoritairement la réponse incompatibiliste : notre monde ne ressemble pas à l'univers déterministe A, et dans un tel univers, il ne serait pas possible que les gens fassent des choix vraiment libres. Quelle que soit l'interprétation que l'on donne de ces résultats, ils constituent manifestement un « conflit intuitif naturel » entre la description de l'univers déterministe A et notre conception de nous-mêmes et de nos actions. Une divergence d'interprétation que j'ai eu l'occasion de discuter concerne spécifiquement la question abstraite : certains auteurs ont considéré qu'en répondant à la question abstraite, les gens ne rapportaient pas vraiment un *jugement* appliquant le concept « libre », mais qu'ils rapportaient en fait leur *théorie* à propos de ce concept. Si cela est vrai, alors le résultat de la question abstraite fait apparaître une véritable « théorie naturelle incompatibiliste », ou en tout cas une propension naturelle à former une théorie incompatibiliste lorsque la question est posée ; pour reprendre le vocabulaire de l'expérience doxastique, cela voudrait donc dire qu'il y a une expérience doxastique naturelle en faveur de l'incompatibilisme. Et dans ce cas, nous avons à notre disposition une première sorte de justification initiale en faveur de l'incompatibilisme.

Si l'on reprend la thèse du « conservatisme phénoménal charitable » de Michael Huemer (voir chapitre 2 p.172), toute expérience doxastique, quelle qu'elle soit, apporte à son contenu un certain degré de justification *prima facie*. Si donc il y a en nous naturellement une expérience doxastique selon laquelle le concept de libre arbitre est incompatible avec le déterminisme, alors il y a aussi en nous naturellement une justification *prima facie* en faveur de l'incompatibilisme. Cette justification *prima facie* aura deux caractéristiques importantes. D'une part, ce sera une justification *basique*, et non pas une justification fondée sur une « preuve » au sens inférentiel. D'autre part, ce sera une justification disponible pour tout le monde ou presque, c'est-à-dire que chaque individu qui aura pour lui-même l'expérience doxastique naturelle verra *de facto* la charge de la preuve retomber (pour lui) du côté de la thèse compatibiliste. Telle est en tout cas la première manière dont le conflit intuitif naturel pourrait fournir une justification initiale en faveur de l'incompatibilisme. Il est important de ne pas oublier cependant que cette théorie repose sur une certaine interprétation de la réponse à la question abstraite, l'interprétation selon laquelle cette question abstraite invite les gens à donner une *théorie* à propos du concept de libre arbitre et non pas à produire un *jugement* appliquant ce concept à un cas général. Si la question abstraite ne révèle rien d'autre qu'un jugement général, alors l'expérience doxastique qui est ainsi manifestée ne sera pas une expérience doxastique de second ordre, ou *à propos d'un concept* (de la forme : « le concept de libre arbitre est incompatible avec le concept de l'univers déterministe A ») mais une expérience doxastique de premier ordre (de la forme : « les gens dans l'univers A ne pourraient pas faire des choix libres »). Dans ce cas, il n'y a pas d'expérience doxastique *directement* en faveur de l'incompatibilisme, et il faut donc passer par un argument du type Inférence à la meilleure explication.

L'inférence à la meilleure explication (IME) en faveur de l'incompatibilisme repose sur une idée très simple. Supposez que j'observe une espèce rationnelle dont je ne connais pas le vocabulaire conceptuel. J'observe que, dans cette espèce, une large majorité d'individus sont d'accord pour dire que « les F ne peuvent pas être des G ». Bien que j'ignore les concepts de F et de G, il semble que j'aie à ma disposition une explication possible et très simple d'un tel consensus : les concepts de F et de G utilisés par cette espèce sont incompatibles l'un avec l'autre. Bien sûr, avec de plus amples observations, je *pourrais* être amené à réviser cette hypothèse. L'idée est seulement que cette hypothèse est disponible dans ce genre de circonstances, et en outre qu'il s'agit d'une explication extrêmement simple et efficace du fait observé. Ainsi, lorsque nous observons qu'il y a, de manière inter-culturelle et stable, une propension à juger que « les gens dans l'univers déterministe A ne pourraient pas faire des

choix libres », une explication simple et efficace est disponible : le concept de l'univers déterministe A est incompatible avec le concept d'un choix libre. En attente de plus amples informations, cette explication est manifestement la plus simple, peut-être même la seule. Aussi, l'observation du conflit intuitif naturel nous apporte une justification par IME en faveur de l'incompatibilisme. A la différence du modèle précédent, cette justification n'est pas basique : elle repose sur une inférence. Par ailleurs, elle n'est pas disponible à toute personne qui dispose en première personne des intuitions naturelles : elle n'est disponible qu'à celui qui a pu observer le consensus d'intuitions. Enfin, cette justification fonctionne quelle que soit l'interprétation qu'on adopte pour la question abstraite : que cette question révèle un simple jugement de premier ordre, ou qu'elle révèle une intuition d'incompatibilité conceptuelle, dans les deux cas, elle fait apparaître un *explanandum* dont une explication extrêmement simple et immédiatement disponible est l'incompatibilité réelle des deux concepts.

Pour ma part, je m'appuierai ici simplement sur la justification par IME, et non pas sur la justification basique qui suppose l'existence d'une intuition commune (de second ordre) proprement incompatibiliste. Il ne me semble pas du tout impossible qu'il y ait effectivement une telle intuition, dès le niveau de la réflexion commune³¹⁵ ; mais cela ne me semble pas non plus évident, ni dialectiquement nécessaire. Ce choix a une conséquence directe sur mon plan général. En effet, comme je l'ai expliqué à la fin du chapitre précédent, mon plan général a pour but de poser dans la première partie le « point de départ épistémologique » de la discussion, c'est-à-dire la situation justificationnelle d'un sujet qui n'aurait pas encore commencé l'investigation philosophique et empirique. Un tel sujet, ai-je défendu, sera justifié à croire au libre arbitre (JLA), et à croire qu'il connaît l'existence du libre arbitre par l'expérience qu'il en a (JKLA). Si l'on admet l'existence d'une intuition naturelle proprement incompatibiliste, alors il faudra compléter le tableau de la situation initiale en disant que nous avons en outre une justification pré-philosophique (et pré-empirique) à croire à l'incompatibilisme. Autrement dit, une première analyse incompatibiliste devrait se retrouver dès ma première partie. Dans la mesure où je reste neutre sur l'existence d'une telle intuition, en m'appuyant plutôt sur une Inférence à la Meilleure Explication, je ne présenterai pas ici la « présomption incompatibiliste » comme quelque chose qui est disponible à tout un chacun avant toute investigation philosophique ou empirique. La présomption incompatibiliste est plutôt, à mes yeux, une hypothèse qui constitue la toute première étape de l'investigation réflexive, et qui repose sur une observation minimale des intuitions des gens (d'abord une observation informelle reposant sur des lectures personnelles et des discussions avec les gens, puis une observation plus systématique en X-Phi, qui, comme on l'a vu, est venue confirmer l'observation informelle). C'est pour cette raison que je la situe au début de la deuxième partie, partie dont le but est de voir si les apports philosophiques et empiriques viennent remettre en cause la situation épistémique de départ (JLA et JKLA).

Ce parti pris de présentation est aussi lié au fait que le « compatibilisme d'ajustement » représente pour moi une solution de repli envisageable, ainsi que je l'ai annoncé à la fin du chapitre 3. De fait, si on considère que l'intuition incompatibiliste fait partie, au même titre que l'intuition JLA et l'intuition JKLA, de la situation initiale du Sens Commun, alors on aura tendance à ranger dans la même catégorie « sceptique » (au sens de « violant le sens commun ») aussi bien les thèses compatibilistes que les thèses niant l'existence du libre

³¹⁵ On verra en particulier, dans la quatrième section de ce chapitre, que certains résultats de X-Phi semblent bel et bien révéler une disposition naturelle à formuler proprement une *théorie* incompatibiliste à propos de notre *expérience phénoménale* agentive (lorsqu'on invite les sujets à réfléchir consciemment sur cette expérience phénoménale). Mais pour l'instant, je reste neutre sur le rôle de l'expérience agentive dans le conflit intuitif, et par conséquent ce résultat « de second ordre » à propos de l'expérience agentive n'est pas directement utilisable.

arbitre (ou sa connaissabilité). Or, à mon sens, les solutions qui rejettent absolument le libre arbitre (ou sa connaissabilité) représentent une violation du Sens Commun bien plus radicale que les solutions compatibilistes ; ces dernières me semblent préserver un certain « cœur » fondamental de notre conception de nous-mêmes, et c'est ce cœur que j'ai voulu pour ainsi dire « sanctuariser » dans la première partie.

Quand bien même le compatibilisme ne violerait pas le cœur du Sens Commun, il s'accommode mal, comme on vient de le voir, avec une observation assez élémentaire pour quiconque commence à s'intéresser à la philosophie du libre arbitre, à savoir l'observation du conflit intuitif naturel, dont la meilleure explication, au moins initialement, est qu'il y a incompatibilité entre déterminisme et libre arbitre. Voyons donc à présent s'il est possible pour un compatibiliste de relever ce défi, c'est-à-dire d'établir la vérité du compatibilisme en dépit de la charge de la preuve que pose contre lui le conflit intuitif naturel.

II. Comment être (et comment ne pas être) compatibiliste

Pour voir si le compatibiliste peut relever ce défi, je procèderai en deux étapes. Dans cette section, je présenterai les conditions nécessaires pour relever le défi. Dans la section suivante, j'examinerai si certaines théories compatibilistes ont rempli les conditions nécessaires. Et pour mieux faire comprendre, dans cette section, ce qu'un philosophe compatibiliste doit faire, il sera utile de commencer par souligner un certain nombre de choses qu'il ne *doit pas faire* (ou qui, en tout cas, ne servent pas la cause du compatibilisme à proprement parler).

Lorsqu'on aborde la (très vaste) littérature compatibiliste, on remarque immédiatement deux choses. D'une part, une profusion d'*analyses* de la liberté, toutes plus fines et détaillées les unes que les autres ; et d'autre part une assez grande diversité dans le *rôle dialectique* que ces analyses sont censées jouer. Il n'y a donc pas simplement une grande variété de concepts compatibilistes, mais il y a aussi une grande variété « d'attitudes » ou de positionnements que l'on regroupe généralement sous le label « compatibilisme ».

Pour ce qui est de la variété de *concepts compatibilistes*, j'ai donné au chapitre 1 (p.43) les exemples de l'analyse conditionnelle de Hume, et de l'analyse hiérarchique de Frankfurt. Voici à présent une classification plus complète des analyses aujourd'hui les plus influentes. Les notions compatibilistes de liberté peuvent être classées en fonction de leur capacité à faire droit à « l'intuition de contrôle », c'est-à-dire à l'intuition selon laquelle une action est libre seulement si l'agent la « contrôle » dans un certain sens. Certaines analyses défendent la compatibilité avec le déterminisme d'une forme forte de contrôle, d'autres se reportent sur une forme plus faible de contrôle, d'autres enfin abandonnent la notion de contrôle au profit d'une notion de simple « concordance » (*mesh theories*). Dans la première catégorie entrent les analyses néo-conditionalistes, héritières de Hobbes et Hume : pour un néo-conditionaliste, par exemple pour Kadri Vihvelin (2004), la liberté suppose le contrôle des alternatives (l'agent doit pouvoir faire autrement que ce qu'il fait en effet), mais la possibilité d'agir autrement s'analyse en des termes conditionnels qui sont compatibles avec le déterminisme. Dans la deuxième catégorie se trouve la stratégie de John Fischer, défendue dans (Fischer 1994) puis (Fischer and Ravizza 1998) : pour Fischer, le déterminisme n'est pas compatible avec le contrôle des alternatives (qu'il appelle « contrôle régulateur », *regulative control*), mais il est compatible en revanche avec une forme plus faible de contrôle, le « contrôle de guidage » (*guidance control*). Le contrôle de guidage, tel qu'il est défini par Fischer, consiste dans le fait que l'agent (ou plutôt son « mécanisme » d'action) est « sensible aux raisons », c'est-à-dire que son action a été produite en réponse à des considérations rationnelles et non

pas en réponse à simple stimulus psychologique, ou une suggestion subconsciente, ou toute autre forme de causation de l'action qui court-circuiterait le processus rationnel. Enfin, certains auteurs compatibilistes ont développé des notions de liberté qui ne reposent pas sur l'idée d'un contrôle de l'action, mais plutôt sur l'idée que l'action effectuée, qu'elle soit contrôlée ou non, s'est produite en « conformité » ou en « concordance » avec certaines normes. La première intuition de ce genre est qu'une action peut être légitimement attribuée à l'agent lorsqu'elle exprime son « moi profond » (*real self view*), même si l'agent n'a pas de contrôle particulier sur ce moi profond. La conception hiérarchique de Frankfurt (concordance des désirs de premier ordre par rapport aux désirs d'ordres supérieurs) reste le meilleur exemple de théorie du moi profond. Un autre type de théorie de la concordance est la « Théorie de la Raison » (*Reason View*) défendue par Susan Wolf (1990) : pour Wolf, la capacité fondamentale pour qu'un agent soit libre n'est pas la capacité d'agir en conformité avec ses désirs (subjectifs) de quelque ordre que ce soit, mais la capacité d'agir en conformité avec le Vrai et le Bien (objectifs). Pour résumer, il y a donc aujourd'hui quatre grandes familles de notions compatibilistes : le Pouvoir Conditionnel d'Agir Autrement, la Sensibilité aux Raisons, la Conception Hiérarchique, et enfin la Théorie de la Raison.

Théories du contrôle	... des alternatives	analyses (néo-)conditionnalistes <i>Hobbes, Hume, Vihvelin</i>
	... de guidage	sensibilité aux raisons <i>Fischer et Ravizza</i>
Théories de la concordance	... avec le « moi profond »	conception hiérarchique <i>Frankfurt</i>
	... avec le Bien	Théorie de la Raison <i>Wolf</i>

Comme on l'a vu au chapitre 1 (p.46), cette liste de « notions compatibilistes » ou de « types de libre arbitre compatibilistes » peut être mise en parallèle avec une liste non moins importante de « types de libre arbitre incompatibilistes ». Et la véritable question qui oppose compatibilistes et incompatibilistes ne porte donc pas sur le caractère compatibiliste ou non de telle notion (*cette* question est triviale pour tout le monde) ; la véritable question consiste à savoir *laquelle* de ces notions techniques de libre arbitre correspond à *notre* concept commun. Compte tenu de la « solution descriptiviste » que nous avons adoptée pour déterminer de quel concept il s'agit (voir p.52), le vrai débat entre compatibilistes et incompatibilistes se réduit donc à la question suivante :

(Question d'Origine du Conflit) Quel type de libre arbitre (LA1, LA2, ...) est à l'origine du conflit intuitif commun avec le déterminisme ?

Une position véritablement compatibiliste (telle que définie p.56) consiste à répondre ceci : le type de libre arbitre à l'origine du conflit intuitif commun est LAn – où LAn appartient à la liste des notions (clairement) compatibilistes. Telle est la « bonne attitude » à adopter pour être compatibiliste, la bonne façon « d'utiliser » l'une ou l'autre des notions compatibilistes. A la fin de cette section, je ferai quelques remarques plus détaillées sur cette attitude. Avant cela, je vais présenter une liste d'attitudes qui ne sont *pas* des formes de compatibilisme, mais qui sont parfois classées comme telles parce qu'elles font un usage important des « notions compatibilistes » présentées plus haut.

Le point commun des attitudes faussement compatibilistes, c'est qu'elles utilisent la liste de notions compatibilistes pour répondre à toutes sortes de questions intéressantes, mais qui n'ont pas de rapport direct avec le problème traditionnel du libre arbitre. On a vu en effet

au chapitre 1 (p.47) que la liste des « types de libre arbitre » pouvait être utilisée pour poser un ensemble de questions très variées (outre la Question d'origine du conflit qui permet de fixer « le » sens privilégié de « libre arbitre »). Pour l'économie de cette section, je retiendrai particulièrement les questions suivantes :

(Question de fait) Quels types de libre arbitre avons-nous : LA1, LA2, etc. ?

(Question de moralité) Quels types de libre arbitre sont (réellement) requis par la responsabilité morale : LA1, LA2, etc. ?

(Question de valeur) Quels types de libre arbitre ont une réelle importance ou une réelle valeur pour nous : LA1, LA2, etc ?

Dans les trois premières sous-sections, je présenterai les attitudes qui consistent à répondre à ces questions par une « notion compatibiliste », en soulignant qu'il ne s'agit pas, en fait, d'une forme de compatibilisme. Dans les deux sous-sections suivantes, je présenterai deux autres attitudes faussement compatibilistes qui répondent à des questions plus subtiles mais tout aussi indépendantes du « problème traditionnel », tel que l'ai défini au chapitre 1. Enfin dans la dernière sous-section, je montrerai de manière plus précise ce que *doit* être la vraie attitude compatibiliste. Les premières sous-sections, qui montrent négativement ce que le compatibilisme n'est pas, sembleront peut-être superflues, et pourront effectivement être omises, pour les lecteurs qui sont d'ores et déjà en accord avec ma caractérisation du compatibilisme ; seulement, étant donné l'importance des positions de John Fischer, de Frankfurt ou de P.F. Strawson (par exemple) dans les catalogues compatibilistes habituels, il m'a semblé nécessaire d'expliquer brièvement pourquoi ces positions ne sont pas ici pertinentes.

II.1. L'attitude naturaliste : non pertinence de la question factuelle

Une première attitude, qui est parfois prise pour une forme de compatibilisme, pourrait être appelée « l'attitude naturaliste ». L'attitude naturaliste provient des développements considérables réalisés à la fois en neuro-sciences et en psychologie expérimentale depuis une trentaine d'années, sous le label unifié des jeunes « sciences cognitives ». L'étude empirique de l'action humaine et du contrôle volontaire est un domaine des sciences cognitives qui est lui-même en expansion depuis une quinzaine d'années, comme en témoignent un certain nombre de conférences internationales aujourd'hui imprimées. Trois volumes particulièrement pertinents sont : *Agency and Self-Awareness* (Roessler and Eilan 2003), *Disorders of Volition* (Sebanz and Prinz 2006), et *Does Consciousness Cause Behavior?* (Pockett, Banks, and Gallagher 2006). Cette approche naturaliste est également développée de manière systématique dans la monographie de Joëlle Proust (2005), qui offre une remarquable synthèse philosophique de résultats en sciences cognitives de la volonté.³¹⁶

De manière générale, on peut caractériser l'approche naturaliste comme un effort pour fonder la philosophie de l'action ou du contrôle volontaire sur les données empiriques des sciences cognitives. L'attitude naturaliste ne se *réduit* pas à une collection des données empiriques (voir Proust 2005, 15) : la philosophie garde une position fondamentale, à la fois comme réservoir de concepts et de distinctions, et comme cadre général qui définit l'enjeu des résultats empiriques. Mais les *données* qui permettront de choisir entre telle et telle théorie de l'action seront les données des sciences cognitives, et non pas des données proprement

³¹⁶ Nous verrons cependant que Joëlle Proust a une attitude philosophique précise qui reste très prudente à l'égard du compatibilisme et ne prête donc pas le flanc à la critique opérée dans cette section.

philosophiques ou de sens commun. L'idée qui se trouve derrière ce genre d'attitude, c'est que les sciences cognitives ont apporté des données si nombreuses et si précises qu'elles ont tout bonnement supplanté les approches plus traditionnelles de la question du libre arbitre.

[L]es travaux effectués par les neurosciences, la psychologie du développement, la neuropsychologie et par la philosophie de l'esprit bouleversent le paysage classique et paraissent peu compatibles avec les présupposés ordinairement admis dans l'analyse conceptuelle de l'action. (Proust 2005, 12)

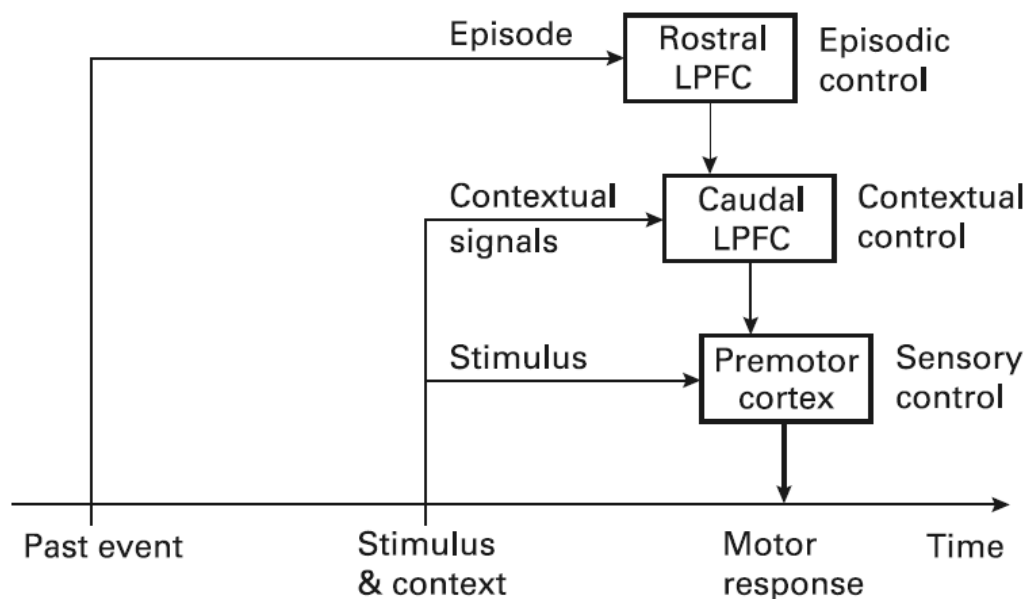
Quelles sont ces données et comment sont-elles utilisées ? Une source majeure d'information se trouve dans les dysfonctionnements du contrôle agentif, soit pathologiques soit bénins. Il est bien connu en sciences expérimentales que pour comprendre et modéliser les différentes composantes d'un système complexe (par exemple, la capacité de parler), on peut s'appuyer sur les cas où ces différentes composantes sont déficientes *séparément* (par exemple, déficience de la capacité articulatoire, de l'association mot-idée, de la composition grammaticale, etc.). Dans le cas du contrôle agentif, les études cliniques ont apporté un grand nombre de cas pathologiques qui permettent des modélisations théoriques de plus en plus fines. Une des pathologies les plus connues et les plus étudiées est celle de la « main anarchique » (Pacherie 2007) dans laquelle un sujet fait avec l'une de ses mains une certaine action (taper avec un ustensile, déboutonner sa chemise) mais en ayant l'impression que c'est la main qui agit d'elle-même et contre la volonté de l'agent. D'autres échecs de contrôle agentif sont observés dans les cas de schizophrénie (Sebanz and Prinz 2006, pt. II), de dépression (idem, pt. III), ou d'abus de substances nocives (idem, pt. V). Face à ces pathologies, une première méthode consiste donc à induire une théorie des différents « modules » de l'agentivité, en fonction des pathologies où ces modules tombent en panne séparément. Une autre méthodologie consiste à identifier grâce aux neurosciences (et particulièrement à l'imagerie cérébrale) la localisation cérébrale de ces modules. Par exemple, on a pu mettre en évidence la responsabilité du lobe pariétal inférieur dans la capacité à détecter la différence entre mouvement observé et mouvement exécuté volontairement (laquelle capacité est perturbée chez certains patients schizophrènes) (Jeannerod 2006).³¹⁷

Un exemple précis permettra de mieux comprendre l'articulation entre analyse philosophique et données empiriques dans l'attitude naturaliste. Comme on l'a vu, Harry Frankfurt a établi la distinction conceptuelle entre désirs de premier ordre (désir de faire ceci ou cela) et désirs de deuxième ordre (désir d'avoir tel ou tel désir), et a fondé sur cette distinction une analyse des différents types de concordance dans les désirs d'un agent : la concordance entre les actions d'un agent et ses désirs de premier ordre définit un certain type de « liberté » ; mais un agent dont les désirs de premier ordre s'accordent à ses désirs de deuxième ordre sera clairement doué d'un type de « liberté » plus élaboré. Chez Frankfurt, cette distinction était en outre illustrée par une expérience de pensée. Frankfurt (1971, 12–13) nous invite à imaginer deux sujets dépendants aux narcotiques. Les deux sujets ont en commun d'avoir le désir de premier ordre de consommer des narcotiques ; ce qui les distingue, c'est que le premier sujet a un désir de deuxième ordre de *ne pas* avoir ce désir de premier ordre (il est donc « dépendant contre son gré », *unwilling addict*), tandis que le second n'a aucune sorte de désirs de deuxième ordre, il agit de manière totalement irreflexive en fonction de ses désirs de premier ordre (c'est donc un « dépendant impulsif », *wanton addict*). On peut remarquer que cette expérience de pensée utilise, d'une certaine manière, la méthodologie des « cas pathologiques », mais les cas dont il s'agit sont purement imaginaires, et la distinction

³¹⁷ Pour une autre classification et une discussion des différentes formes d'échecs du contrôle agentif, voir (Roessler and Eilan 2003, 3–5).

conceptuelle entre désirs de premier et de second ordre est elle-même une distinction essentiellement logique, ouvrant la voie à une régression infinie d'ordres de désirs (désirs de troisième ordre, de quatrième ordre, etc.) Mais qu'en est-il dans la réalité ? Il est peu probable qu'on puisse identifier chez les agents humains des désirs du dix-septième ordre, mais peut-on néanmoins utiliser les outils conceptuels de Frankfurt pour identifier une hiérarchie de niveaux de contrôle agentif ? La réponse semble bien être oui.

Ainsi que l'a montré Joëlle Proust (2005, 245–248), le schéma hiérarchique de Frankfurt trouve une forme de confirmation empirique dans les travaux en imagerie cérébrale de l'équipe d'Etienne Koechlin. En observant les zones d'activités cérébrales impliquées dans l'effectuation de tâches plus ou moins complexes, (Koechlin, Ody, and Kouneiher 2003) ont permis de mettre en évidence trois véritables modules cérébraux du contrôle agentif, de niveaux hiérarchiques différents. Pour les tâches simplement *motrices* (réaction à un unique stimulus), le module de « contrôle sensoriel » se situe dans le cortex prémoteur. Pour les tâches qui impliquent la sensibilité au *contexte* (par exemple un feu vert permettant ou non la réaction au stimulus), un autre module, situé dans le cortex préfrontal latéral caudal, joue un rôle de contrôle de deuxième ordre que Koechlin et al. appellent « contrôle contextuel ». Enfin, pour les tâches qui impliquent quelque chose comme des règles du jeu définies pour un *épisode* donné, un troisième module, situé dans le cortex préfrontal latéral rostral, représente pour ainsi dire un contrôle de troisième ordre, le « contrôle épisodique ». Le schéma suivant résume cette structure hiérarchique :



Cette étude empirique permet donc de donner une application et un contenu réel à la notion frankfurtienne de niveaux hiérarchiques du contrôle agentif, laquelle notion resterait spéculative sans ces données empiriques – et inversement, les concepts de Frankfurt permettent de donner à cette étude empirique un cadre théorique et une portée philosophique qu'elles n'auraient pas à elles seules : elles permettent en effet d'intégrer ces données sur les modules de contrôle agentif à la discussion philosophique sur les degrés et les modes de la liberté humaine.

Maintenant que nous avons une idée plus nette de l'attitude naturaliste, on peut envisager son rapport avec la question du compatibilisme. A première vue, le rapport n'est pas évident. Ce que recherche la démarche naturaliste, c'est manifestement une réponse à la

question factuelle (et plus précisément une réponse tirée des sciences cognitives) : « parmi les différents concepts de libre arbitre que les philosophes ont élaborés, quels sont ceux qui trouvent dans les sciences cognitives une application réelle ou une preuve empirique d'existence ? » Or il est clair que la question factuelle est logiquement indépendante de la question du compatibilisme – c'est-à-dire de la question de savoir si le concept de libre arbitre (celui qui est en jeu dans le débat traditionnel et donc dans le conflit intuitif originel) est l'un des concepts compatibilistes ou l'un des concepts incompatibilistes. Bien sûr, il se pourrait que le concept commun de libre arbitre s'avère compatibiliste, et qu'en outre nous ayons effectivement un tel type de libre arbitre. C'est la case A dans le tableau suivant. Mais les trois autres cas de figure sont également possibles.

		le type de LA que nous avons factuellement :	
		compatibiliste	incompatibiliste
le concept de LA (celui du débat traditionnel) :	compatibiliste	A	C
	incompatibiliste	B	D

Il se pourrait (B) que le concept de sens commun soit incompatibiliste, et que nous n'ayons factuellement que des types de liberté compatibilistes. Il se pourrait (C) que le concept de sens commun soit compatibiliste, mais que nous ayons factuellement des types de liberté plus forts que ceux que nous attribue le sens commun, et en particulier des types de liberté incompatibilistes. Il se pourrait enfin (D) que notre concept de sens commun soit incompatibiliste, et que nous ayons factuellement ce type de libre arbitre incompatibiliste. Dans la mesure où elle cherche une réponse à la question factuelle, l'attitude naturaliste devrait donc être totalement indépendante de la question de compatibilité. Et de fait, on trouve un certain nombre de philosophes qui adoptent cette attitude naturaliste en philosophie du libre arbitre sans pour autant estimer qu'elle implique une forme quelconque de compatibilisme. C'est le cas de Joëlle Proust, qui rejette explicitement le compatibilisme (2005, 238–242) en s'appuyant sur un des deux grands arguments que nous verrons dans la section suivante, à savoir « l'argument de la manipulation ». Sa conclusion est claire : « [si le déterminisme] est vrai, dans ce cas les volitions, comme les autres attitudes, sont le produit causal d'autres états et ne peuvent être dits compatibles avec la liberté qu'en changeant le sens du mot « liberté » de manière à servir l'argument » (Proust 2008, 148). Certes, Joëlle Proust utilise également un concept graduel ou « relatif » de liberté, un concept selon lequel deux agents totalement déterminés peuvent être dits « plus ou moins libres » (en fonction de leurs degrés de contrôle). Mais Joëlle Proust est très claire sur le fait qu'il ne s'agit là que d'un « choix terminologique » consistant à donner au mot liberté un « sens résiduel non illusoire » (Proust 2008, 148), par opposition au « terme ordinaire, celui qui donne lieu à des intuitions fortes », ce dernier étant fondamentalement incompatibiliste et donc (pour Joëlle Proust) irréconciliable avec une théorie naturaliste de l'agent.³¹⁸

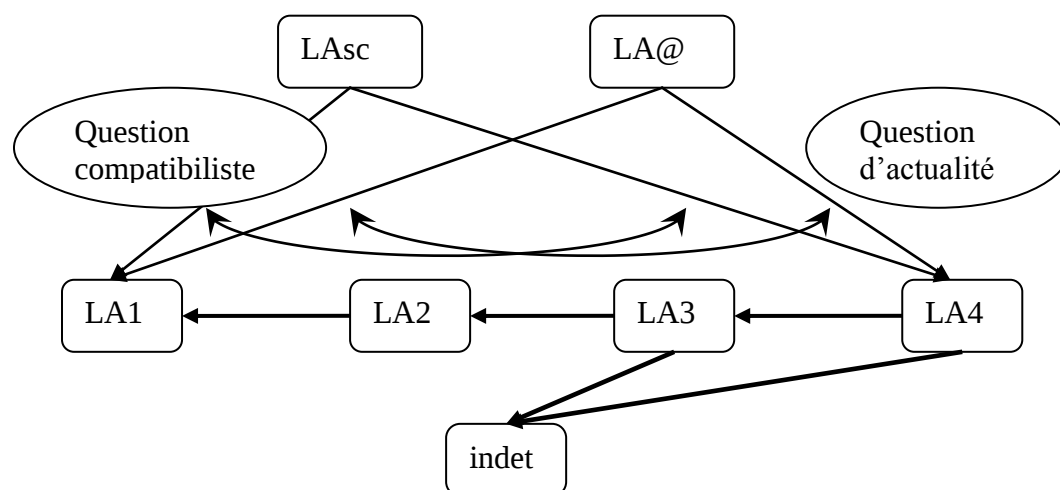
Pourtant l'attitude naturaliste est souvent présentée ou conçue comme étant associée à une forme de compatibilisme. On trouve cette idée notamment dans la critique virulente de (Prinz, Dennett, and Sebanz 2006, 2) contre les concepts incompatibilistes de libre arbitre : « Certains philosophes continueront peut-être à préférer leurs notions fantastiques de 'causalité agentive' ou d'amplificateurs quantiques, commodément cachés dans le brouillard au centre du cerveau, mais la science de la volonté est en train de les rendre absolument dépassées ». Dire que les neuro-sciences rendent *inapplicables* (ou introuvables dans la réalité) les notions incompatibilistes de libre arbitre, c'est une chose ; dire qu'elle les rendent

³¹⁸ Nous retrouverons cette idée d'un choix terminologique *a minima* dans la position révisionniste de Manuel Vargas, au point II.4.

absolument sans intérêt et que seuls les concepts compatibilistes restent significatifs, c'est une thèse plus forte que ne partagent pas, on vient de le voir, tous les naturalistes. Il me semble qu'on peut comprendre le glissement du naturalisme au compatibilisme de la manière suivante : premièrement, il est probable que les seules facultés de contrôle agentif qui pourront être mises en évidence et étudiées au moyen des sciences cognitives seront des facultés compatibles avec le déterminisme ; par conséquent, si un philosophe veut se *restreindre* aux données des sciences cognitives, le portrait du contrôle agentif qu'il pourra donner sera assurément un portrait « compatibiliste ». On en tire alors la conclusion que les conceptions incompatibilistes de l'agentivité n'auront pas leur place dans une démarche naturaliste. Mais cette conclusion est illégitime : comme on le voit chez Joëlle Proust, une conception incompatibiliste peut faire partie d'une démarche naturaliste au titre de théorie de notre *expérience* de l'action, sans constituer une théorie de notre contrôle agentif *réel*.

La démarche naturaliste est donc, en tant que telle, totalement indépendante de la question compatibiliste. A cause des considérations épistémologiques de la première partie, il devrait être clair que je n'adhère pas à cette démarche naturaliste : c'est-à-dire que j'accorde un statut justificationnel (et donc potentiellement « informatif » pour la théorie philosophique de l'agentivité) à des données autres que les données scientifiques (en particulier à notre expérience commune de l'action). Mais ce qui m'importe ici n'est pas tant de montrer que l'attitude naturaliste est insatisfaisante (ou incomplète, ou trop restrictive) ; mon but ici est de souligner qu'elle ne constitue absolument pas une forme de compatibilisme : si elle tend à restreindre son attention à *des* concepts compatibilistes, c'est seulement pour affirmer leur existence factuelle, et pas pour affirmer qu'ils décrivent adéquatement *notre* concept commun de libre arbitre.

Si l'on note LAsc pour notre concept de sens commun de libre arbitre et LA@ pour le type de libre arbitre que nous avons effectivement, on peut représenter l'indépendance des deux questions à l'aide du schéma suivant :



Le schéma se lit ainsi : sachant que certains types précis de libre arbitre (LA3 et LA4) impliquent l'indéterminisme et d'autres pas (LA1 et LA2), la question compatibiliste consiste à déterminer l'étendue d'implications de LAsc parmi ces types de libre arbitre et à voir si des concepts incompatibilistes tombent dans cette étendue. Quant à la question d'actualité, elle consiste, de manière indépendante, à connaître l'étendue d'implications de LA@, c'est-à-dire à savoir quels types de liberté (parmi LA1 à LA4) nous avons effectivement.

II.2. Frankfurt et le semi-compatibilisme : non pertinence de la question morale

Une autre attitude qui est parfois tenue pour une attitude compatibiliste consiste à défendre que, même si une certaine forme de liberté (typiquement le pouvoir d'agir autrement) requiert l'indéterminisme, la responsabilité morale quant à elle ne requiert pas d'avoir cette forme de liberté (et donc, sauf argument indépendant, ne requiert pas l'indéterminisme). Ce type d'attitude remonte à un autre article séminal de Harry Frankfurt, « Alternate Possibilities and Moral Responsibility » (Frankfurt 1969). Dans cette section, je ne pourrai évoquer que très brièvement cet article et la discussion considérable qu'il a généré, mon intention étant simplement de montrer que cette discussion n'est pas pertinente pour la question du compatibilisme proprement dit.

En simplifiant quelque peu, on peut présenter l'apport dialectique de (Frankfurt 1969) de la manière suivante. D'après Frankfurt, la discussion traditionnelle sur le compatibilisme cherchait à savoir si le type de libre arbitre requis par la responsabilité morale était compatible ou non avec le déterminisme. Pour défendre sa position, l'incompatibiliste procédait en deux étapes : (1) la responsabilité morale (pour un acte donné) suppose d'être capable d'agir autrement, et (2) le pouvoir d'agir autrement est incompatible avec le déterminisme. Les compatibilistes traditionnels ont déployé des efforts considérables pour contester la deuxième étape (notamment en proposant des analyses conditionnelles du pouvoir d'agir autrement). En s'attachant exclusivement à la deuxième étape, les compatibilistes traditionnels manifestaient leur accord implicite avec la première étape. C'est là que l'article de Frankfurt a opéré un tournant révolutionnaire : Frankfurt y propose en effet une expérience de pensée dans laquelle un agent semble moralement responsable bien qu'il n'ait pas la possibilité d'agir autrement. Si donc l'argument incompatibiliste échoue dès la première étape, les compatibilistes traditionnels ont eu tort de s'attaquer exclusivement à l'étape 2, en cherchant des interprétations complexes du pouvoir d'agir autrement : le compatibiliste peut très bien abandonner ce type de pouvoir sans pour autant abandonner le libre arbitre (dans le sens requis par la responsabilité morale).

Dans l'article de Frankfurt, le principe (1) était introduit sous le nom de « principe des possibilités alternatives » (ou PAP : « *principle of alternate possibilities* ») et formulé de la manière suivante :

Une personne n'est moralement responsable de ce qu'elle a fait que si elle pouvait [/aurait pu] faire autrement. (Frankfurt 1969, 829)

Et le contre-exemple originellement imaginé par Frankfurt était celui-ci :

Supposez que quelqu'un – Black, disons – veut que Jones accomplisse une certaine action. Black est prêt à aller très loin pour arriver à ses fins, mais il préfère éviter de s'impliquer directement si ce n'est pas nécessaire. Donc il attend jusqu'à ce que Jones soit sur le point de se décider, et il ne fait rien tant qu'il n'est pas clair pour lui (Black est un très bon juge en la matière) que Jones va décider *autre chose* que ce qu'il veut le voir faire. S'il devient clair que Jones va décider de faire autre chose, Black prend les moyens de s'assurer que Jones décide de faire, et fait effectivement, ce qu'il veut le voir faire. Quelles que soient les préférences et inclinations initiales de Jones, Black arrivera donc à ses fins. [...] Mais supposez à présent que Black n'a jamais besoin de s'impliquer directement parce que Jones, pour des raisons qui lui sont propres, décide de faire, et fait effectivement, l'action même que Black veut le voir faire. (Frankfurt 1969, 835–836)

Dans un tel scénario, il semble qu'on doive admettre à la fois (a) que Jones n'aurait pas pu faire autre chose que ce qu'il a fait (à cause de la présence de Black), et (b) que Jones est moralement responsable de son action (car Black n'a pas eu à intervenir ; Jones a agi de lui-même, exactement de la manière dont il aurait agi si Black n'avait pas été là).

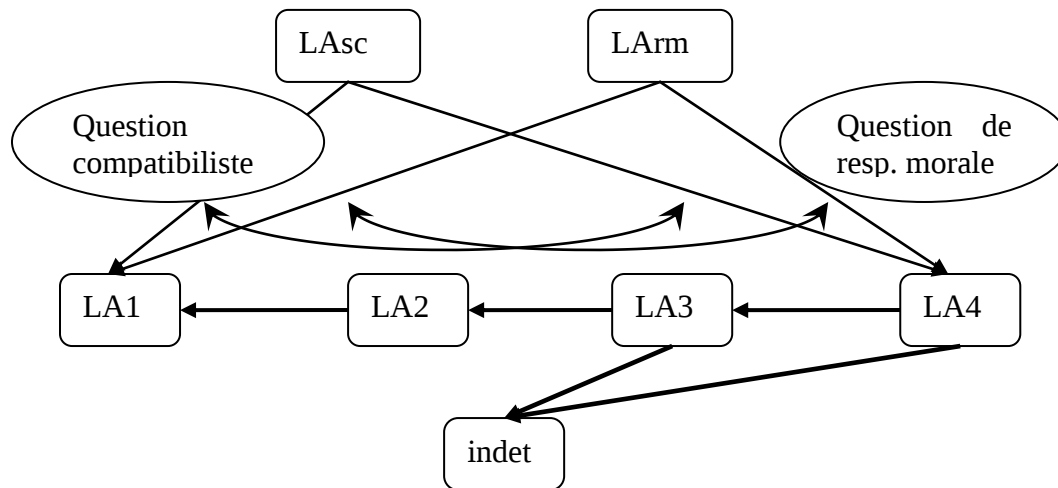
Depuis l'article de Frankfurt, la discussion sur le PAP a pris des proportions considérables ; toutes sortes d'objections ont été proposées contre la concevabilité du scénario ou contre sa valeur de contre-exemple ; et d'innombrables scénarios améliorés ont été imaginés pour répondre à ces objections. Toutefois, il est aisé de voir que ces discussions ne sont pas pertinentes pour trancher la question du compatibilisme, telle que nous l'avons posée ici : en effet, l'ensemble de ces discussions concerne le *type de libre arbitre requis par la responsabilité morale*, et non pas le concept de sens commun de libre arbitre, ou le concept « qui fait partie de l'image que nous avons de nous-mêmes et génère le conflit intuitif avec le déterminisme ». Quand bien même on concéderait toutes les conclusions de Frankfurt, il pourrait très bien se faire que la responsabilité morale (et le type de liberté qu'elle requiert) soit compatible avec le déterminisme *mais* que notre concept commun du libre arbitre exige des possibilités alternatives dans un sens incompatible avec le déterminisme. De fait, cette position est celle que nous avons déjà rencontrée au chapitre 1 (p.57) sous le nom de « semi-compatibilisme », et qui est défendue par John M. Fischer : compatibilisme pour la responsabilité morale, mais incompatibilisme pour le libre arbitre à proprement parler.³¹⁹ Dans la présentation de Fischer, cette combinaison de thèse est encore présentée comme une forme de « compatibilisme » – Fischer est même pris comme le *représentant* de la position compatibiliste dans l'ouvrage à quatre auteurs *Four views on Free Will* (Fischer et al. 2007). Mais cette présentation est trompeuse : pour être compatibiliste, il ne suffit pas de défendre que « quelque chose » ou « quelque concept » (par exemple, la responsabilité morale) est compatible avec le déterminisme, il faut que le concept de *libre arbitre* (le concept de libre arbitre) soit compatible avec le déterminisme. Dans la mesure où le semi-compatibilisme de Fischer est incompatibiliste à propos du concept expérientiel de libre arbitre, cette position n'est pas un compatibilisme du tout, c'est un incompatibilisme.³²⁰

Pour conclure, la discussion des cas de Frankfurt et du principe des possibilités alternatives est totalement indépendante de la question du compatibilisme, car elle concerne seulement la « question de moralité » et pas la « question de l'origine du conflit » (ou la

³¹⁹ Bien que Fischer défende effectivement la thèse que j'appelle ici « semi-compatibilisme », il faut noter, pour être exact sur la terminologie, que son usage du terme est légèrement différent. Pour Fischer, est semi-compatibiliste toute personne qui défend que la responsabilité morale est compatible avec le déterminisme, *indépendamment* de la question de savoir si le *pouvoir d'agir autrement* est compatible ou non avec le déterminisme. Le semi-compatibiliste (i) n'a donc pas besoin d'être engagé à l'incompatibilisme pour l'autre « moitié » du débat, et (ii) cette autre « moitié » est d'abord définie à partir du pouvoir des alternatives (le « contrôle régulateur ») et non pas à partir du concept expérientiel de libre arbitre. Cependant, (i) Fischer lui-même est incompatibiliste à propos du pouvoir des alternatives (Fischer 2006, 7; Fischer et al. 2007, 56) ; s'il est *semi*-compatibiliste, il l'est donc au sens « strict » ou « exclusif » de l'autre moitié. Et par ailleurs, (ii) si l'autre moitié concerne d'abord le pouvoir des alternatives, Fischer est très clair sur le fait que l'*intérêt* de ce concept (la *raison* d'en faire une moitié du débat, et de considérer comme incomplètement compatibilistes ceux qui resteraient compatibilistes à propos de ce concept) vient précisément du fait qu'il s'agit du concept expérientiel : « Some philosophers have claimed that semicompatibilism is merely old wine in new bottles. They have stated that [...] no one really cares (or for that matter ever has cared) about [...] regulative control. [...] For Watson (and others, especially traditional compatibilists), semicompatibilism is not so much an innovation as something always presupposed by traditional compatibilists. [...] These remarks puzzle me for various reasons. In our phenomenology as agents, it is quite natural to think that in deliberating about the future, we are selecting from among various options that are genuinely metaphysically open to us. » (Fischer et al. 2007, 72)

³²⁰ Dans le chapitre 1, j'ai présenté la position semi-compatibiliste comme étant « une des solutions » possibles au problème traditionnel du libre arbitre ; dans le contexte présent, l'important est seulement de noter qu'il ne s'agit pas de la *même* solution que la solution compatibiliste.

question de notre concept commun de libre arbitre). C'est à cette deuxième question qu'il faut apporter une réponse compatibiliste si l'on veut être compatibiliste à propos *du* libre arbitre. On peut représenter l'indépendance des deux questions sous la forme de schéma suivant (où LArm désigne le type de libre arbitre requis par la responsabilité morale).



II.3. Dennett, ou l'Optimisme : non pertinence de la question de valeur

Un autre ensemble de thèses ou de réflexions que l'on trouve fréquemment sous la plume des auteurs « compatibilistes » concerne la « question de valeur » : Quels types de libre arbitre ont une réelle importance ou une réelle valeur pour nous : LA1, LA2, etc ? Ou pour reprendre le sous-titre de (Dennett 1984a) : quelles variétés de libre arbitre vaut-il la peine de souhaiter ? (*the varieties of free will worth wanting*). Et pour Dennett, la réponse est claire : les types de libre arbitre « exigeants », ceux qui impliquent l'indéterminisme, n'ont aucune importance pour nous ; nous n'avons aucune bonne raison de souhaiter les posséder. Admettons que nous ne les avons pas, cela n'a aucune importance : « Je n'aurais pas pu agir autrement, et alors ? » (Dennett 1984b). Le corollaire de cette première thèse est une forme « d'optimisme » à propos du libre arbitre : tous les types de libre arbitre qui *ont* une importance sont des types « peu exigeants », et même très certainement des types de libre arbitre que nous avons *effectivement*.

Optimisme : nous avons toutes les sortes de libre arbitre qu'il vaut la peine de souhaiter.

Ma réaction à ce genre de thèse sera la même que dans les sections précédentes : la question de valeur est indépendante de la question compatibiliste proprement dite. Mais avant de voir cette réponse, il convient de présenter un peu plus la position « optimiste ».

Pour bien comprendre la position optimiste, il faut exposer brièvement les arguments qui y mènent : pourquoi Dennett dit-il que les formes incompatibilistes de libre arbitre n'ont aucune importance pour nous ? Dennett avance plusieurs types de raisons, sans proposer de hiérarchie bien claire entre ces raisons.

Une première raison, c'est que « la question de savoir si un agent aurait pu faire autrement n'a tout simplement aucune importance pour la responsabilité morale » (Dennett 1984b, 553). On retrouve ici le point de la section précédente, mais cette fois comme *prémisse* vers une conclusion optimiste : « donc la capacité d'agir autrement n'a aucune importance pour nous ». ³²¹ Bien qu'on puisse avoir d'*autres* raisons de s'intéresser à une certaine faculté que des raisons morales, il me semble qu'en l'occurrence, l'argument est assez fort si du moins la prémisse était vraie : s'il est vrai que la responsabilité morale ne requiert pas un libre arbitre de type indéterministe, alors cela nous enlève certainement une forte raison de *souhaiter* avoir un tel type de libre arbitre. Mais comme nous le verrons, cette conclusion même si elle était vraie n'est *pas* une forme de compatibilisme.

Une deuxième raison, étroitement associée à l'importance *morale*, c'est que la capacité d'agir autrement n'aurait pour nous aucune importance *pratique* : apprendre qu'un agent est causalement indéterminé dans ses choix « serait apprendre quelque chose de totalement inutile, du point de vue de l'évaluation de son caractère ou de la planification du futur. [...] Donc si quelqu'un s'intéresse un tant soit peu à cette question [...] ce devra être par pure curiosité métaphysique – c'est-à-dire par une curiosité si pure qu'elle n'aura aucun motif ultérieur. » (Dennett 1984b, 559). Je serais plus réservé sur ce deuxième argument : si l'on apprenait que tel agent agit de manière déterministe, et tel autre non, il semble qu'on pourrait chercher à « déterminer » le choix de l'un, tandis qu'on pourrait seulement chercher à « influencer » (probabilistiquement) le choix de l'autre. Il semble donc que notre attitude pratique à leur égard pourrait différer.

Une troisième raison de ne pas nous intéresser à la question de l'indéterminisme de nos choix libres, c'est « la pure et simple impossibilité de mener une enquête sensée sur la question » (Dennett 1984b, 556). Autrement dit : nous n'avons aucun moyen de savoir si nous avons ou pas ces types exigeants de libre arbitre. Pourquoi donc se soucier d'une chose qu'on ne pourra jamais savoir ? On retrouve ici le problème épistémologique qui est au cœur de cette étude. Dennett fait partie des auteurs qui n'accepteraient pas d'attribuer une valeur informative à notre expérience du libre arbitre si celle-ci était incompatibiliste. Ma réponse principale, contre cette raison, c'est donc de dire que notre expérience du libre arbitre constitue bel et bien une source d'information pour savoir que nous avons une forme incompatibiliste de libre arbitre. Et la défense de cette réponse est l'ensemble de cette étude. Mais plus simplement, on peut faire deux remarques à ce nouvel argument. Premièrement, à supposer que nous n'ayons aucun moyen de déterminer si nous avons oui ou non un libre arbitre indéterministe, il ne s'ensuit pas que la question n'a aucun intérêt. Bien sûr, si nous ne pouvons espérer le savoir, nous ne devons pas chercher à le savoir ; mais nous pourrions très bien *l'espérer* néanmoins. Par comparaison, imaginons que j'ai un doute sur la sincérité d'un ami qui est décédé depuis plusieurs années ; peut-être n'aurai-je aucun moyen de jamais déterminer si son amitié était sincère ; mais cela implique-t-il que cette question n'a aucune importance pour moi dans le sens où il serait irrationnel pour moi de souhaiter ou espérer que son amitié était sincère ? Il ne semble pas. De même, même si l'existence d'un libre arbitre indéterministe échappait à notre connaissance, il pourrait être rationnel de *souhaiter* avoir un tel libre arbitre. Deuxièmement, à supposer qu'il soit *irrationnel* d'attribuer de l'importance à cette question, il ne s'ensuivrait pas pour autant que cette question n'a aucune importance pour nous. En effet, il se pourrait que nous lui attribuions *irrationnellement* de l'importance. Auquel cas, l'absence d'un tel libre arbitre serait peut-être une frustration *irrationnelle*, mais

³²¹ On peut penser que ce genre de raisonnement est implicitement présent chez tous les auteurs qui prennent le semi-compatibilisme post-frankfurtien pour un compatibilisme de plein droit : « si l'on a sauvé la responsabilité morale, on peut dire qu'on a *tout* sauvé puisque le reste (capacité d'agir autrement, etc.) n'a aucune importance. »

elle serait une frustration tout de même. Ou pour le dire autrement, un véritable optimisme, un optimisme fort, devrait défendre que :

Optimisme fort : nous avons toutes les sortes de libre arbitre auxquelles nous attribuons effectivement (à tort ou à raison) de l'importance.

Mais ce que pourrait établir l'argument de l'inconnaissabilité, c'est au mieux une forme d'optimisme « raisonné » :

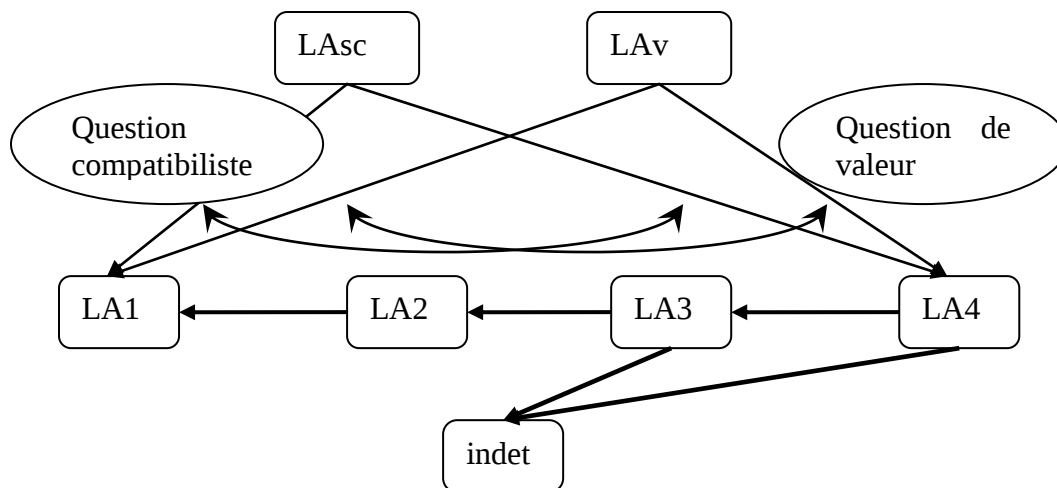
Optimisme raisonné : nous avons toutes les sortes de libre arbitre auxquelles il est raisonnable d'attribuer de l'importance (ou qu'il vaut la peine de souhaiter avoir si on se raisonne).

Or il n'est pas clair que cet optimisme raisonné soit si optimiste que cela étant donné que nous n'avons pas (d'une manière générale) la faculté de rendre nos désirs raisonnables.

La quatrième raison avancée par Dennett en faveur de l'optimisme est également une défense de l'optimisme « raisonné ». Elle repose sur l'idée qu'il est irrationnel de désirer l'impossible. « Si l'idéal de liberté que nous attendons est simplement contradictoire, nous ne devrions vraiment pas nous sentir privés de quoi que ce soit si nous ne pouvons pas l'avoir. Il n'y a aucun sens à se torturer les mains parce que nous ne pouvons pas défaire le passé, et ne pouvons pas empêcher un événement qui a lieu effectivement, et ne pouvons pas nous créer nous-mêmes *ex nihilo*, et ne pouvons pas choisir les deux alternatives d'une décision donnée, et ne pouvons pas être parfaits » (Dennett 1984a, 172). Les exemples utilisés ici par Dennett montrent clairement la faiblesse de l'optimisme raisonné : bien sûr, il est irrationnel de désirer changer le passé, mais cela n'empêche pas que nous en ayons, parfois, un très fort désir, et que la frustration puisse en être pénible. Ce type d'argument, si on le combine à l'idée de Dennett d'une absence de pouvoir de faire autrement, louche dangereusement vers un optimisme digne de maître Pangloss : comme nous n'avons aucun pouvoir de faire que les événements soient autrement qu'ils ne sont, nous devrions nous réjouir de tout, puisqu'il est irrationnel de désirer ce que nous ne pouvons changer. Cette dernière raison me semble donc clairement la plus faible.

En somme, l'argument le plus fort me semble être l'argument moral : si la responsabilité morale ne requiert pas un libre arbitre de type indéterministe, cela diminue considérablement la motivation de désirer un tel type de libre arbitre. Autrement dit, la question de moralité et la question de valeur semblent être assez étroitement associées. Sont-elles aussi associées à la question compatibiliste proprement dite, c'est-à-dire à la question d'origine du conflit ? La réponse est non.

La question compatibiliste reste clairement indépendante de la question de valeur : il pourrait très bien se faire que nous ayons une expérience d'être libres de manière incompatibiliste, et que pourtant ce type de libre arbitre n'ait aucune importance particulière pour nous. Dans ce cas, le libre arbitre (celui dont nous faisons l'expérience) serait bel et bien incompatibiliste, et pourtant les conclusions optimistes de Dennett seraient vraies. Ces conclusions n'ont donc rien de compatibiliste en elles-mêmes. En notant LAV pour le type de libre arbitre auquel nous attribuons (ou devrions attribuer) de la valeur, on peut une nouvelle fois résumer ce propos à l'aide d'un schéma :



Implicitement, Dennett pose la question rhétorique suivante : si telle forme de libre arbitre n'a aucune valeur pour nous (si nous ne la désirons pas et n'avons pas de bonne raison de la désirer), à quoi bon s'intéresser à la question de savoir si nous l'avons ou non ? A cette question rhétorique, je défendrai la réponse suivante : il est important de s'y intéresser parce que nous avons l'*expérience* d'avoir un tel type de libre arbitre, que cette expérience fixe le sens du mot « libre arbitre » dans toute la littérature philosophique, et que c'est donc à partir de ce sens qu'on pourra dire si le libre arbitre est ou non une illusion – *quel que soit* l'enjeu moral ou pratique ou rationnel de cette question.

Par ailleurs, je noterai brièvement que je ne crois pas à la thèse dennettienne de l'absence de valeur des types de libre arbitre incompatibilistes. A mon sens, une des raisons pour lesquelles les libres arbitres de type incompatibiliste ont une certaine valeur est *précisément* parce que nous en faisons l'expérience. De manière évidente, il est toujours préférable d'être dans la vérité que d'être dans l'illusion, et donc pour une expérience donnée, *a fortiori* pour une expérience extrêmement fréquente et commune, il est préférable qu'il ne s'agisse pas d'une illusion. Mais il y a un deuxième aspect qui donne une importance particulière à l'expérience du libre arbitre : on a pu mettre en évidence expérimentalement que des agents qui étaient convaincus que le libre arbitre est une illusion avaient tendance à se comporter moralement plus mal que des agents qui croient à la vérité du libre arbitre.³²² Ces

³²² Deux études expérimentales tendent vers cette conclusion. Dans l'étude de (Vohs and Schooler 2008), des sujets sont invités à lire un court texte d'un scientifique affirmant que la science a prouvé l'absence de libre arbitre, tandis que d'autres sujets témoins sont invités à lire un texte neutre. Après ces lectures, les uns et les autres sont invités à réaliser un petit exercice sur ordinateur ; on leur annonce que le logiciel a un léger défaut qui fait que le résultat s'affiche tout seul si l'on ne maintient pas enfoncée la barre d'espace, mais on leur demande de répondre honnêtement et donc d'appuyer sur la barre d'espace pour ne pas « tricher ». Le résultat est le suivant : les personnes qui ont lu le texte anti-libre arbitre avant cet exercice ont une plus forte propension à tricher que celles qui ont lu le texte neutre. L'étude de (Baumeister, Masicampo, and DeWall 2009) arrive à un résultat comparable. Leur premier dispositif expérimental de cette étude soumet également les sujets aux deux lectures (anti-libre arbitre et neutre) que dans l'expérience de Vohs et Schooler. Mais à l'issue de ces lectures, on présente aux sujets la situation d'une personne qui aurait besoin d'aide, et on leur demande s'ils se sentiraient prêts à aider la personne en question. Les sujets qui avaient lu le texte anti-libre arbitre se sont avérés beaucoup moins disposés à adopter une attitude socialement généreuse. Dans l'étude de Vohs et Schooler, comme dans celle de Baumeister, une question supplémentaire a permis de s'assurer que l'effet s'ensuivait bel et bien du fait que les sujets ayant lu le texte anti-libre arbitre se mettaient pour certains à *croire* à l'absence de libre arbitre. Il semble donc que la croyance en l'absence de libre arbitre ait tendance à générer plus de comportements de triche et moins de comportements socialement généreux. Les résultats de ces expériences (et de quelques autres) ont été réunis dans (Schooler 2010).

données sont probablement liées au fait mentionné au chapitre 1 (p.52), à savoir que les gens ont tendance à penser (*peut-être à tort*) que sans libre arbitre il n'y a pas de responsabilité morale.³²³ Quand bien même les gens auraient *tort* à ce sujet, il n'empêche qu'une révélation « illusionniste » ou « déterministe dure » (par exemple sous la forme d'une instruction publique enseignant aux enfants qu'ils ne sont pas libres) aurait très probablement des conséquences pratiques et morales importantes.

Pour conclure, la question de valeur est logiquement indépendante de la question compatibiliste : il pourrait se faire que le libre arbitre dont nous faisons l'expérience soit incompatibiliste même s'il n'y a aucun intérêt particulier à avoir effectivement un tel type de libre arbitre. Tous les discours sur l'importance ou la valeur des libres arbitres de type incompatibiliste devraient donc être éliminés de la discussion sur le compatibilisme proprement dit ; seules importeront les conditions de vérité de notre expérience commune. Notez que Dennett lui-même défend véritablement le compatibilisme (à d'autres endroits)³²⁴ et pas seulement l'optimisme. Le reproche que je lui adresse est plutôt d'avoir mêlé et confondu les deux argumentations plutôt que d'avoir totalement manqué la cible du compatibilisme.

II.4. P. F. Strawson, ou le Quiétisme : non pertinence de la question de résilience

La position de P. F. Strawson est également présentée traditionnellement comme une forme de compatibilisme ; Kane (2005, 111), notamment, l'appelle « le compatibilisme des attitudes réactives ». On pourrait résumer cette position de la manière suivante : (i) le problème du libre arbitre concerne avant tout les conditions dans lesquelles nous attribuons aux agents la responsabilité morale (en particulier la question de savoir si nous l'attribuerions ou devrions l'attribuer en cas de déterminisme causal) ; (ii) mais attribuer la responsabilité morale n'est pas avant tout une affaire de théorie, ni même d'activités délibérées, c'est avant tout une affaire d'attitudes réactives – telles que le ressentiment, l'admiration, la gratitude, l'indignation, la culpabilité, le blâme, l'approbation, etc. ; or (iii) toutes ces attitudes sont si profondément ancrées dans notre nature humaine que nous ne pourrions pas raisonnablement les abandonner si l'on venait à apprendre que le déterminisme est vrai ; par conséquent, (iv) la question du déterminisme n'a pas d'impact sur notre conception de la liberté humaine.

Par les thèses (i) et (ii), c'est-à-dire l'insistance sur la responsabilité morale comme enjeu principal du débat, la position de P. F. Strawson se rapproche de la position de Frankfurt ou de Fischer dont nous avons parlé un peu plus haut. Dans cette sous-section, je m'attacherai plutôt à la deuxième grande intuition de P. F. Strawson, à savoir l'idée selon laquelle (iii) notre engagement au libre arbitre (et à la responsabilité morale) est si profondément ancré que nous ne l'abandonnerions pas (et ne devrions pas l'abandonner) si d'aventure les scientifiques

³²³ Voir cependant l'interprétation différente que propose (Schooler 2010, 198–200) : la croyance en l'absence de libre arbitre n'aurait pas pour conséquence de supprimer notre sens de la responsabilité morale, mais aurait plutôt pour conséquence de diminuer « l'énergie mentale » que nous serions prêts à dépenser, ce qui entraînerait moins de résistance à la tentation de tricher, et moins de disponibilité pour faire l'effort d'aider son prochain.

³²⁴ Dennett exprime son compatibilisme notamment dans l'affirmation suivante : « free will is not an illusion, not even an irrepressible and life-enhancing illusion. When we look closely at the sources of our suspicion and dread, we find again and again that they are not indisputable axioms or overwhelmingly well-supported, empirical discoveries, but unfocused images, hastily glanced at. » (Dennett 1984a, 169) D'après cette citation, l'origine de l'incompatibilisme ne tient pas à notre expérience du libre arbitre – laquelle n'est *pas* illusoire, car elle a des conditions de vérité compatibilistes – mais à des métaphores (les fameux « bugbears » dont nous parlerons un peu plus loin), c'est-à-dire à des représentations théoriques qui biaisent notre expérience commune.

établissaient la vérité du déterminisme. C'est ce qu'on peut appeler, à la suite de John Fischer, l'idée de « résilience »³²⁵ de notre engagement à la découverte éventuelle du déterminisme. Dans l'argumentation de Strawson, c'est cet aspect de résilience qui est fondamental pour arriver à la conclusion d'une « non pertinence » du déterminisme par rapport à la question du libre arbitre, laquelle conclusion est ensuite interprétée (à tort) comme une forme de compatibilisme. Cette conclusion, je l'appellerai « quiétiste », parce qu'elle revient à dire ceci : nous n'avons pas à *être inquiet* des éventuelles informations concernant le déterminisme, car quoi qu'elles puissent nous apprendre, nous ne changerons pas nos attitudes et engagements fondamentaux à l'égard du libre arbitre.³²⁶ Dans cette sous-section, je m'attacherai donc à montrer que la thèse de résilience, et la position « quiétiste » à laquelle elle mène, ne sont absolument pas des formes de compatibilisme. La meilleure preuve sera qu'il existe des incompatibilistes « résilients » et « quiétistes ».

Chez P. F. Strawson, comme on vient de le voir, l'élément résilient est l'ensemble de nos attitudes réactives d'imputation morale. Pour Strawson, ces attitudes constituent l'essentiel de notre engagement inter-personnel dans les relations humaines : les abandonner reviendrait à adopter sur les gens « l'attitude objective » avec laquelle nous considérons et utilisons les objets inertes, les êtres irrationnels, et parfois les êtres humains mentalement déficients. Adopter ce genre d'attitude objective temporairement ou pour certains agents, nous le pouvons parfois, mais abandonner absolument toute attitude inter-personnelle, c'est ce que notre nature ne nous permettrait pas de faire.

Pour nous, tels que nous sommes, c'est inconcevable en pratique. L'engagement humain à participer à des relations inter-personnelles est, je crois, trop universellement et profondément enraciné pour que nous prenions au sérieux l'idée qu'une conviction théorique d'ordre général puisse changer le monde d'une façon telle qu'il n'y ait plus du tout de relations inter-personnelles telles que nous les comprenons normalement ; et être engagé dans une relation inter-personnelle comme nous le comprenons normalement, c'est précisément être sujet aux attitudes et sentiments réactifs dont nous parlons.
(P. F. Strawson 2008, 12)

Dans sa première formulation, cette thèse est d'abord une thèse *factuelle* : nous ne pourrions pas abandonner nos attitudes réactives « même si une vérité générale nous donnaient un fondement théorique de le faire » (*ibid.*) Mais, ainsi que le remarque Strawson un peu plus loin, la question principale ne concerne pas tant ce que nous *ferions* ou pourrions faire en cas de preuve du déterminisme ; la question principale « concerne ce qu'il serait *rationnel* de faire si le déterminisme était avéré » (P. F. Strawson 2008, 14). Et Strawson défend ensuite la thèse forte, ou normative, selon laquelle non seulement nous préserverions nos attitudes réactives, mais nous serions rationnellement fondés à le faire, essentiellement sur la base de considérations pragmatistes.³²⁷

³²⁵ Voir notamment (Fischer 1999, 129) : “This resiliency idea is a major motivation for my acceptance of semicompatibilism.”. Voir aussi (Fischer 2000, 323) et (Fischer 2006, 6).

³²⁶ Ce sens du mot « quiétisme » a peu de choses à voir avec le sens que l'on trouve, par exemple, chez McDowell, Brandom ou Huw Price, c'est-à-dire le « quiétisme wittgensteinien ». Le quiétisme wittgensteinien est une thèse sur le rôle de la philosophie en général ; ce rôle serait purement thérapeutique et négatif, et non pas constructif. J'espère que cette divergence de vocabulaire n'entraînera pas de confusion. Si cette terminologie ne convient pas, on pourrait parler à la place des « philosophes confiants », ou simplement des « philosophes résilients ».

³²⁷ Je n'entre pas ici dans le détail de l'argumentation de Strawson. Les deux considérations qui, pour lui, font passer d'une résilience factuelle à une résilience rationnelle sont (i) la prise en compte des conséquences pratiques, gains et pertes, qu'aurait un abandon des relations inter-personnelles, et (ii) de nouveau la place

Lorsqu'on s'intéresse à la thèse de résilience en elle-même, on s'aperçoit qu'elle peut être envisagée de manière plus générale, c'est-à-dire indépendamment de la question des attitudes réactives et même de la question de la responsabilité morale. On pourrait en effet défendre de manière directe que le libre arbitre, c'est-à-dire notre croyance au libre arbitre, est en lui-même résilient à une éventuelle découverte du déterminisme. Cette thèse a été explicitement défendue par John Fischer :

Ne serait-il pas étrange que notre conception de nous-mêmes comme libres [...] dépende de la question de savoir si les probabilités de tels conditionnels [les conditionnels d'événements microscopiques dont dépendent nos choix] sont à 99.9% ou à 100% ? (Fischer et al. 2007, 48)

Mes conceptions profondes de moi-même et des autres comme libres [...] sont et devraient être résilientes à l'égard d'une telle découverte sur les faits mystérieux et abscons relatifs aux généralisations de la physique. (Fischer et al. 2007, 44-45)

On retrouve ici la différence entre une résilience faible ou factuelle et une résilience forte ou normative (« sont et devraient être résilientes ») et la défense explicite de la thèse forte. Chez Fischer, la thèse normative repose sur l'intuition d'une incommensurabilité entre les conséquences gigantesques d'un abandon de nos conceptions de nous-mêmes et d'autre part le caractère infime des faits physiques dont on voudrait faire dépendre ces conceptions.³²⁸

L'idée de résilience peut donc être généralisée, et n'est pas intrinsèquement dépendante d'une stratégie qui passe par la notion de responsabilité morale. Et quel que soit le chemin qu'elle emprunte, la conclusion souhaitée sera une forme de « quiétisme ». Cette motivation quiétiste est très claire chez Fischer :

Une de mes motivations principales pour être compatibiliste, c'est que je ne veux pas que notre statut de personnes et notre responsabilité morale tiennent, pour ainsi dire, à un fil, ou soient pris en otage par l'éventuelle découverte scientifique que le déterminisme est en fait vrai. [...] Donc c'est une motivation intuitive pour moi, cette résilience de notre statut de personnes et de notre responsabilité. (Fischer 2000, 323)

A l'attitude quiétiste, Fischer oppose ici l'image inquiétante d'une « prise en otage » de notre conception de nous-mêmes par la science future. Pour être à l'abri d'une telle prise en otage, il faut accepter une forme de résilience de notre conception de nous-mêmes... mais faut-il nécessairement accepter le *compatibilisme* ainsi que Fischer le dit ici ? Il est évident que si l'on s'en tenait à une résilience strictement factuelle, on n'aurait aucunement le droit de conclure à une forme de compatibilisme. En effet, une résilience strictement factuelle voudrait dire ceci : si le déterminisme était avéré, nous continuerions *peut-être à tort et irrationnellement* de nous considérer comme libres. Un incompatibiliste peut tout à fait défendre cette position ; elle reviendrait simplement à dire qu'en cas de preuve du déterminisme nous vivrions tous dans une illusion irrépressible et irrationnelle de liberté. Mais Strawson et Fischer défendent une forme forte de résilience, une résilience normative : ce n'est pas seulement que nous *garderions* notre conception de nous-mêmes, mais que nous *devrions* (épistémiquement) garder notre conception de nous-mêmes ; nous serions

fondamentale de ces attitudes dans notre mode de vie, qui les mettent hors de tout calcul rationnel de révision. Voir (P. F. Strawson 2008, 14).

³²⁸ Compte tenu de ce que nous avons vu plus haut du semi-compatibilisme de Fischer, il faut noter que sa motivation de résilience n'est pas entièrement satisfaite : en fin de compte, Fischer *doit* abandonner une certaine image de nous-mêmes comme libres, mais il préserve la résilience de nos conceptions de contrôle qui fondent l'imputation morale.

rationnellement fondés à le faire. Si tel est le cas, n'est-il pas vrai que la thèse déterministe serait sans pertinence rationnelle pour la question du libre arbitre ? Et cela n'est-il pas évidemment une thèse compatibiliste ? Comment une telle thèse ne serait-elle pas compatibiliste si elle est préparée à accepter (éventuellement) *à la fois* le déterminisme et le libre arbitre ?

En réalité quiétisme et compatibilisme sont indépendants. On peut défendre la résilience du libre arbitre, être quiétiste face à la possibilité d'une découverte du déterminisme, et néanmoins être incompatibiliste. Le plus simple est ici de partir de l'exemple de Peter van Inwagen qui présente exactement ces caractéristiques.

Ainsi que j'ai eu l'occasion de l'évoquer au chapitre 3 (p.242, et note 82), van Inwagen ne se contente pas de donner une liste de thèses qu'il accepte *effectivement* : il donne en outre (van Inwagen 1983, 219–220) une classification de ces thèses par ordre intuitif, ce qui lui permet de dire *conditionnellement* quelles thèses il serait d'avantage prêt à abandonner si on l'amenait à changer de position. Pour simplifier, disons que van Inwagen soutient actuellement la conjonction de thèses suivante :

- (i) nous sommes moralement responsables
- (ii) la responsabilité morale suppose d'être libres
- (iii) être libre implique l'indéterminisme

Ce qui l'amène à accepter la conséquence :

- (iv) donc le déterminisme est faux.

Il est clair que cette conjonction de thèse est incompatibiliste. Mais van Inwagen ajoute que ces différentes thèses n'ont pas pour lui le même degré de certitude ou de force intuitive, et qu'en particulier (i) est plus forte que (ii) qui est plus forte que (iii). Cette classification permet à van Inwagen de déduire ce qu'il ferait, ou plutôt ce qu'il devrait faire rationnellement, si on lui apportait un démenti formel pour sa croyance (iv), c'est-à-dire une preuve du déterminisme. Si on lui apportait une preuve du déterminisme, van Inwagen se trouverait devant un conflit d'intuition comme nous en rencontrons souvent, en l'occurrence un conflit entre les intuitions (i), (ii), (iii) et (~iv) – un tétralemmme pour ainsi dire. Que faire rationnellement en cas de conflit d'intuition ? Abandonner l'intuition la plus faible. Quelle intuition est la plus faible pour van Inwagen ? L'intuition (iii), c'est-à-dire l'intuition incompatibiliste à proprement parler. Autrement dit, si on lui apportait une preuve du déterminisme, van Inwagen maintiendrait sa croyance à la responsabilité morale, ainsi que sa croyance résultante au libre arbitre, et donc il *deviendrait* compatibiliste. Cela ne veut évidemment pas dire qu'il est d'ores et déjà compatibiliste – tout son livre n'est qu'un long argument en faveur de l'*incompatibilisme*. Mais il est parfaitement rationnel d'être prêt à réviser ses croyances en cas de nouvelles informations, et en l'occurrence, van Inwagen déclare qu'il réviserait sa croyance incompatibiliste en cas d'information déterministe. Certes il *pourrait* se faire que, pour d'autres auteurs incompatibilistes, l'intuition incompatibiliste est plus forte que l'intuition du libre arbitre. Dans ce cas, *pour ces auteurs*, l'attitude la plus rationnelle serait d'abandonner la croyance au libre arbitre et de rester incompatibiliste. Mais étant donné l'ordre de force des intuitions de van Inwagen, le plus rationnel serait d'abandonner l'incompatibilisme et de préserver la croyance au libre arbitre.

La stratégie défendue par van Inwagen est donc une stratégie qui est à la fois clairement incompatibiliste, et clairement quiétiste. Van Inwagen défend la résilience de sa croyance au libre arbitre par rapport à l'éventuelle découverte du déterminisme, et par conséquent sa conception de nous-mêmes et de nos actions n'est pas « inquiétée » par l'éventualité d'une découverte scientifique du déterminisme. Comme on l'a vu (p.242), John Fischer a critiqué cette stratégie en lui reprochant de « retourner sa veste métaphysique » en cas d'information empirique ; on peut en effet avoir des réticences à faire dépendre certaines de nos conceptions

métaphysiques (et donc *a priori*) d'informations empiriques. Mais quelles que soient les réticences que l'on puisse avoir face à cette stratégie de van Inwagen, il est incontestable (et Fischer ne conteste pas) qu'elle respecte la résilience de la croyance au libre arbitre (et à la responsabilité morale) et qu'elle est donc une forme de quiétisme, bien qu'elle soit également une forme d'incompatibilisme.

Pour conclure cette sous-section, il sera peut-être utile de lever une difficulté de compréhension que semble susciter fréquemment la position de van Inwagen. On peut la formuler ainsi : si van Inwagen est prêt à changer d'avis comme il le dit, c'est qu'il doit bien considérer une preuve du déterminisme comme *possible*, et cette *possibilité* qu'il envisage, c'est une possibilité où le déterminisme sera vrai, et où sa croyance au libre arbitre sera vraie également ; donc dès aujourd'hui il considère comme *possible* que le déterminisme soit vrai en même temps que l'existence du libre arbitre. Mais n'est-ce pas dire qu'il est d'ores et déjà compatibiliste ? Cette objection confond en fait la notion métaphysique de possibilité avec la notion épistémique. Dire que *p* est épistémiquement possible (pour moi), c'est dire que je ne suis pas certain que $\sim p$. C'est donc compatible avec la croyance que $\sim p$ et même avec la croyance que $\sim p$ est une nécessité métaphysique. Par exemple, si je crois que Hespérus est Phosphorus mais que je n'en suis pas certain, je peux dire qu'il est épistémiquement possible que Hespérus ne soit pas Phosphorus, mais il ne s'ensuit pas que, pour moi, il soit métaphysiquement possible que Hespérus ne soit pas Phosphorus. Etant donné que je les crois identiques (bien que sans certitude), je les crois aussi *métaphysiquement nécessairement* identiques (bien que sans certitude). De même, lorsque van Inwagen envisage la *possibilité* d'une preuve future du déterminisme, il ne s'agit que de possibilité *épistémique* : il manifeste par là qu'il n'est pas absolument *certain* que l'indéterminisme soit vrai, et ainsi qu'il nous l'apprend, il n'est pas *certain* non plus que l'incompatibilisme soit vrai. Mais cela ne veut pas dire que l'incompatibilisme est pour lui une thèse qui pourrait métaphysiquement être fausse : cette thèse qu'il affirme (sans certitude absolue), il la tient évidemment pour *métaphysiquement nécessaire*. Pour dire les choses de manière formelle et explicite, le compatibilisme consiste à soutenir que :

(Comp) Il est métaphysiquement possible que à la fois 'nous soyons libres' et 'le déterminisme soit vrai'.

tandis que la stratégie prudente de Peter van Inwagen consiste simplement à reconnaître ceci :

(PvI) Il est épistémiquement possible qu'il soit métaphysiquement possible que à la fois 'nous soyons libres' et 'le déterminisme soit vrai'.

Bien que van Inwagen considère le compatibilisme comme épistémiquement possible, dans le sens faible où il n'est pas certain qu'il soit faux (ou encore, dans le sens où il pourrait être *amené* à penser qu'il est vrai dans certaines conditions), néanmoins dans l'état actuel de ses informations, et sans certitude, il le tient néanmoins pour faux, et pour nécessairement faux. La position de van Inwagen est donc bien une position qui préserve la résilience de la croyance au libre arbitre tout en restant pleinement incompatibiliste.

La thèse de résilience et sa conclusion quiétiste sont donc indépendantes de la thèse compatibiliste. Ce qui donne l'impression que le quiétisme est un compatibilisme, c'est l'idée vague selon laquelle « l'information déterministe n'aurait aucune pertinence pour notre conception de notre liberté ». Mais cette idée est trop ambiguë, comme on peut le voir chez un auteur à la fois quiétiste et incompatibiliste comme van Inwagen : pour ce dernier, il est vrai que l'éventuelle *information* du déterminisme n'aurait pas d'incidence sur sa *croyance* au

libre arbitre (n'entraînerait pas la révision de cette croyance) ; mais il est faux de dire que, en l'état actuel de ses croyances, la *thèse* du déterminisme n'a aucun lien logique (d'incompatibilité) avec la *thèse* du libre arbitre. La confusion entre quiétisme et compatibilisme repose sur une confusion plus fondamentale entre le niveau épistémique des croyances et de leurs révisions éventuelles, et le niveau métaphysique des propositions.

II.5. Vargas, ou le Révisionnisme : non pertinence des pratiques linguistiques

Une dernière attitude qui peut être parfois prise pour une forme de compatibilisme est l'attitude « révisionniste » défendue par Manuel Vargas (Fischer et al. 2007; Vargas 2011). Cette attitude peut se résumer de la manière suivante : (i) notre concept de sens commun de libre arbitre est un concept libertarien (incompatibiliste) ; par ailleurs, (ii) il est très probablement faux que nous ayons un tel type de libre arbitre ; mais devant ces deux premiers constats, la bonne réaction ne consiste pas à tirer une conclusion sceptique ou incompatibiliste dure, (iii) la bonne réaction consiste à réviser notre concept de libre arbitre dans un sens compatibiliste, de telle sorte que nous puissions continuer à affirmer que nous avons le libre arbitre.

Cette stratégie originale a clairement une affinité avec l'incompatibilisme dur par ses thèses (i) et (ii), mais par sa thèse fondamentale, la thèse (iii), elle peut aisément être prise pour une forme de compatibilisme. Cette stratégie de « révision » pourrait aussi être rapprochée de la démarche de van Inwagen, qui est prêt à « revoir » son incompatibilisme en cas de preuve du déterminisme, et il sera important de bien les distinguer.

Pour bien comprendre la position de Vargas, il faut se demander ce que signifie l'idée de « réviser notre concept de libre arbitre ». Que peut signifier l'idée de « réviser notre concept » de libre arbitre ? Vargas donne quelques exemples de révision de concept : notre concept « d'eau », avant les découvertes de la chimie, désignait une substance indécomposable (Fischer et al. 2007, 146),³²⁹ mais comme nous nous sommes aperçus que nos verres d'eau ne contenaient pas une telle substance indécomposable mais plutôt une grande quantité de molécules d'H₂O, nous avons dû réviser notre concept pour prendre en compte cette information. En découvrant l'absence d'une substance indécomposable, nous n'avons pas conclu « donc l'eau n'existe pas », nous avons révisé notre concept d'eau pour qu'il reste vrai que « l'eau existe ». De même, dit Vargas, nous avons révisé notre concept de « magicien » (*ibid.*) : celui-ci désignait auparavant les personnes dotées de pouvoirs surnaturels, mais nous l'avons révisé pour désigner les personnes capables de nous divertir en donnant l'*illusion* de tels pouvoirs. Dans les deux cas, nous avons révisé notre concept parce que « le monde ne coopérait pas », c'est-à-dire que la réalité ne présentait aucune chose qui corresponde exactement à notre concept de départ. Vargas estime que nous sommes dans la même situation pour le libre arbitre : le monde ne coopère pas et ne présente (probablement) aucune réalité correspondant exactement à notre concept de sens commun de libre arbitre, et par conséquent, nous devrions réviser notre concept pour désigner quelque chose que la réalité présente effectivement.

Malgré ces exemples, il me semble que l'idée de Vargas reste fondamentalement ambiguë entre deux possibilités (que Vargas envisage toutes les deux sans chercher à trancher). Pour voir cette ambiguïté, on peut repartir de la distinction de Dana Nelkin entre

³²⁹ Ou plus exactement « homéomère », c'est-à-dire qu'on lui attribuait intuitivement la propriété suivante : toute sous-partie d'un volume d'eau, quel qu'il soit est encore un volume d'eau.

concept et *conception* (par exemple entre concept du libre arbitre et conception du libre arbitre, voir chapitre 1, p.66). Il se pourrait par exemple que nous ayons un *concept* compatibiliste de libre arbitre, mais que nous nous trompions dans l'analyse réflexive de ce concept et que nous nous en fassions une *conception* incompatibiliste. Si par la suite nous étions amenés à prendre conscience de notre erreur d'analyse, nous réviserions alors notre *conception* du libre arbitre : nous jugerions que le libre arbitre est en fait compatible avec le libre arbitre. Nous pouvons donc réviser notre *conception* du libre arbitre. Mais en révisant notre conception, nous n'aurions évidemment pas révisé notre *concept* : au contraire, l'idée même de réviser notre conception de X, notre jugement à propos de X, suppose que c'est bien à propos d'un seul et même X (de X sous un seul et même concept) que nous avons modifié notre jugement. L'exemple de l'eau rentre probablement dans cette catégorie de la révision de conception : si l'on considère que le concept d'eau est un concept d'espèce naturelle (signifiant quelque chose comme « la substance quelle qu'elle soit que l'on trouve dans les lacs, les sources et les mers »), alors on peut dire que nous avons une *conception* de l'eau comme étant indécomposable, mais que nous avons été détrompés à ce sujet, car la substance que l'on trouve dans les lacs, les sources et les mers n'est *pas* en fait une substance indécomposable. Il s'agit là d'une simple révision de jugement à propos d'une même chose que l'on continue de désigner par le même concept. De façon similaire, dit Vargas, « il se pourrait que [le libre arbitre] soit différent de ce que nous supposons habituellement qu'il est » (Fischer et al. 2007, 146).

Après la révision de conception, il faut envisager ce que serait une révision du *concept* à proprement parler. Quel sens pourrait avoir une telle expression ? L'évolution du mot « magicien » donne probablement le meilleur exemple d'une révision de concept au sens strict. A la différence du cas de « l'eau », il semble qu'on doive dire, dans le cas de « magicien », que le mot a tout simplement changé de signification. Si l'on regarde dans un dictionnaire, on trouvera deux définitions distinctes, deux concepts distincts, pour l'entrée « magicien », l'une se rapportant à une personne aux facultés surnaturelles (et dans ce sens-là, il n'existe pas de magiciens), l'autre désignant simplement un illusionniste (et dans ce sens-là, il y a évidemment des magiciens). Autrement dit, « réviser notre concept » veut simplement dire ici « réviser notre usage du mot », de telle sorte que le même mot qui signifiait un certain concept c1 (un concept non instancié dans la réalité) soit désormais utilisé pour signifier un autre concept c2 (lequel concept est, pour sa part, instancié). Vargas envisage également cette possibilité dans le cas du libre arbitre : supposons que notre concept de libre arbitre soit *en lui-même* un concept incompatibiliste (c'est-à-dire que ce n'est pas seulement notre *conception* du libre arbitre qui est incompatibiliste mais bien le concept), dans ce cas, il faudrait reconnaître (d'après Vargas) que la réalité ne contient aucune chose correspondant à ce concept. Donc nous ne pourrions pas dire que « le libre arbitre existe » si nous continuions à utiliser le mot libre arbitre pour signifier ce concept. Mais nous aurions tout de même la possibilité de changer notre usage du mot « libre arbitre » pour désigner un nouveau concept.

[P]our quelle raison n'appellerions-nous pas simplement 'libre arbitre' tout ce qui satisfait les capacités liées à la liberté et requises pour l'éloge et le blâme [?] S'il y a une justification pour nos pratiques et attitudes caractéristiques de la responsabilité, indépendamment du libertarianisme, alors pourquoi ne pas penser que la capacité liée à la liberté qui apparaît dans ces pratiques devrait compter comme un bon ré-ancrage pour notre usage et notre compréhension du libre arbitre ? [...]

Quels que soient leurs autres fonctions, nos concepts de libre arbitre et de responsabilité morale sont importants pour nous aider à organiser, à suivre et à justifier différentes manières de nous traiter les uns les autres. Si l'on peut montrer qu'il y a quelque chose qui joue ces rôles, qui fait le travail qui est censé être fait ces concepts, [...] cela

montrerait qu'il y a quelque chose de fonctionnellement équivalent (ou presque) au libre arbitre et à la responsabilité morale. [...] [P]ourquoi une théorie de cette chose proche ne pourrait-elle *pas* servir de concept de remplacement adéquat [?]
(Fischer et al. 2007, 146–147)

Ce que propose ici Vargas est bien un changement « d'usage » linguistique : il s'agit d'utiliser le même terme « libre arbitre » dans un sens nouveau. On pourrait donc lui faire l'objection de Ayer, citée en exergue de ce chapitre :

Il est effectivement évident que si l'on a le droit de donner au mot « liberté » n'importe quel sens, à notre bon plaisir, nous pouvons trouver un sens qui le réconciliera avec le déterminisme : mais ce n'est pas plus une solution pour notre problème que le fait que le mot « cheval » puisse être arbitrairement utilisé pour signifier ce que l'on signifie habituellement par « moineau » n'est une preuve que les chevaux ont des ailes.
(Ayer 1954, 6)

Pour être exact, il faut reconnaître que le changement linguistique proposé par Vargas n'est pas tout à fait aussi « arbitraire » que l'idée d'utiliser le mot « cheval » pour désigner les moineaux. En effet, ce qui motive la révision linguistique proposée par Vargas, c'est l'idée qu'il y a une grande proximité fonctionnelle entre notre concept de départ et le « concept de remplacement ». Le raisonnement de Vargas est le suivant : comme l'ancien concept n'a aucune application dans la réalité, utiliser le mot « libre arbitre » pour désigner ce concept rendrait quasiment inutile le mot « libre arbitre » ; ce mot inutile peut donc retrouver une nouvelle vie si on lui donne un nouvel usage, une nouvelle signification ; et en lui réattribuant une signification fonctionnellement proche de la signification d'origine, on suit la pente de l'évolution naturelle des langues, dans laquelle les mots peuvent changer progressivement de sens par proximité fonctionnelle. Cette révision de vocabulaire semble donc être une bonne solution pragmatique au problème de l'inapplicabilité du terme « libre arbitre » dans son sens d'origine.

Telles sont donc les deux possibilités envisagées par Vargas : réviser notre *conception* du libre arbitre (nous nous étions trompés, le concept de libre arbitre est en fait un concept compatibiliste), ou si cela est impossible (c'est-à-dire s'il est établi que notre concept lui-même est bel et bien incompatibiliste) « réviser notre concept », c'est-à-dire en réalité réviser notre usage linguistique, pour que le terme « libre arbitre » désigne désormais un concept compatibiliste. Il faut noter que ces deux possibilités sont totalement hétérogènes : l'une présente une raison théorique de réviser un jugement, l'autre une raison pragmatique de changer l'usage des mots, sans aucune révision de jugement. Bien que Vargas choisisse de rester *neutre* entre ces deux possibilités, il est indispensable de les évaluer séparément.

L'idée de réviser notre conception du libre arbitre est celle qui se rapproche le plus de la stratégie de Peter van Inwagen. Il est clair que lorsque van Inwagen envisage de devenir compatibiliste en cas de preuve scientifique du déterminisme, il n'envisage pas un simple changement dans ses pratiques linguistiques. Il envisage une situation qui l'obligerait à « changer d'avis » de manière importante. Alors que, en l'état présent, van Inwagen considère le libre arbitre comme incompatible avec le déterminisme, il pourrait être amené à réviser cette conception. La différence avec le révisionnisme de Vargas tient seulement au caractère potentiel de cette révision : van Inwagen *croit* à l'incompatibilisme et *réviserait* ce jugement si telles informations l'y obligeaient. Mais d'après Vargas, les informations qui empêchent de croire à l'incompatibilisme libertarien sont *déjà* présentes, et par conséquent il nous faut *dès maintenant* réviser notre conception du libre arbitre pour devenir compatibilistes (si du moins

nous le pouvons, c'est-à-dire s'il n'est pas établi que notre *concept* de libre arbitre est en lui-même incompatibiliste). Supposons que nous le puissions, quelle serait la thèse alors obtenue ? La thèse obtenue serait la suivante : notre concept de libre arbitre est (en fait) compatible avec le déterminisme (contrairement à ce que pense le sens commun à propos de ce concept, ce qui signifie que nous avons été obligés d'opérer une révision de jugement). Autrement dit, la thèse obtenue n'est, sur le fond, rien de plus que le compatibilisme. Reste à savoir *pourquoi* la thèse compatibiliste serait admise en dépit de notre conception initiale. La motivation de Vargas semble être simplement de préserver notre croyance (de sens commun) en l'existence du libre arbitre. Dans ce cas, le révisionnisme est en fait équivalent à ce que j'ai appelé plus haut le « compatibilisme d'ajustement ». Or ce que nous cherchons dans ce chapitre, ce sont d'éventuels arguments en faveur du « compatibilisme indépendamment motivé ». De ce point de vue, Vargas n'apporte aucun élément ; pour lui, l'analyse incompatibiliste de notre concept est clairement une position qui a une certaine force en elle-même, tant qu'on ne cherche pas à « sauver » l'existence du libre arbitre, et il semble même que Vargas accorde une plausibilité très sérieuse à l'idée que cette analyse de notre concept soit ultimement la bonne, puisque toute l'originalité de sa stratégie est justement d'envisager la possibilité d'un « remplacement du concept » lui-même.

L'idée de réviser notre *concept*, ou plutôt de réviser notre usage du mot « libre arbitre » pour lui faire désigner un nouveau concept, n'a clairement plus rien à voir avec la question du compatibilisme proprement dit. Il s'agit strictement d'une question pragmatique-linguistique (comment devrions-nous utiliser nos termes dans telles ou telles circonstances ? Quelles raisons pratiques peut-on avoir pour nos stipulations terminologiques ?) Certes, si la réponse à cette question pragmatique-linguistique était celle décrite par Vargas, alors nous devrions utiliser le mot « libre arbitre » pour désigner un « concept compatibiliste ». Mais cela ne veut absolument pas dire que le concept de libre arbitre est compatible avec le déterminisme : au contraire, c'est précisément parce que le concept de libre arbitre n'est *pas* compatible avec le déterminisme qu'il faut, d'après Vargas, se mettre à utiliser le même mot pour désigner un nouveau concept. Autrement dit, si l'on se concentre sur la question de fond, en laissant de côté les simples questions de langage, le « révisionnisme conceptuel » de Vargas est tout simplement un incompatibilisme dur. Une fois qu'on a accepté cette thèse de fond, on est peut-être fondé pragmatiquement à changer de langage – peut-être, et peut-être pas – mais quoi qu'il en soit, un simple changement de langage n'est jamais un changement de thèse, or c'est bien de la *thèse* compatibiliste ou de la *thèse* incompatibiliste que nous souhaitons parler ici.

Pour conclure, le révisionnisme de Vargas est ambigu entre deux lectures. Comme révisionnisme de notre conception, il correspond à ce que j'ai appelé le compatibilisme d'ajustement et non pas un compatibilisme « indépendamment motivé ». Il n'est donc pas pertinent pour ce chapitre. Deuxièmement, comme révisionnisme du concept, la position de Vargas est fondamentalement une question pragmatique-linguistique consistant à se demander en quel langage il faut s'exprimer *étant posé que l'incompatibilisme est vrai* et qu'il mène au scepticisme du libre arbitre. Sur le fond, la position est donc incompatibiliste, et la question pragmatique-linguistique ultérieure, bien qu'intéressante, est totalement périphérique à notre enquête.

II.6. Le compatibilisme comme Théorie de l'Illusion

Nous avons à présent fait le tour de tout ce que le compatibilisme *n'est pas*. Il ne suffit pas pour être compatibiliste « d'utiliser » un certain concept compatibiliste comme réponse à une certaine question. Être compatibiliste, ce n'est pas dire que tous les types de libre arbitre que nous observons scientifiquement sont compatibilistes, que tous les types de libre arbitre requis par la responsabilité morale sont compatibilistes, que tous les types de libre arbitre ayant une véritable valeur sont compatibilistes ; être compatibiliste, c'est dire que le type de libre arbitre *dont nous faisons l'expérience et qui est à l'origine du conflit intuitif* avec le déterminisme est en réalité un concept compatible avec le déterminisme. Être compatibiliste, ce n'est pas non plus croire de manière « résiliente » à l'existence du libre arbitre, quitte à réviser potentiellement un jugement initialement incompatible, voire à changer de vocabulaire pour préserver la phrase « nous avons le libre arbitre » ; être compatibiliste, au moins un compatibiliste « indépendamment motivé », c'est dire que *dès maintenant*, et indépendamment de notre motivation à préserver l'existence du libre arbitre, nous avons des raisons de tenir notre concept pour compatible avec le déterminisme.

Après ces caractérisations négatives, je terminerai cette section par une caractérisation plus précise de ce qu'est un véritable compatibiliste, ou de ce qu'il « doit faire ». Nous avons déjà vu deux éléments essentiels de ce que doit faire le compatibiliste :

- (a) il doit proposer un concept de libre arbitre explicitement analysé et dont la compatibilité avec le déterminisme ne fait pas débat.
- (b) il doit soutenir que ce concept est précisément *le* concept de libre arbitre, c'est-à-dire le concept que nous avons de nous-mêmes et de nos actions et qui est à l'origine du conflit intuitif avec le déterminisme.

Il est évident qu'il y a une tension entre ces deux tâches du compatibiliste : d'un côté il doit décrire un concept manifestement compatible avec le déterminisme, mais de l'autre il doit nous montrer que c'est précisément ce concept qui a généré un conflit intuitif avec le déterminisme. La question qui se pose immédiatement est la suivante : pourquoi le concept mis en avant génèrerait-il un conflit intuitif avec le déterminisme s'il est manifestement compatible avec le déterminisme ? Si un concept est manifestement compatible avec le déterminisme, on ne devrait pas s'attendre *prima facie* à ce qu'il soit à l'origine d'un conflit intuitif avec le déterminisme. Cette attente n'est que la contraposée du principe que nous avons défendu dans la section I : si notre concept de libre arbitre est à l'origine d'un conflit intuitif avec le déterminisme, l'explication la plus simple *prima facie* est qu'il est réellement incompatible avec le déterminisme. Compte tenu de cette attente, dès qu'un philosophe compatibiliste produira un certain concept manifestement compatibiliste (tâche a), il sera *prima facie* en mauvaise posture pour remplir la tâche b et nous convaincre qu'il s'agit bien du concept à l'origine de notre conflit intuitif. S'il renonce à remplir cette tâche b, alors il renonce au compatibilisme proprement dit et change de sujet (il défendra par exemple que le concept mis en avant est seul pertinent pour la responsabilité morale ou seul digne de nos espoirs, etc.). S'il veut remplir la tâche b, s'il veut nous convaincre que ce concept est bel et bien à l'origine de notre conflit intuitif *malgré* son caractère manifestement compatibiliste, la tâche essentielle sera pour lui de nous expliquer pourquoi ou comment un tel concept a pu malgré tout générer un conflit intuitif avec le déterminisme. Autrement dit, il lui faudra apporter ce qu'on pourrait appeler une *Théorie de l'illusion de conflit*, ou une théorie de la confusion, ce qui pour (Berofsky 2002, 182) constitue « l'autre élément standard du répertoire compatibiliste ». Plusieurs auteurs compatibilistes ont clairement identifié et admis cette tâche :

Si le libre arbitre et la responsabilité sont compatibles avec le déterminisme, alors il nous faudra plus d'information sur ce que sont le libre arbitre et la responsabilité pour expliquer pourquoi l'apparence d'incompatibilité est si persistante. (Wolf 1990, 24)

Qu'est-ce qui produit donc l'illusion, s'il s'agit d'une illusion, que la liberté requiert l'indéterminisme ? (Peacocke 1999, 334)

La deuxième tâche du compatibiliste – celle de montrer que son concept est bel et bien le concept recherché – se ramène donc essentiellement à ceci :

(b') il doit montrer que ce concept, en dépit de son contenu manifestement compatibiliste, est susceptible de produire en nous de manière persistante et répandue un conflit illusoire avec l'idée de déterminisme.

Avec sa clarté habituelle, C.A. Campbell avait déjà donné, en 1931, cette description des deux tâches du compatibiliste :

Il devrait être capable de résoudre l'expérience en question [i.e. l'expérience du libre arbitre], sans aucun reste, en un ensemble d'éléments qui, lorsqu'on les considère réflexivement, n'entraînent clairement avec eux aucune implication libertarienne ; et par ailleurs [...] de montrer clairement comment une expérience de ce type génère « l'illusion libertarienne ». (Campbell 1931, 159)

Pour bien comprendre l'importance de la deuxième tâche, prenons l'exemple du concept conditionnaliste. D'après ce concept, un agent est libre si ses actions sont conditionnellement dépendantes de ses volontés (ou de ses tentatives, ou de ses choix, selon les versions). Il est manifeste que cette capacité est compatible avec le déterminisme. Si nos volontés (nos tentatives, nos choix) sont totalement déterminés en elles-mêmes, alors nos actions qui en dépendent conditionnellement seront elles aussi totalement déterminées ; on possèdera donc la liberté conditionnaliste dans une configuration d'action totalement déterministe. Par ailleurs, il est évident que ce concept fait partie de notre image de nous-mêmes et de nos actions : nous avons certainement l'expérience que nos actions dépendent de nos volontés. Pour le dire autrement, il est certain que cette condition est une condition *nécessaire* pour être libre dans le sens métaphysique. Mais le compatibiliste ne doit pas simplement défendre qu'il a produit un concept intéressant qui fait *partie* de notre image de nous-mêmes : il doit défendre que son analyse *épuit* le concept de nous-mêmes tel qu'il génère en nous un conflit avec le déterminisme. C'est là sa tâche essentielle et son principal défi. Faute de relever ce défi, la subtilité conceptuelle de son analyse ne permettra jamais d'éviter l'objection faite par C.A. Campbell et résumée ainsi par Berofsky :

A l'égard de l'analyse hypothétique [conditionnaliste], C.A. Campbell observa que sa vérité est *au mieux* une condition nécessaire du type de liberté qui est essentiel pour la responsabilité morale. (Berofsky 2002, 182)

Peut-être que ce point sera encore plus clair à partir d'une caricature. Supposons qu'un philosophe déclare la chose suivante : « le concept de libre arbitre n'est rien d'autre que le concept de n'être pas physiquement emprisonné ; mais il est évident que ne pas être

physiquement emprisonné est compatible avec le déterminisme ; donc il est évident que le libre arbitre est compatible avec le déterminisme ». Un tel propos est clairement inacceptable dès la première phrase : il est clair que le concept de libre arbitre (celui utilisé dans la discussion métaphysique depuis plusieurs siècles) n'est *pas* le concept de ne pas être physiquement emprisonné. Ce dernier concept, certes, est clairement « compatibiliste », mais il n'est clairement pas pertinent pour la discussion ; on pourrait dire qu'il est *trop* clairement compatible avec le déterminisme (à vrai dire sans aucun rapport avec le déterminisme) pour qu'il puisse être le concept qui fait l'objet d'un débat séculaire sur la compatibilité avec le déterminisme. Au lieu de montrer sa subtilité analytique, un tel auteur ne révélera alors que son incapacité à comprendre de quoi nous parlons.

Parmi les deux tâches du compatibiliste que je viens de présenter – (a) produire un concept manifestement compatibiliste et (b') expliquer comment ce concept a pu générer une illusion incompatibiliste – la seconde a souvent été présentée comme dialectiquement moins importante. Souvent les théories de la confusion incompatibiliste sont présentées comme des appendices servant simplement à « aider » un lecteur empêtré dans son illusion incompatibiliste, mais non nécessaire pour un lecteur qui aurait bien compris la « preuve directe » du compatibiliste que représenterait l'analyse compatibiliste. Ou bien, elles peuvent être présentées comme une curiosité théorique supplémentaire : il est toujours intéressant, une fois qu'on a établi sa thèse, d'essayer de comprendre l'origine de l'erreur adverse. Pour Peter van Inwagen, les théories de la confusion représentent même un symptôme du « niveau pas très élevé » des discussions habituelles sur le compatibilisme.³³⁰ Il choisit donc de court-circuiter totalement la question de l'illusion : qu'il y ait ou non des arguments fallacieux ou illusoire dans un sens ou dans l'autre, tout ce qui importe pour aborder sérieusement le fond d'une problème, c'est de peser les *bons* arguments. Aussi van Inwagen se contente-t-il de présenter ses arguments incompatibilistes, qui quant à eux ne reposent sur aucune confusion.³³¹

Mais si ma présentation du problème est exacte, alors la tâche (b') n'est ni inutile, ni même « secondaire » : elle est à vrai dire la tâche centrale et essentielle du compatibiliste comme tel. Présenter des concepts compatibilistes est en effet une tâche très aisée (et la liste des concepts aujourd'hui disponibles dans la littérature est extrêmement longue), et surtout une tâche qui en elle-même n'a rien de « compatibiliste » (un incompatibiliste peut s'y intéresser pour dresser la liste de différentes conditions *nécessaires* mais non suffisantes au libre arbitre) ; ce qui est compliqué et proprement compatibiliste, c'est de relever le « défi de la pertinence », de montrer que l'un de ces concepts, malgré son caractère compatibiliste, est néanmoins *le* concept pertinent dans le débat, c'est-à-dire le concept qui a pu générer un conflit intuitif avec le déterminisme. Contrairement à van Inwagen, je considère donc que les théories de la confusion ou de l'illusion incompatibiliste sont au *cœur* de la stratégie compatibiliste : à défaut d'une telle théorie, on doit en rester (pour tout concept compatibiliste) à l'idée que ce concept n'est probablement pas le concept à l'origine du conflit intuitif. Autrement dit, c'est la théorie de l'illusion, et *seulement* la théorie de l'illusion, qui permettra au compatibiliste de relever la charge de la preuve que fait peser sur lui l'IME de la section I.

Cette présentation du problème donne à mon « argument » incompatibiliste un tour dialectique assez inhabituel. Dans la suite de ce chapitre, je ne chercherai pas en effet à établir directement que tel concept préalablement déterminé (par exemple le concept de « pouvoir faire autrement ») se révèle après analyse être incompatible avec le déterminisme. Je ne

³³⁰ « Discussions of this question are usually not on a very high level. In the great majority of cases, they are the work of compatibilists and consist to a large degree in the ascription of some childish fallacy or other to incompatibilists » (van Inwagen 1983, 55).

³³¹ « It is not my purpose in this book to defend any previous writer against a charge of fallacious argument. My own arguments will be explicit, and any fallacies they commit should be correspondingly visible. » (*ibid.*).

chercherai pas non plus à « critiquer » les différents concepts compatibilistes qui ont été mis en avant, historiquement ou récemment. Je m'attacherai simplement à montrer qu'aucune théorie de l'illusion satisfaisante n'a été apportée pour l'un ou l'autre de ces concepts : les divers concepts compatibilistes sont sans doute intéressants en eux-mêmes, mais il y a fort peu de chance qu'ils aient pu générer un conflit intuitif (illusoire) ; ce qui veut dire que le conflit intuitif qui existe *effectivement* a toutes les chances de reposer sur un *autre* concept, un concept qui, quant à lui, présente une incompatibilité réelle avec le concept de déterminisme.

III. Echec des théories compatibilistes du conflit illusoire

Les théories de l'illusion n'ont pas encore fait, à ma connaissance, l'objet d'une classification systématique. Dans cette section, je présenterai ce qui me semble être les trois grandes théories de l'illusion disponibles dans la littérature contemporaine ou classique. Et je montrerai que chacune de ces théories de l'illusion est peu plausible en elle-même ; ceci renforcera donc notre justification *prima facie* selon laquelle le conflit intuitif ne vient pas d'une illusion mais d'un conflit conceptuel véritable, c'est-à-dire d'une véritable incompatibilité entre les concepts de libre arbitre et de déterminisme.

Ces théories consistent donc à dire que nos concepts de déterminisme (Det) et de libre arbitre (LA) sont, en eux-mêmes, compatibles, mais que nous apercevons une incompatibilité à cause d'une confusion conceptuelle (soit confusion de LA avec un autre concept de liberté plus exigeant LA*, soit confusion de Det avec un concept de déterminisme plus radical Det*). Les théories de l'illusion *présupposent* donc certaines analyses minimales du concept de libre arbitre (ou certaines conditions nécessaires minimales). Ainsi, les trois grandes théories que je vais considérer présupposent que notre concept de libre arbitre implique (respectivement) : un *contrôle alternatif* de nos actions, le caractère *non nécessaire* de nos décisions, et enfin l'idée que nous sommes en un certain sens la *source* de nos actions. Ces conditions nécessaires me semblent parfaitement légitimes mais à ce stade de la discussion, je ne m'engagerai pas sur leur vérité. Je les admettrai simplement à titre conditionnel : *si* le compatibiliste a raison de reconnaître telle condition nécessaire au concept de libre arbitre, peut-il construire une théorie de l'illusion qui soit plausible ?

III.1. La théorie classique : l'erreur de catégorie de la « volonté libre »

La théorie compatibiliste classique, celle de Hobbes, Locke, Hume, ou encore Stuart Mill, part d'une analyse conditionnaliste du libre arbitre. Pour reprendre la définition de Hume déjà présentée au chapitre 1 (p.43) :

Par liberté, nous pouvons seulement entendre un pouvoir d'agir ou de ne pas agir selon les déterminations de la volonté ; c'est-à-dire, si nous choisissons de rester en repos, nous le pouvons, si nous choisissons de nous mouvoir, nous le pouvons aussi.
(Hume 1983, 164)³³²

³³² Voir également chez Locke la définition du libre arbitre comme « pouvoir d'agir ou de ne pas agir par la détermination de la pensée préférant l'un à l'autre » (Locke 1690 II, ch.21, §21).

Il est important de noter que cette analyse classique du libre arbitre inclut l'idée d'un pouvoir des contraires : je suis libre lorsque je peux agir *ou ne pas agir*, et c'est ce pouvoir des contraires – ou ce « contrôle des possibilités alternatives » – qui s'analyse *ensuite* en termes conditionnels : si je veux agir, j'agis, et si je veux ne pas agir, je n'agis pas.³³³

A partir de cette analyse conditionnaliste, la tâche du compatibiliste classique est ensuite de montrer (1) que cette analyse rend le libre arbitre compatible avec le déterminisme, et (2) que ce concept présente un risque de confusion qui explique l'*illusion* d'incompatibilisme. La première tâche est très simple : l'analyse conditionnaliste dit seulement que nos actions doivent dépendre de nos volontés ou volitions, mais si nos volitions, quant à elles, suivent un cours parfaitement déterministe, alors nos actions elles aussi suivront un cours parfaitement déterministe. Pour le dire autrement, nos actions ont une marge de manœuvre *par rapport* à nos volitions, mais comme nos volitions n'ont aucune marge de manœuvre, l'ensemble est parfaitement déterministe. Par exemple, imaginons que je sois déterminé par mon passé et les lois de la nature à vouloir prendre un verre d'eau maintenant, et que je le fasse effectivement. Dire que j'étais libre de ne pas le faire, c'est simplement dire que si je ne l'avais pas voulu (si mon passé et les lois de la nature m'avaient déterminé à ne pas le vouloir), alors je n'aurais pas pris de verre d'eau. C'est donc parfaitement compatible avec le fait de dire que mon passé et les lois de la nature me déterminaient, en fait, à vouloir prendre un verre d'eau. De fait, pour les compatibilistes classiques, nos volitions sont entièrement déterminées ; par exemple, chez Locke, c'est le « malaise » ou « désir » qui détermine notre volition d'agir, et inversement la satisfaction détermine notre volition de ne pas agir (Locke 1690 II, ch.21, §29). L'analyse conditionnaliste met donc en évidence une propriété qui est très clairement compatible avec le déterminisme. Mais cette propriété a-t-elle des chances de correspondre à notre concept de libre arbitre, c'est-à-dire au concept qui a généré un conflit intuitif avec l'idée de déterminisme ?

A première vue, si notre concept de libre arbitre n'était que le concept d'une dépendance de nos actions par rapport à nos volitions – si c'était *tout* ce que nous avons à l'esprit lorsque nous nous concevons comme libres – on ne devrait pas s'attendre à ce que ce concept génère en nous un conflit avec le déterminisme. En effet, il suffit de comparer avec d'autres concepts du même type : lorsque je comprends le mécanisme de l'horloge, je comprends que la rotation de l'aiguille dépend du mouvement pendulaire du balancier, de telle sorte que si le balancier bouge, alors l'aiguille tourne, et si il ne bouge plus, l'aiguille ne tourne plus. Lorsque j'utilise ce concept, suis-je tenté de croire que le mouvement de l'aiguille d'une horloge est incompatible avec le déterminisme simplement parce qu'il est conditionnellement dépendant du mouvement du balancier ? Absolument pas. Pourquoi donc le concept d'une dépendance conditionnelle volitions-actions aurait-il produit en nous une illusion incompatibiliste ?

La théorie classique de l'illusion repose sur « l'erreur de catégorie des volitions libres ». L'idée centrale est que nous avons une tendance à essayer d'appliquer l'attribut « libre » non seulement à des agents ou des actions, mais aussi à notre volonté elle-même ou à nos volitions, comme l'atteste l'expression « libre arbitre » ou « free will ». Mais il s'agit d'une erreur de

³³³ G.E. Moore présente très clairement l'analyse conditionnelle comme une analyse du *pouvoir d'agir autrement* et non pas directement une analyse du mot « libre » : « There is much reason to think that when we say that we *could* have done a thing which we did not do, we *often* mean merely that we *should* have done it, *if* we had chosen. » Moore en conclut (validement) que le pouvoir d'agir autrement est clairement compatible avec le déterminisme : « And if so, then it is quite certain that, in *this* sense, we often really *could* have done what we did not do, and that this fact is in no way inconsistent with the principle that everything has a cause. » (Moore 2005, 112–113).

catégorie car l'arbitre ou la volonté ne sauraient être dits « libres ». On trouve cette théorie exposée de manière particulièrement claire chez Locke :

L'erreur est venue de ce que l'on a parlé des facultés comme d'autant d'agents distincts et qu'on les a ainsi représentés : si l'on demandait ce qui digérait la viande dans l'estomac, la réponse immédiate et suffisante était *la faculté digestive* ; [...] qu'est-ce qui meut ? *la faculté motrice*. Et de même pour l'esprit : la *faculté intellectuelle* ou entendement, entendait, et la *faculté élective* ou volonté, voulait ou commandait. (Locke 1690 II, ch.21, §20)

Il est donc évident que la *volonté* n'est qu'un pouvoir ou une capacité, et la *liberté* un autre pouvoir ou capacité. Aussi, demander si *la volonté a la liberté*, c'est demander si un pouvoir a un autre pouvoir, une capacité une autre capacité, question d'emblée trop absurde pour être discutée, pour demander une réponse. (Locke 1690 II, ch.21, §16)

Je pense que la question n'est pas à proprement parler *la volonté est-elle libre ?* mais *un homme est-il libre ?* (Locke 1690 II, ch.21, §21)

Hobbes critique également l'expression « libre arbitre » ou « free will » comme une erreur de catégorie :

Lorsque les mots libre et liberté sont appliqués à toute autre chose que des corps, c'est un abus de langage ; [...] de l'usage de l'expression « libre arbitre » [free will], on ne peut aucunement déduire une liberté de la volonté [will], du désir, ou de l'inclination, mais seulement la liberté de l'homme. (Hobbes 1998, chap. 21)³³⁴

La force de John Locke est qu'il apporte un début *d'explication* de notre tendance à former l'expression (pourtant impropre) de libre arbitre (free will) : c'est que nous avons une tendance générale à parler des facultés comme de quasi-agents.³³⁵ Si nous avons une tendance à considérer la volonté comme un quasi-agent qui veut et ne veut pas, on pourra aussi se demander si elle est *libre* de vouloir ou de ne pas vouloir. A partir de cette première erreur de catégorie, on en vient à exiger que les *volitions* soient libres (et non plus seulement les actions), et c'est cette exigence qui n'est pas compatible avec l'idée que nos volitions soient entièrement déterminées par nos états passés et les lois de la nature :

la question suivante est : « Dispose-t-on de la liberté de vouloir ce qui plaît entre mouvement et repos ? » La question est en elle-même si manifestement absurde qu'elle nous convainc que la liberté ne concerne pas la volonté. Car [...] c'est demander si on peut *vouloir* ce qu'on veut ou se plaire à ce qui plaît : une question qui à mon sens n'a pas besoin de réponse. Ceux qui en font un problème doivent supposer une volonté qui déterminerait les actes d'une autre, et une autre qui déterminerait celle-là, et ainsi de suite à l'infini. (Locke 1690 II, ch.21, §25)

³³⁴ Il y a quelques nuances importantes entre la théorie de Locke et celle de Hobbes : plus que sur l'analyse conditionnaliste, Hobbes insiste sur la liberté comme « non empêchement » ; de cette analyse, il s'ensuit que Hobbes restreint l'usage du mot libre aux seuls *corps*, tandis que Locke accepte de l'employer pour les *esprits* (corps et esprits étant tous deux des substances et non des *facultés*).

³³⁵ En parlant de la « faculté digestive » et de la « faculté motrice », Locke semble s'appuyer ici sur la critique – très répandue à son époque – des pseudo-explications apportées par les formes substantielles, telles que la « vertu dormitive » de l'opium chez Molière et autres « vertu horodictique de l'horloge » chez Leibniz. Locke se donnerait ainsi le grand avantage de pouvoir ranger l'incompatibilisme dans la famille de sophisme la plus célèbre en son temps. Il faut noter cependant que la critique formulée ici par Locke est d'une nature différente à la critique des vertus dormitives : le problème de la vertu dormitive n'est *pas* un problème d'erreur de catégorie, mais un problème de vacuité de l'explication proposée.

Il faut noter que cette nouvelle question (la liberté de la *volition*) n'est pas aussi clairement une erreur de catégorie que la première (la liberté de la *volonté*), car les volitions ne sont pas des *facultés* (ce sont sans doute des événements, peut-être même des actes). Mais Locke identifie un autre problème d'inapplicabilité conceptuelle : compte tenu de l'analyse conditionnaliste, « être libre de faire A » veut dire « pouvoir faire A si je veux faire A, et pouvoir ne pas faire A si je ne veux pas faire A ». Donc « être libre de vouloir » voudrait dire « pouvoir faire vouloir si je veux vouloir etc. » or cette notion semble inintelligible.

Pour résumer, le compatibilisme classique analyse le libre arbitre comme un contrôle des possibilités alternatives de nos *actions* et non pas un contrôle des possibilités alternatives de nos *volitions* : contrôler nos alternatives d'action signifie justement que nos actions dépendent de nos volitions, mais nos volitions, quant à elles, peuvent dépendre de manière déterministe de l'ordre naturel. Et ce concept de libre arbitre, selon la théorie classique, peut produire une illusion incompatibiliste à cause de notre tendance générale à hypostasier les facultés : en prenant la volonté comme un quasi-agent, nous inventons une contrainte de « liberté de la volonté » qui se comprend ensuite comme « liberté des volitions », c'est-à-dire un contrôle des possibilités alternatives de nos volitions. Mais une telle notion est à la fois difficilement intelligible et non requise par notre concept initial de l'action libre. L'illusion consiste donc à passer du contrôle alternatif des actions (par les volitions) au contrôle alternatif des volitions : c'est cette transition illégitime que nous avons naturellement tendance à faire et qui explique la genèse du problème intuitif du libre arbitre.

Que penser de cette théorie classique de l'illusion ? Le réquisit de non-détermination des volitions est-il la conclusion illégitime provenant d'une erreur de catégorie entre agents et facultés ? On peut montrer que ce n'est pas le cas, à partir de deux considérations : tout d'abord, l'erreur de catégorie entre agents et facultés ne semble pas pouvoir être vraiment l'origine du réquisit de contrôle des volitions ; et deuxièmement, le contrôle des volitions semble être une conséquence parfaitement légitime et non pas fallacieuse *si* l'on exige un véritable contrôle des actions, ainsi que les conditionnalistes actuels le reconnaissent eux-mêmes.

Premièrement, donc, l'erreur de catégorie qui fait de la volonté un quasi-agent susceptible d'être libre a peu de chances d'être véritablement l'origine du conflit intuitif entre liberté et déterminisme. Certes, il est incontestable que nous avons une tendance à faire cette « erreur de catégorie » – les expressions de libre arbitre, de *free will*, ou même de « volonté libre » l'attestent. Mais ces expressions linguistiques n'ont pas besoin d'être employées pour faire apparaître le conflit intuitif : les expériences de X-phi ces dernières années ont permis de mettre en évidence une grande stabilité des intuitions de conflit avec l'idée d'un monde déterministe, quelles que soient les expressions employées.³³⁶ En se focalisant sur la langue anglaise, on pourrait facilement penser que l'expression « *free will* », qui est d'usage commun, a profondément inscrit dans les esprits anglo-saxons la notion d'un volonté libre, mais c'est là une particularité linguistique que l'on ne retrouve pas dans toutes les langues : « libre arbitre », en français, reste une expression rare et technique, et « volonté libre » n'apparaît probablement qu'en contexte philosophique. De fait, lorsqu'il compare la volonté libre à la faculté digestive ou à la faculté motrice, Locke semble situer l'origine du problème dans un certain langage savant. Si tel était le cas, on ne rencontrerait le conflit intuitif que chez les savants, ou chez les philosophes habitués à hypostasier les facultés. Mais il est clair, au moins d'après les résultats de X-phi, que le conflit intuitif, légitime ou non, n'apparaît pas uniquement chez les savants ou les philosophes, mais beaucoup plus largement. Ce conflit n'a pas son origine dans une certaine formulation linguistique, pas *a fortiori* dans une formulation

³³⁶ Voir notamment la recension proposée par (Sarkissian et al. 2010, 349) qui donne une idée de la diversité de formulations linguistiques de ces expériences.

linguistique *savante*. Pourrait-on étendre la théorie classique et dire que nous avons une tendance naturelle à *concevoir* les facultés comme des quasi-agents (quelle que soit la manière dont on les décrit linguistiquement), et donc à *concevoir* la notion d'une volonté libre, et partant de volitions libres ? Ce n'est pas impossible, mais c'est postuler des étapes de raisonnement implicite bien complexes chez des sujets qui rapportent simplement et immédiatement l'intuition que « l'univers déterministe A n'est pas tel que le nôtre ». En réalité, l'hypothèse de loin la plus plausible c'est que l'exigence d'un contrôle des volitions vient directement de l'exigence d'un vrai contrôle de nos actions.

Deuxièmement, si l'on accepte que le libre arbitre requiert un contrôle des possibilités alternatives de nos actions (ce qu'accepte le compatibiliste classique), alors il y a de bonnes chances pour que cette notion elle-même nous amène à un réquisit de contrôle de nos *volitions*. L'idée d'un contrôle de nos actions *sans* contrôle de nos volitions est en effet paradoxale, comme le montre un trait d'humour du film *Some like it Hot* (Wilder 1959). Sortant une bouteille de Bourbon, Marilyn Monroe se défend :

Je ne veux pas que vous pensiez que je suis une buveuse. Je peux m'arrêter si je veux, ... c'est seulement que je ne veux pas... surtout quand j'ai le cafard.

Il est clair que la défense de Marilyn Monroe n'est pas très convaincante : elle manifeste clairement qu'elle est dépendante à l'alcool, non pas certes parce que ses actions ne dépendraient pas de sa volition de boire, mais parce que manifestement elle n'est capable d'avoir *que* des volitions de boire. Pour avoir un vrai contrôle sur ses actions vis-à-vis de l'alcool, il lui faudrait récupérer un minimum de contrôle sur ses volitions, il lui faudrait être capable de *vouloir ne pas boire*. Autrement dit, le principe suivant semble intuitivement valide :

(CV) pas de contrôle sur les actions sans contrôle sur les volitions.

Pour le conditionnaliste classique, comme on vient de le voir, le conflit intuitif entre libre arbitre et déterminisme vient de ce que nous passons implicitement de (i) « X aurait pu faire autrement » à (ii) « X aurait pu *vouloir* (ou *choisir de*) faire autrement ». Et le conditionnaliste ajoute que cette transition est une simple confusion, car la phrase (i) s'analyse en termes purement conditionnels (si X avait voulu faire autrement, X aurait fait autrement). Mais si le principe (CV) est vrai – et intuitivement il semble bien qu'il le soit – alors la transition de (i) à (ii) ne repose pas sur une « confusion » mais sur une déduction valide, et c'est l'analyse purement conditionnelle qui se trouve elle-même réfutée.

Suite à ce genre de considérations, le conditionnalisme classique est aujourd'hui largement considéré comme réfuté. Historiquement, l'article le plus important a été celui de Keith Lehrer « Cans without Ifs » (Lehrer 1968), dont l'argument repose sur un scénario similaire à celui de Marilyn. Lehrer imagine qu'il est frappé d'une aversion pathologique pour les bonbons rouges (qui lui rappellent des gouttes de sang) ; quelqu'un lui propose l'un de ces bonbons, et comme on peut s'y attendre il ne choisit pas de le prendre, parce qu'il ne *peut pas* choisir de le prendre. Mais il reste vrai que *si* il avait choisi de le prendre (si il avait pu choisir de le prendre), alors il l'aurait pris. Dans ce scénario de Lehrer, l'incapacité de choisir un bonbon rouge entraîne logiquement une incapacité à le prendre effectivement :

- (1) Lehrer n'aurait pas pu choisir de prendre un bonbon rouge.
- (2) Si Lehrer ne choisit pas de prendre un bonbon rouge, alors il n'est pas possible que Lehrer prenne un bonbon rouge.
- (C) Lehrer n'aurait pas pu prendre un bonbon rouge.

Mais les deux prémisses (et donc leur conclusion) restent logiquement compatibles avec la vérité du conditionnel :

(3) Si Lehrer avait *choisi* de prendre un bonbon rouge, alors il aurait pris un bonbon rouge.

Ceci suffit à montrer que le conditionnel (3) n'est *pas* une analyse de la capacité à faire autrement (Lehrer aurait pu prendre un bonbon rouge), puisque la négation de cette capacité (C) est logiquement compatible avec (3). L'exemple de Keith Lehrer, ou d'autres semblables, ont convaincu la communauté philosophique que la dépendance conditionnelle volition-action ne pouvait pas être une condition *suffisante* pour attribuer à un agent le pouvoir d'agir autrement : pour attribuer une telle capacité, il faut *également* que l'agent ait un minimum de contrôle sur ses volitions. La plupart des auteurs compatibilistes ont donc totalement abandonné la stratégie conditionnaliste.

Certains auteurs cependant, en particulier Kadri Vihvelin (2004) ont proposé une solution inspirée du conditionnalisme et qui répond à l'objection de Lehrer, une solution que Ghislain Le Gousse (Le Gousse 2011) a baptisée néo-conditionnalisme. En suivant la formulation simplifiée qu'en donne (Le Gousse 2011, 165), l'analyse néo-conditionnaliste de Vihvelin peut être présentée de la manière suivante :

- (NC) S aurait pu faire B (au lieu de A) =_{df}
- (i) si S avait choisi de faire B (au lieu de A), S aurait fait B (au lieu de A)
 - (ii) S avait la base causale requise pour choisir de faire B (au lieu de A)

On voit clairement que la clause (ii) sert ici à pallier l'insuffisance de la clause conditionnelle traditionnelle (i), et on peut remarquer également que cette clause supplémentaire consiste à conférer à l'agent S une certaine sorte de « contrôle » sur le *choix* de l'action B (une sorte de contrôle dont ne dispose pas Lehrer à l'égard des bonbons roses). Autrement dit, l'analyse néo-conditionnaliste tire les leçons des exemples à la Lehrer en acceptant une certaine version du principe (CV) :

(CV') pas de contrôle de l'action sans la *base causale requise* pour choisir autrement.

Une telle analyse ne revient-elle pas à abandonner l'esprit de l'analyse conditionnaliste ? Pas nécessairement ; cette nouvelle analyse préserve de l'analyse traditionnelle un élément important : elle préserve le fait que l'agent n'a pas besoin de « pouvoir vouloir » (ou de « pouvoir choisir »). En effet, la clause (ii) de l'analyse (NC) doit être distinguée d'une clause de contrôle alternatif des volitions : « S *aurait pu* choisir de faire B (au lieu de A) ». La clause (ii) est beaucoup plus faible que cela. On pourrait dire que, dans cette analyse, le contrôle alternatif des actions requiert *seulement* une base causale pour les volitions et *non pas* un véritable contrôle alternatif des volitions. Et par conséquent, le néo-conditionnaliste peut préserver l'esprit de la théorie classique de l'illusion :

Théorie néo-conditionnaliste de l'illusion

Pour qu'un agent soit libre à l'égard de l'action A, il faut qu'il dispose d'un contrôle fort à l'égard de son action A (qu'il ait pu faire autrement que A), et ce contrôle fort de l'action suppose une certaine forme de contrôle (faible) à l'égard de sa volition (il faut qu'il ait eu la *base causale* requise pour former la volition de faire autrement que A). Cependant, nous avons tendance à confondre les deux types de contrôle et à exiger un

contrôle fort à l'égard de nos volitions, ce qui serait incompatible avec le déterminisme causal et entraîne donc l'illusion incompatibiliste.

L'esprit de cette critique reste donc le même que dans le conditionnalisme classique : on aurait à tort transféré sur les volitions la notion de contrôle qui s'applique proprement aux actions (ce qui n'empêche pas qu'une *autre* notion de contrôle est requise pour les volitions).

On peut apporter contre la théorie néo-conditionnaliste une objection très semblable à l'objection de Lehrer : il n'est pas clair du tout que les conditions (i) et (ii) de (NC) soient *suffisantes* pour attribuer à un agent le contrôle de son action. En particulier, si l'on imagine un agent qui a la base causale requise pour choisir de faire B (au lieu de A) *mais* qui pour d'autres raisons ne *peut pas* faire B (au lieu de A), il n'est pas clair du tout qu'on serait tenté de décrire cet agent comme pouvant faire B (au lieu de A). Pour bien voir cette objection, on peut partir d'un scénario à la Frankfurt :

Le neurochirurgien maléfique Black veut la mort de Smith. Il apprend que Jones a justement le projet d'assassiner Smith. Afin de s'assurer qu'il exécutera ce projet, Black a implanté dans le cerveau de Jones (à l'insu de ce dernier) un appareil permettant de contrôler son comportement. Il compte surveiller la délibération de Jones : si Jones choisissait de ne pas tuer Smith, il le laisserait s'abstenir ; mais cela n'arrivera pas, car si Black se rend compte que Jones est sur le point de choisir de ne pas tuer Smith, il le forcera à choisir d'appuyer sur la gâchette ; sinon, il le laissera faire. Finalement, Jones prend de lui-même et exécute la décision de tuer Smith. (Cf. Le Gousse 2011, 166–167)

Dans ce scénario, la condition (i) est explicitement respectée (si Jones choisissait de s'abstenir, il s'abstiendrait), et la condition (ii) l'est également (Jones a la base causale requise pour choisir de s'abstenir). Pourtant, on ne serait pas tenté de dire que, dans cette situation, Jones *pourrait* ne pas tuer Smith, et la raison de cette réticence semble bien être le fait que Jones ne *pourrait pas choisir* de ne pas tuer Smith. Vihvelin est obligée de dire que dans ce scénario Jones pourrait ne pas tuer Smith, et elle accepte explicitement cette conséquence de sa théorie. Mais l'important est de noter que cette conséquence est au minimum contre-intuitive ; intuitivement, il semble bien qu'on ait le principe suivant :

(CV'') pas de contrôle alternatif de l'action sans *contrôle alternatif* des volitions.

Et par conséquent, si le conflit intuitif entre libre arbitre et déterminisme repose sur le passage d'un réquisit de contrôle alternatif de l'action à un réquisit de contrôle alternatif de nos volitions, il y a toutes les chances (au moins à première vue) pour que ce passage soit une inférence valide et non pas une simple confusion entre l'action et la volition. Autrement dit, le néo-conditionnaliste, comme le conditionnaliste, ne permet pas de mettre en avant une possibilité de confusion initialement plausible. A défaut d'une possibilité plausible de confusion, nous devons en rester à l'explication initialement la plus simple, à savoir que le conflit intuitif repose sur une *véritable* incompatibilité conceptuelle et non pas sur une illusion.

Pour conclure cette section, rappelons qu'il n'est pas absolument évident que le concept de libre arbitre requiert un contrôle alternatif des actions (le pouvoir d'agir autrement) ; nous ne l'avons admis ici qu'à titre d'hypothèse parce que le compatibilisme conditionnaliste en fait le fondement de sa théorie de l'illusion. Pour lui, notre concept de libre arbitre implique le pouvoir d'agir autrement, et l'intuition incompatibiliste viendrait de ce que nous passons implicitement du pouvoir d'agir autrement à un pouvoir de *vouloir* agir autrement. Il ajoute enfin que cette déduction est fallacieuse. Le problème, c'est que nous n'avons aucune raison

indépendante de considérer cette déduction comme fallacieuse : au contraire, lorsque nous l'étudions avec attention, tout porte à croire que cette déduction est parfaitement légitime et repose sur un principe vrai : pas de contrôle des actions sans contrôle des volitions. Le compatibilisme conditionnaliste échoue donc à présenter une théorie de l'erreur plausible et indépendamment motivée. On pourra *éventuellement* accepter sa théorie de l'erreur *si* on doit ultimement recourir au compatibilisme pour sauver le libre arbitre ; mais à défaut de cette motivation, cette théorie de l'erreur ne semble pas initialement plausible et nous en restons à l'idée d'un conflit probablement profond entre libre arbitre et déterminisme.

III.2. La théorie modale de Robert Audi : confusion de la nécessité du conséquent

Une deuxième théorie de l'illusion est celle proposée par Robert Audi dans l'article « Modalities of Knowledge and Freedom » (Audi 1993). De même que Locke appuyait sa théorie sur une confusion bien connue indépendamment du libre arbitre (la confusion consistant à hypostasier des facultés et à en faire des quasi-agents), de même Robert Audi s'applique à montrer que l'intuition incompatibiliste pourrait reposer sur un cas de confusion classique, la confusion de la nécessité du conséquent et de la nécessité de la conséquence (ou pour faire plus court : confusion NCNC). Je montrerai cependant que l'intuition incompatibiliste a peu de chance de reposer sur une telle confusion : au contraire, si on admet avec Audi que la nécessité de la conséquence joue un rôle dans notre intuition incompatibiliste, il y a toutes les raisons de penser que le raisonnement en jeu soit un cas de « transfert de la nécessité », c'est-à-dire un raisonnement parfaitement valide.

Qu'est-ce qu'une confusion NCNC et comment aurait-elle pu générer notre intuition incompatibiliste ? Une confusion NCNC est un cas très célèbre de sophisme modal qui repose, au moins en partie, sur une ambiguïté linguistique. Lorsque j'affirme une proposition de la forme « si p alors nécessairement q » – par exemple, « si Jean vient, alors nécessairement Paul viendra aussi », ou « si $2+2=4$ est une vérité mathématique, alors nécessairement $2+2=4$ » – je peux vouloir dire deux choses différentes selon la proposition sur laquelle porte l'opérateur de nécessité. En apparence, l'opérateur de nécessité porte sur la proposition q (le conséquent), et on a donc une proposition de la forme :

$$(NCt) p \supset \Box q$$

Le deuxième exemple peut se lire de cette manière là : s'il est vrai que « $2+2=4$ est une vérité mathématique », alors la proposition « $2+2=4$ » est une proposition marquée par la nécessité (les vérités mathématiques ne sont pas vraies de manière simplement contingente). Mais cette lecture n'est pas acceptable pour le premier exemple : dans la phrase « si Jean vient, alors nécessairement Paul viendra aussi », il ne s'agit pas d'attribuer la nécessité à la proposition « Paul viendra » ; cette proposition, qu'elle soit vraie ou fausse, est clairement une proposition contingente. La nécessité, en l'occurrence, porte non pas sur le conséquent (si p , alors nécessairement- q), mais sur le lien de conséquence (si p , alors-nécessairement q). On a donc une proposition de la forme :

$$(NCce) \Box (p \supset q)$$

Une confusion NCNC consiste alors à prendre une nécessité de la conséquence (NCce) pour une nécessité du conséquent (NCt). C'est une confusion extrêmement classique qui a été identifiée dès l'époque médiévale.

Robert Audi donne comme premier exemple une ambiguïté en philosophie de la connaissance que nous avons déjà rencontrée au chapitre 2 (p.132). La phrase suivante semble dire quelque chose de vrai :

(~PE) Si S sait que p , alors S ne peut pas se tromper.

Mais cette phrase est ambiguë. Si on donne à l'opérateur modal la portée étroite, on arrive à principe épistémologique de nécessité qui est extrêmement coûteux :

(N) $Kp \supset \sim \Diamond \sim p$, ou encore $Kp \supset \Box p$

Ce principe signifie qu'il ne peut y avoir de connaissance *que* des vérités nécessaires, par exemple des vérités mathématiques ou métaphysiques. Il n'y a donc aucune sorte de connaissance pour les vérités contingentes (pas même de ma propre existence, qui est certainement contingente). Mais si on lit la phrase (~PE) selon la nécessité de la conséquence, on obtient un principe qui est incontestablement vrai et très peu coûteux, à savoir le principe de factivité de la connaissance :

(F) $\sim \Diamond (Kp \wedge \sim p)$, ou encore $\Box (Kp \supset p)$

C'est-à-dire que si une proposition peut être dite « connue » par S, alors elle est vraie. Si l'on accepte intuitivement (~PE), c'est sans doute à cause de la vérité évidente de (F), mais une confusion NCNC pourrait nous amener à accepter la lecture (N).

Un cas sans doute plus typique de confusion NCNC est celui qui peut mener au fatalisme logique en philosophie du libre arbitre. Un argument simpliste (et fallacieux) en faveur du fatalisme est le suivant :

Sophisme fataliste

Entre les deux propositions « il y aura une bataille navale demain » et « il n'y aura pas de bataille navale demain », une seule est vraie. Supposons que ce soit la première (le même raisonnement s'appliquera si c'est la seconde). Il est donc vrai que « il y aura une bataille navale demain ». Mais si cette proposition est vraie, alors il n'est pas possible qu'il n'y ait pas de bataille navale demain. Donc il n'est pas possible qu'il n'y ait pas de bataille navale demain. C'est-à-dire que cette proposition concernant le futur est aussi une proposition *nécessaire*. Et le raisonnement s'applique ainsi à toute autre proposition concernant nos actions futures. Donc toutes nos actions futures sont nécessaires, aucune n'est libre.

Le sophisme, dans ce raisonnement, concerne évidemment la phrase « si cette proposition est vraie, alors il n'est pas possible qu'il n'y ait pas de bataille navale demain ». Selon la nécessité de la conséquence, cette phrase exprime simplement le truisme de la vérité :

(T) $\Box (Tp \supset p)$

Mais le raisonnement fataliste ne peut pas fonctionner si l'on emploie ce principe. Le raisonnement repose en effet sur une lecture de nécessité du conséquent :

(Fat) $Tp \supset \Box p$
 or Tp
 donc $\Box p$

Il est clair que le principe (Fat) est extrêmement coûteux (il veut dire que *toute* vérité est une vérité nécessaire) et que bien peu de gens l'admettraient. S'il *semble* qu'il y ait là un argument fort, c'est donc seulement à cause de la confusion entre (T) (qui est trivialement vrai) et (Fat).

Pour Robert Audi, l'intuition incompatibiliste repose sur le même genre de sophisme. La confusion concernerait en l'occurrence le lien que le déterminisme établit entre mon action et ses causes (et notamment ses causes déterminantes lointaines, puisque, d'après la thèse déterministe, on peut trouver pour n'importe quel événement présent un ensemble de causes déterminantes à n'importe quel moment du passé, si lointain soit-il). Le déterminisme permet donc d'affirmer quelque chose comme le « principe de nécessité causale » suivant :

(1) « Si mon action (A, par exemple) est causée, alors elle doit se produire (d'une nécessité causale, par opposition à une nécessité logique) » (Audi 1993, 257)

Mais ce principe est ambigu. Ce que la thèse déterministe affirme vraiment, c'est la lecture de nécessité de la conséquence, que Audi appelle « principe de détermination causale » :

(2) « Il est causalement nécessairement que : si les causes du fait que j'accomplis A se produisent, alors j'accomplis effectivement A ». (*ibid.*)

Ce que la thèse déterministe n'affirme *pas* en revanche, c'est la lecture de nécessité du conséquent, que Audi appelle « principe de nécessité causale » :

(3) « Si les causes [du fait que j'accomplis A] se produisent, alors il est (causalement) nécessaire que je fasse A. » (*ibid.*)

On peut voir aisément pourquoi le principe (3) ne s'ensuit pas du principe (2) : le fait qu'il y ait une connexion nécessaire entre un événement et sa cause n'implique pas que l'événement soit nécessaire, tout simplement parce que la cause elle-même peut être contingente. Par exemple, si la détonation d'un revolver est la conséquence nécessaire du fait que j'ai appuyé sur la gâchette, il ne s'ensuit clairement pas que la détonation était un événement nécessaire puisque j'aurais très bien pu ne pas appuyer sur la gâchette ; la conséquence, aussi bien que sa cause, est donc en elle-même contingente, bien que le lien causal entre les deux soit nécessaire. Pour Robert Audi, c'est la confusion entre (2) et (3) qui génère l'intuition incompatibiliste³³⁷ : sur la base du principe (1), nous concluons hâtivement que, dans le déterminisme, nos *actions* elles-mêmes « doivent » se produire de nécessité causale, alors que la nécessité (causale ou nomologique) porte sur la *relation* entre nos actions et leurs causes, et pas sur nos actions. « Mon action peut donc être nomologiquement causée, c'est-à-dire 'déterminée', sans être nomologiquement nécessaire » (Audi 1993, 259).

Un point qui reste implicite, parce que assez évident, dans la théorie de Audi, c'est que si une action était en elle-même nécessaire, *alors* elle ne serait pas libre. Autrement dit, la théorie complète de Robert Audi consiste à dire ceci : notre concept d'action libre est

³³⁷ Plus précisément, Robert Audi admet qu'il puisse y avoir d'autres arguments ou raisonnements menant à l'incompatibilisme, mais il soutient que la confusion NCNC est la manière « naturelle » ou commune de parvenir à l'intuition incompatibiliste : « I believe that the modal considerations discussed above undermine some of the plausibility of incompatibilism, not because there is no case for it that does not run afoul of them, but because the position simply becomes less plausible when it cannot be viewed in the way it is naturally seen » (p.262)

logiquement incompatible avec le concept de *nécessité* de l'action ; le concept de déterminisme quant à lui implique la *détermination* causale (ou *nécessitation*) de l'action et non pas sa nécessité, mais une confusion bien connue génère en nous l'impression que le déterminisme implique la nécessité de l'action et qu'il est donc incompatible avec le libre arbitre.

Que penser de cette théorie de l'illusion ? Le problème de cette théorie, pour le dire de manière générale, c'est qu'il existe dans le voisinage du sophisme NCNC une autre forme de raisonnement (le transfert de nécessité) qui est quant à lui parfaitement légitime, et qu'il semble y avoir de bonnes chances que ce raisonnement soit ici en jeu et non pas le sophisme NCNC. Je présenterai donc rapidement cet autre type de raisonnement avant de montrer pourquoi l'intuition incompatibiliste a de bonnes chances d'en être un exemple.

On vient de voir, à partir de l'exemple de la détonation de revolver, que la conséquence nécessaire d'un événement pouvait fort bien n'être pas, en elle-même, nécessaire. En l'occurrence, ce qui *permet* à la détonation du revolver de n'être pas un événement nécessaire, bien qu'il soit *nécessité* par le mouvement de la gâchette, c'est que le mouvement de la gâchette n'est *pas* quant à elle nécessaire. Autrement dit, si p est contingent, alors q une conséquence nécessaire de p – telle que $\Box(p \supset q)$ – peut très bien être contingente également. Mais que se passe-t-il si p n'est *pas* contingente. Que se passe-t-il lorsqu'on a *à la fois* la nécessité de la conséquence et la nécessité de l'antécédent ? Dans cette situation, s'il s'agit d'un opérateur de nécessité logique dans les deux cas, alors la logique modale nous enseigne qu'on peut *validement* tirer la conclusion de la nécessité du conséquent : $\Box q$. Autrement dit, le schéma d'argument suivant est un schéma logiquement valide, qu'on appelle un « transfert de nécessité » (car la nécessité de l'antécédent se « transfère » au conséquent, *via* la nécessité de la conséquence) :

Transfert de la nécessité logique :

(NCc)	$\Box(p \supset q)$	(prémisse 1)
(Na)	$\Box p$	(prémisse 2)
(C)	$\Box q$	

Comparons ce schéma avec le sophisme NCNC :

Sophisme NCNC

(NCc)	$\Box(p \supset q)$	(prémisse 1)
#(NCt)	$p \supset \Box q$	(par confusion avec NCc)
(Na)	p	(prémisse 2)
(C)	$\Box q$	

On voit rapidement que la différence entre le sophisme NCNC et le raisonnement valide de transfert de la nécessité tient à la présence (ou à l'absence) de la modalité de nécessité pour l'antécédent (prémisse 2).

Le raisonnement incompatibiliste repose-t-il alors sur le sophisme ou sur le raisonnement valide ? Tout dépendra de la présence ou de l'absence d'une prémisse de nécessité de l'antécédent. Remarquons tout d'abord que le sophisme NCNC n'est pas si universel que Audi semble le penser : il y a des cas dans lesquels nous ne sommes pas du tout enclins à faire un tel sophisme ; ce sont les cas où il est *évident* que l'antécédent est contingent. Prenons l'exemple de quelqu'un qui annonce son intention d'aller à la piscine demain, et raisonne ainsi : « je vais aller à la piscine demain ; si je vais à la piscine demain, alors inévitablement j'aurai les cheveux mouillés ; donc inévitablement j'aurai les cheveux

mouillés demain ». Personne ici ne serait enclin à tirer la conclusion (le caractère inévitable dès aujourd'hui des cheveux mouillés demain) ; on répondra plutôt « tu n'as qu'à changer d'avis, ne pas aller à la piscine, et dans ce cas tu n'auras pas les cheveux mouillés demain ». Autrement dit, dans ce genre de cas où l'antécédent est clairement contingent, il nous saute aux yeux que sa conséquence nécessaire est également contingente. Pour que nous déduisions la nécessité du conséquent, il semble donc qu'il doive y avoir au moins un doute sur la nécessité de l'antécédent. Est-ce le cas pour le raisonnement qui, d'après Audi, est à l'origine de l'incompatibilisme, c'est-à-dire le raisonnement suivant :

- (4) (d'après le déterminisme) si les causes lointaines de mon action A se produisent, alors nécessairement j'accomplis l'action A
- (5) *donc* nécessairement j'accomplis l'action A

L'occurrence des causes lointaines de mon action présente-t-elle au moins un risque d'être nécessaire, au moins en un certain sens ? La réponse est clairement oui : il y a un sens évident dans lequel les événements passés sont « nécessaires », même si, au moment où ils ont eu lieu, ils étaient contingents et pouvaient ne pas se produire. Une fois qu'ils se sont produits (en tant qu'événement passés), ils ne peuvent plus être altérés. C'est ce qu'on a appelé traditionnellement la « nécessité historique ». Cette nécessité historique a des conséquences importantes en philosophie de l'action : comme aucun événement passé ne peut être modifié par notre action présente, il n'est pas rationnel de chercher à modifier le passé, et on ne peut pas non plus être tenu pour « libre » à l'égard d'un événement passé.³³⁸ Donc si l'on prête attention au conditionnel sur lequel repose, d'après Robert Audi lui-même, l'intuition incompatibiliste, on s'aperçoit que l'antécédent de ce conditionnel est marqué par une sorte de nécessité qui est incompatible avec le libre arbitre. Et si on compare ce cas avec celui de la piscine, on a toutes les raisons de penser que c'est *précisément* à cause de ce caractère nécessaire de l'antécédent que nous tirons la conclusion de la nécessité du conséquent – c'est-à-dire la nécessité de toutes nos actions si le déterminisme est vrai. Autrement dit, notre raisonnement implicite n'aurait pas la forme d'un sophisme NCNC (raisonnement sans nécessité de l'antécédent), mais bien plutôt la forme d'un transfert de nécessité.

Pour être précis, il faut remarquer que les opérateurs de nécessité que l'on peut ainsi mettre en évidence ne sont pas l'opérateur de nécessité logique. Et donc le raisonnement de transfert de la nécessité, quant bien même ce serait *effectivement* le raisonnement que nous faisons, ne peut pas être considéré comme trivialement valide. D'après ce que nous venons de dire, la conséquence est marquée par une nécessité d'ordre causal ou nomologique, et l'antécédent (les événements passés) est marqué par la « nécessité historique ». Ce que ces deux types de nécessité ont en commun, du point de vue de la philosophie de l'action, c'est qu'ils impliquent tous les deux une forme de « nécessité pratique », l'inaltérabilité (untouchability) : les lois de la nature, du fait qu'elles sont nomologiquement nécessaires, sont absolument inaltérables pour nous ; il n'y a rien que nous puissions faire pour les modifier ; elles ne dépendent en rien de nous. De même, du fait que les événements passés sont historiquement nécessaires, ils sont aussi hors de portée des agents, *a fortiori* des agents humains. Si l'on note *N* cette nécessité pratique ou inaltérabilité, *P* une proposition décrivant l'état complet du monde à un moment passé, et *A* pour n'importe quelle action humaine, on retrouve donc les deux prémisses suivantes :

³³⁸ S'il s'agit d'une action que j'ai accomplie autrefois, j'étais peut-être libre à *ce moment-là*, mais je ne peux plus être tenu pour libre *maintenant* à l'égard de cette action. On peut me reprocher de l'avoir commise, mais on ne peut pas *m'enjoindre* de ne pas l'avoir commise.

(6) $N P$

(7) $N (P \supset A)$

Deux ces deux prémisses, il semble intuitif, au moins initialement, qu'on puisse tirer la conséquence :

(8) $N A$

c'est-à-dire que nos actions sont inaltérables *justement* dans le sens qui exclut la liberté. Accepter ce raisonnement, c'est en appeler à un principe de déduction que Peter van Inwagen a appelé la règle (β) :

(β) $N p \ \& \ N (p \supset q) \supset N q$

Et, à quelques modifications près, l'argument qui établit l'incompatibilisme à partir des prémisses (6), (7) et (β) correspond au fameux « argument de la conséquence » du même Peter van Inwagen, un des arguments incompatibilistes les plus célèbres et influents dans la discussion contemporaine. Remarquez que mon propos n'est pas ici réellement de défendre positivement l'argument de la conséquence. Il me suffit qu'il soit *possiblement* valide. En effet, mon propos est d'abord et avant tout de réfuter la théorie de l'illusion NCNC. Pour cela, il me faut montrer que notre raisonnement incompatibiliste intuitif a beaucoup plus de chances de ressembler à l'argument de la conséquence qu'à une confusion NCNC – parce que nous sommes dans un cas où nous *avons* une forme de nécessité de l'antécédent. Une fois qu'on a établi que le raisonnement avait probablement cette forme, il reste à se demander si cette forme d'argument est une « pure confusion ». Et la réponse est clairement non : certes, on peut chercher à mettre en doute la validité de la règle (β) – et il y a sur cette question de nombreuses discussions. Mais on ne peut pas dire qu'accepter cette règle soit un pur sophisme ou une confusion évidente. Si cette règle est discutable, il semble même qu'elle soit initialement plausible pour au moins deux raisons.

Une première raison d'accepter la validité de la règle (β) est d'ordre formel. L'opérateur N , on l'a dit, est un opérateur de nécessité. En logique modale contemporaine, la règle (β) correspond donc à l'axiome (K) :

(K) $[\Box (p \supset q) \ \& \ \Box p] \supset \Box q$

Or cet axiome (K) est un axiome si fondamental qu'il est vérifié dans *tous* les systèmes modaux normaux, que ce soit pour la modalité logique, pour la modalité métaphysique, pour la modalité déontique, etc. Les différents opérateurs modaux diffèrent pour certains axiomes importants, mais pour ce qui est de l'axiome (K), il semble qu'on ne puisse pas s'en passer si l'on veut construire un opérateur modal cohérent. Il est donc très naturel de penser que l'opérateur de nécessité pratique, ou d'inévitabilité, vérifie ce principe fondamental des logiques modales.

Une deuxième raison d'accepter (β) est intuitive : lorsqu'on considère ce que ce principe signifie, il est difficile de ne pas avoir au moins initialement l'intuition de la vérité de ce principe. En effet, ce principe signifie simplement ceci : si je ne peux absolument rien changer à un certain fait, je ne peux rien changer non plus à ses conséquences inévitables. Par exemple, si je ne peux absolument rien changer à la chute d'une météorite, et si cette chute de météorite implique inévitablement (d'après les lois de la nature) qu'un cratère de dimensions D se formera, alors je ne peux rien changer au fait qu'un cratère de dimensions D se formera.

Ces deux raisons ne sont pas suffisantes pour établir de manière certaine la validité de (β). De fait, c'est sur cette règle que les compatibilistes ont fait porter l'essentiel de leurs critiques (Voir notamment Kapitan 2002). Une partie de cette discussion consiste à interroger le sens de l'opérateur d'inévitabilité N : dans certains de ces sens, la règle (β) n'est clairement pas valide, dans d'autres, elle est clairement valide mais peut soulever des problèmes pour les prémisses (6) et (7). Pour nous, le problème serait surtout de savoir si la règle (β) est valide pour le sens d'inévitabilité qui correspond à notre concept commun de libre arbitre, c'est-à-dire : $N^{St}p = p$ et S n'est pas libre de faire que $\sim p$ à t . Notre concept de sens commun rend-il valide la règle (β) ? Une manière de répondre à cette question serait d'avoir une caractérisation directe du contenu de notre concept et de voir si ce contenu vérifie (β). Mais pour l'instant nous n'avons pas encore d'analyse directe du concept de libre arbitre. Nous pouvons seulement accéder à notre concept via les inférences qu'il semble légitimer, et c'est de ce point de vue que l'argument de la conséquence est particulièrement fort, ainsi que le souligne Manuel Vargas.

Ce qui rend ces arguments puissants, ce n'est pas tellement qu'ils excluent la possibilité du compatibilisme, mais plutôt qu'ils montrent à quel point l'incompatibilisme semble capturer aisément nos manières ordinaires de penser à propos de notre agentivité.

Pour voir pourquoi ce point assez inoffensif est important, considérer la critique qui est peut-être la plus répandue contre l'argument de la conséquence et ses successeurs. De nombreux critiques ont souligné le fait qu'une lecture antécédemment compatibiliste des termes de pouvoir rend l'argument non concluant contre des compatibilistes antécédemment convaincus. Supposez que ces critiques soient justes. Si c'est le cas, alors cela voudrait dire que les arguments du type argument de la conséquence ne peuvent pas exclure la possibilité du compatibilisme. A partir de là, on pourrait croire que l'état de la discussion rencontre au mieux à une impasse, au pire un vrai problème pour l'incompatibilisme. Ici, la question inoffensive joue un rôle : le caractère « naturel » ou la facilité à lire l'argument de manière incompatibiliste sont en eux-mêmes une raison de penser que l'argument capture une part importante des contenus et de la logique de notre manière de sens commun de penser à ces questions. Même si les contenus de l'argument n'excluent pas une lecture compatibiliste, le caractère naturel des lectures incompatibiliste de l'argument suggère *fortement* que nous pouvons comprendre et comprenons effectivement ces questions de la manière incompatibiliste, au moins quelquefois. (Vargas in Fischer et al. 2007, 132–133)

Pour conclure, il est délicat d'établir de manière certaine si l'argument de la conséquence est ultimement concluant ou non. Mais ce qui nous intéresse dans notre discussion présente, ce n'est pas tant de savoir si l'argument de la conséquence est concluant, c'est de savoir si notre intuition incompatibiliste repose ou non sur un raisonnement clairement fallacieux, et en particulier sur une confusion NCNC. A cette question, la réponse est clairement non : si l'on accepte (avec Audi) que notre intuition incompatibiliste repose sur un raisonnement de type *modal*, alors il est extrêmement peu probable que ce raisonnement modal soit une inférence fallacieuse NCNC ; il est très probable au contraire qu'il s'agisse d'un raisonnement du type « argument de la conséquence ». Et cet argument, qu'il soit ultimement concluant ou non, ne peut pas être considéré comme une évidente confusion ; c'est au contraire un des arguments les plus *forts* en faveur de l'incompatibilisme, et sa simple force intuitive semble indiquer que notre concept commun lui-même est effectivement incompatibiliste.

III.3. Dennett et les « grands méchants loups » (Bugbears) : confusions sur la notion de déterminisme

La dernière théorie de l'illusion consiste à dire que notre intuition incompatibiliste reposerait sur une confusion de la situation déterministe avec des situations plus fortes ou plus radicales qui excluraient effectivement le libre arbitre. Cette dernière théorie est plutôt une famille de théories, compte tenu du large éventail possible de candidats possibles. Daniel Dennett (Dennett 1984a, chap. 1) a donné de ces sortes de théories une caractérisation imagée : pour lui, les situations radicales que l'on a tendance à confondre avec le déterminisme ne sont que des « monstres à effrayer les enfants » (Bugbears) ou des « grands méchants loups ». Ces grands méchants loups sont suscités, consciemment ou inconsciemment, pour produire l'intuition incompatibiliste. Mais il suffit de bien les mettre en lumière pour dissiper leur pouvoir terrifiant. On peut distinguer deux catégories de situations « grand méchant loup », selon qu'elles engagent ou non l'intervention d'un *agent* manipulateur. Dans les situations qui n'engagent pas d'agent manipulateur, on peut classer, notamment, les théories selon lesquelles le déterminisme est confondu avec : le fatalisme, le mécanisme, ou le déterminisme psychologique. Dans les situations qui engagent des agents manipulateurs – ce que Dennett appelle des « croque-mitaines » (*bogeymen*), les deux théories les plus importantes sont les suivantes : soit le déterminisme est comparé à un agent *manipulateur* de mon action (par exemple un scientifique très puissant qui déterminerait chacun de mes faits et gestes), soit il est comparé à une sorte de gendarme qui me donnerait des *ordres*. Je vais présenter rapidement ces différentes théories de la confusion, puis je défendrai que la force intuitive des « grands méchants loups » ne repose probablement pas sur un simple *confusion*, mais plutôt sur l'intuition positive qu'il n'y a pas de différence vraiment pertinente entre ces situations « terrifiantes » et le scénario déterministe.

La théorie du grand méchant loup la plus classique est certainement la théorie de la confusion entre déterminisme et fatalisme. On la trouve par exemple chez John Stuart Mill (1974, bk. vi, ch.2, §3). Mill oppose la « doctrine de la nécessité » (le déterminisme) à la « doctrine du fatalisme ». La doctrine de la nécessité stipule que les actions humaines sont (causalement) déterminées par les circonstances extérieures et par les motifs de l'agent à un moment donné du temps. La doctrine du fatalisme, quant à elle, consiste à dire que les actions humaines sont déterminées « quoi que l'on fasse », c'est-à-dire en fait *indépendamment* de circonstances extérieures antécédentes ou des motivations que nous pourrions affecter. Par exemple, le meurtre de Laios par Œdipe est un cas de fatalisme car, quoi qu'Œdipe ait pu vouloir ou faire, il aurait tué Laios. Cette issue était fixée indépendamment des motivations et des actions antécédentes d'Œdipe.³³⁹ Mais il est clair que nos actions ne sont pas fixées dans

³³⁹ Ce sens du mot « fatalisme » n'est pas le seul que l'on trouve dans la littérature. C'est ce qu'on peut appeler le « fatalisme ponctuel » : il existe un événement ponctuel tel que, quelles que soient les routes (non déterminées) qu'emprunte l'agent, chacune de ces routes mènera d'une manière ou d'une autre à ce même résultat. Van Inwagen (1983, 25) appelle un tel événement « fortement inévitable ». L'autre sens important, qu'on peut appeler « déterminisme du Destin », consiste à dire que tous les événements sont « écrits », déterminés à l'avance par le « petit livre » des Parques. Dans ce sens de déterminisme, il ne s'agit pas de dire que tel événement se produira dans tous les chemins que l'agent pourrait prendre : l'agent ne peut prendre de toutes les façons qu'un seul et unique chemin ; chaque étape du chemin est déterminée, et pas seulement un certain point d'arrivée. Il faut noter que l'exemple d'Œdipe, comme exemple du fatalisme ponctuel, est contestable : d'après (van Inwagen 1983, 24), l'intérêt de ce genre d'histoire réside précisément dans le fait que l'événement prophétisé *n'aurait pas* eu lieu si l'agent n'avait pas essayé de l'éviter (si Œdipe n'avait pas essayé de s'éloigner de son lieu de naissance supposé, il n'aurait pas croisé son véritable père) ; si cela est vrai, alors on ne peut pas dire que le meurtre de Laios aurait eu lieu *quoi qu'Œdipe ait pu faire* ; mais il était déterminé qu'il se produirait parce qu'il était déterminé qu'Œdipe prendrait le chemin de Thèbes pour éviter de croiser son père.

ce sens là : au contraire, nos actions *dépendent* des motivations et des circonstances antécédentes, de telle sorte que *si* nous changions nos motivations, *alors* l'action serait différente. D'après Mill, la confusion entre fatalisme et déterminisme était à l'origine des intuitions incompatibilistes qu'il avait eues lui-même par le passé (Mill 1981, 177). Cette théorie de la confusion est sans doute assez proche de la théorie de Robert Audi : dans l'une et l'autre, l'essentiel repose sur une confusion entre nécessité conditionnelle (des événements par les circonstances) et nécessité inconditionnelle de l'acte. La théorie de la confusion fataliste se distingue seulement par le fait qu'elle donne un modèle de la manière dont les actes sont nécessaires (ils sont nécessaires par destin, quelque chemin qu'on emprunte pour y arriver). Si nous avons une tendance à confondre le déterminisme avec un fatalisme œdipien, alors on peut défendre que nos intuitions incompatibilistes viennent de cette confusion, dans la mesure où le fatalisme œdipien est effectivement incompatible avec le libre arbitre.

Deuxième théorie du grand méchant loup : la confusion du déterminisme avec le « mécanisme ». J'entendrai ici le terme « mécanisme » dans un sens relativement large : la description d'un phénomène pourra être dite « mécaniste » si elle ne fait aucun usage du champ conceptuel de l'action humaine, mais seulement du champ conceptuel de la causalité physique ou biologique. Le mécanisme, à propos de l'action humaine, reviendrait à dire que la bonne explication de nos actions, l'explication fondamentale et complète, est une description qui met à l'écart toute notion d'intentionnalité, de réflexion conceptuelle, de point de vue en première personne, et ne contient que des concepts de causes et d'effets physiques, biologiques, neuronaux, etc. Cela peut vouloir dire deux choses. Soit que les descriptions intentionnelles ou en première personne ne sont absolument pas pertinentes pour décrire la réalité : nous sommes entièrement comparables à des robots, ou à des insectes inconscients ; c'est ce que Dennett a appelé la « sphexishness » (Dennett 1984a, 10–11), c'est-à-dire la « condition de la guêpe ». Soit les descriptions en première personne sont pertinentes pour décrire nos pensées, mais nos pensées n'ont aucun effet dans l'explication causale de nos actions : c'est ce que Nahmias a appelé le « bypassing » ou « court-circuit » (Nahmias, Coates, and Kvaran 2007) ; l'idée est que les processus cérébraux qui causent nos actions sont totalement indépendants des processus neuronaux qui correspondent à notre pensée consciente ou à nos intentions ; la chaîne causale menant à l'action « court-circuite » le module cérébral de la conscience, de telle sorte que notre action en elle-même est purement « mécaniste ». Il est clair que le déterminisme n'implique ni la « condition de la guêpe » (l'absence de toute conscience), ni le « court-circuit » de notre module de conscience. Par ailleurs, il est aussi très consensuel de supposer que le libre arbitre est incompatible avec la condition de la guêpe et avec le modèle du court-circuit. Par conséquent, si nous avons une tendance à confondre le déterminisme avec l'un ou l'autre de ces modèles mécanistes de l'action, cette confusion pourrait être à l'origine de notre intuition incompatibiliste.

Passons à présent aux théories du croque-mitaine, c'est-à-dire aux théories d'après lesquelles le déterminisme a tendance à être confondu avec un *agent* intentionnel. La plus importante théorie de ce genre est la théorie de la confusion avec un agent manipulateur, démon, scientifique, ou martien. Nous l'aborderons en second. Une autre théorie du croque-mitaine est celle qui présente le déterminisme comme une sorte de « gendarme » qui nous donnerait des injonctions ou des ordres à agir. On peut ranger dans cette catégorie la célèbre théorie de Moritz Schlick (Schlick 1939), qui prétend que le « pseudo-problème du libre arbitre » repose sur une confusion entre « lois prescriptives » et « lois descriptives ». L'idée est la suivante : une « loi prescriptive » est une « règle par laquelle l'état prescrit une certaine conduite à ses citoyens. » Certaines de ces lois prescriptives sont « nécessaires » dans le sens où « l'état contraint ses citoyens » à les respecter, par une « inéluctable contrainte ». Par

ailleurs, une « loi descriptive » est une formule qui donne « simplement une description de ce qui se passe. » Si les lois descriptives sont parfois dites « nécessaires », ce n'est donc pas dans le sens de la contrainte mais dans le sens de la généralité : « elle vaut dans tous les cas auxquels elle s'applique ». Notre intuition incompatibiliste viendrait donc de ce que nous avons tendance à confondre les lois (descriptives) nécessaires (c'est-à-dire les descriptions universellement valables) avec des lois prescriptives nécessaires (c'est-à-dire des règles contraignantes), comme si la nature nous donnait elle-même des prescriptions. Dans la mesure où une contrainte inéluctable est incompatible avec le libre arbitre, il est clair qu'une simple confusion entre lois universelles et contrainte inéluctable générerait une intuition incompatibiliste. On peut remarquer un point étonnant de la théorie de Schlick : une loi prescriptive, en tant que telle, n'est jamais qu'une *norme*, c'est-à-dire quelque chose qui n'implique rien quant aux faits et peut donc toujours (peut-être difficilement, mais toujours) n'être pas respectée. Autrement dit, une loi prescriptive, même protégée par un attirail policier et répressif très fort, n'est jamais *absolument* contraignante. A l'inverse, certains *faits* peuvent être absolument contraignants, dans le sens où ne nous pouvons rien pour les changer, et cela semble bien être le cas des lois de la nature. La théorie de Schlick revient donc à dire que nous prenons les lois de la nature pour contraignantes parce que nous les confondons avec d'autres lois... qui sont *moins* contraignantes. Nous prenons les lois de la nature pour incompatibles avec la liberté, parce que nous les confondons avec des lois prescriptives... qui n'excluent pas la liberté mais la supposent. De fait, cette théorie de la confusion entre lois prescriptives et lois descriptives, malgré son autorité historique, n'est pas reprise à ma connaissance par les auteurs contemporains.

Les théories du « croque-mitaine » défendues aujourd'hui sont toutes, d'une manière ou d'une autre, des théories de la « manipulation » – que l'agent manipulateur soit un génie, un scientifique ou un martien. Daniel Dennett (1984a, 7–10) a trouvé dans la littérature du libre arbitre un certain nombre de croque-mitaine : « l'infâme neurochirurgien » (Nefarious Neurosurgeon) qui manipule nos décisions au moyen d'un dispositif qu'il a implanté dans notre cerveau, « l'hideux hypnotiste » (Hideous Hypnotist) qui nous a « programmés » sous hypnose à faire telle ou telle action, « l'irrésistible marionnettiste » (Peremptory Puppeteer) qui tire les ficelles de nos mouvements, « l'enfant cosmique dont nous sommes les poupées », ou encore le « voyant malveillant » (Malevolent Mindreader) qui peut manipuler nos actions non pas directement mais parce qu'il prévoit à la perfection toutes nos réactions. Toutes ces métaphores sont utilisées, d'après Dennett, comme des « pompes à intuitions », c'est-à-dire :

des expériences de pensée [qui] ne sont *pas* supposées suivre des arguments rigoureux, en prouvant des conclusions à partir de prémisses. Leur but est plutôt de susciter chez le lecteur un ensemble de réflexions de l'imagination, qui donneront en fin de compte non pas une conclusion formelle mais un diktat de « l'intuition ». (1984a, 12)

Pour Dennett, les intuitions suscitées par les histoires de croque-mitaine sont clairement illégitimes. Il est évident qu'en comparant le déterminisme à un infâme neurochirurgien, on donne au premier tout le caractère abominable du second... mais si l'on est attentif, le caractère abominable du neurochirurgien lui vient de caractéristiques qui le *distinguent* du déterminisme et non de caractéristiques qui le rendent comparables. En effet, être manipulé par un agent qui veut nous mener à ses fins (et, particulièrement, à ses fins *criminelles*, si l'on pense au scénario de Frankfurt), c'est clairement un sort abominable ; mais le déterminisme causal ne nous met évidemment pas dans cette situation. Le passé et les lois de la nature ne sont pas des « agents » manipulateurs ; ils n'ont aucune intention, aucune fin, certainement pas des fins criminelles. Il n'y a donc pas lieu de dire, dans un scénario déterministe, que nous agissons selon ce que « veut » la nature plutôt que selon ce que *nous* voulons, car la nature ne

veut absolument rien. Dans un scénario déterministe, je reste bel et bien le *seul agent* de mon action. Pour Dennett, ces histoires de croque-mitaines que l'on trouve un peu partout dans la littérature du libre arbitre, sont responsables en bonne partie de nos intuitions incompatibilistes. Compte tenu de l'ancienneté de ce genre d'histoires (Héraclite parle du destin comme d'un enfant qui joue au tric-trac, les stoïciens nous décrivent comme un chien attaché à un chariot, qui suit volontairement ou se laisse traîner), on peut penser qu'il y a en nous une tendance *naturelle* à décrire le déterminisme sous la forme de telles métaphores, et donc une tendance naturelle à attribuer au déterminisme les dangers et méfaits de ces histoires métaphoriques. Si tel est le cas, alors on dispose d'une théorie montrant que notre intuition incompatibiliste originelle repose sur une simple illusion.

Que penser de ces théories du grand méchant loup ? Le conflit intuitif commun entre déterminisme et libre arbitre vient-il simplement du fait que nous confondons le libre arbitre avec divers scénarios, terrifiants en eux-mêmes mais dont le caractère terrifiant vient précisément de ce qui les *distingue* du déterminisme causal ? Je ne reviendrai pas en détail sur chacun des types de scénarios évoqués ci-dessus, notamment parce que nous avons déjà eu l'occasion de voir les limites de certains d'entre eux. En particulier, nous avons vu que la théorie de Searle (confusion entre déterminisme par des lois physiques en général et déterminisme par des lois psychologiques) avait la conséquence étrange de présenter comme « préférable » des causes déterminantes extra-psychologiques (et donc court-circuitante) à des causes psychologiques. Nous avons vu également que la théorie de Schlick (confusion entre lois descriptives et lois prescriptives) revenait à dire que les lois de la nature sont conçues comme contraignantes parce qu'on les confond avec des lois... qui (en tant que normes) ne sont *pas* factuellement contraignantes. Ces deux théories sont peu défendues aujourd'hui. Il vaut donc mieux se concentrer sur les théories qui attribuent notre intuition incompatibiliste à une confusion entre déterminisme et fatalisme ou mécanisme, ou sur une illégitime comparaison avec différents scénarios de manipulation.

Concernant la confusion entre déterminisme et fatalisme ou mécanisme, certains résultats récents en philosophie expérimentale seront particulièrement pertinents. En effet, dans une série d'études, en particulier (Nahmias, Coates, and Kvaran 2007; Nahmias and Murray 2011), Eddy Nahmias s'est attaché à montrer que nous avons une tendance naturelle à confondre le déterminisme avec d'autres théories plus radicales (en particulier le modèle de court-circuit), et que cette confusion *expliquait* nos intuitions incompatibilistes. Pour montrer cela, (Nahmias and Murray 2011) ont cherché à évaluer deux choses, d'une part notre tendance à mal comprendre le déterminisme, et d'autre part une éventuelle *corrélation* entre de telles mécompréhensions et l'intuition incompatibiliste telle qu'on la trouve dans les expériences de X-Phi. Nahmias et Murray ont réutilisé pour cette expérience les scénarios que nous avons décrits au chapitre 1 (p.66), notamment la distinction entre Univers A (déterministe) et Univers B (indéterministe aux endroits des décisions humaines) et la distinction entre scénarios concrets et scénarios abstraits. Après avoir présenté ces scénarios, Nahmias et Murray ont posé deux séries de questions, une série classique concernant la responsabilité morale, le libre arbitre, ou la légitimité du blâme dans ces univers, et une deuxième série de question destinée à tester si les sujets interprétaient la situation considérée comme une situation de court-circuit. Voici la série de question (Nahmias and Murray 2011, sec. III) :

Décisions : Dans l'univers [A/B], les décisions d'une personne n'ont aucun effet sur ce qu'elle est finalement causalement amenée à faire.

Volontés : Dans l'univers [A/B], ce qu'une personne veut n'a aucun effet sur ce qu'elle est finalement causalement amenée à faire.

Croyances : Dans l'univers [A/B], ce qu'une personne croit n'a aucun effet sur ce qu'elle est finalement causalement amenée à faire.

Pas de contrôle : Dans l'univers [A/B], une personne n'a aucun contrôle sur ce qu'elle fait.

Passé différent : Dans l'univers A, tout ce qui arrive *doit* arriver, même si ce qui était arrivé dans le passé était différent.

Il est clair que rien dans le déterminisme n'empêche que les décisions, les volontés, et les croyances des agents aient un effet causal. Ce qui exclut un tel effet, c'est le modèle du court-circuit, pas le modèle déterministe. Par conséquent, une réponse positive à l'une de ces questions manifesterait une confusion entre déterminisme et court-circuit. De même, il est clair que dans un scénario déterministe, un passé différent entraîne un futur différent ; c'est dans un modèle *fataliste* que le futur est fixé quel que soit le passé. Donc une réponse positive à la dernière question manifesterait une confusion entre déterminisme et fatalisme.³⁴⁰

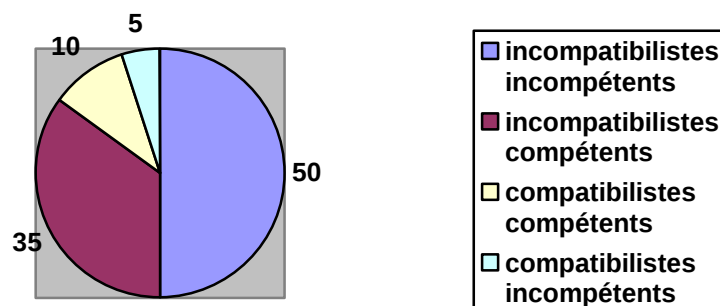
Nahmias et Murray ont mis en évidence le résultat qu'ils souhaitaient trouver, à savoir une forte corrélation entre illusion de court-circuit et intuition incompatibiliste. Plus précisément, si l'on se concentre sur le scénario « abstrait » de Nichols et Knobe, celui dont nous avons vu qu'il révélait la plus forte intuition incompatibiliste, Nahmias et Murray ont retrouvé bien sûr cette intuition incompatibiliste (73% de réponses négatives à la première série de questions, concernant la responsabilité morale, le libre arbitre et le blâme dans l'Univers A), mais ils ont également trouvé un taux très fort de confusion du court-circuit (66% de réponses *positives* – et donc erronées – à la deuxième série de questions). Ce résultat les amène à conclure que sur les 73% d'intuitions « incompatibilistes », il y en a en fait la quasi-totalité (66%) qui sont de *faux* incompatibilistes, c'est-à-dire des personnes qui croient le déterminisme incompatible avec le libre arbitre simplement parce qu'ils comprennent mal le déterminisme. Ainsi, la forte intuition incompatibiliste que nous observons en philosophie expérimentale serait simplement fondée sur une confusion du court-circuit.

Ce résultat de Nahmias et Murray est certainement l'argument le plus fort dont on dispose aujourd'hui en faveur d'une théorie de l'illusion. L'argument n'est pourtant pas concluant pour deux raisons. Premièrement, la conclusion de Nahmias et Murray semble reposer sur la confusion entre deux probabilités conditionnelles. Appelons « compétents » les sujets qui comprennent bien le déterminisme, et « incompetents » ceux qui le confondent avec le modèle du court-circuit. Ce que Nahmias et Murray observent, c'est qu'une large partie des incompatibilistes sont des « incompetents », autrement dit, qu'on a de fortes chances d'être un « incompetent » lorsqu'on est un incompatibiliste.³⁴¹ La probabilité $P(\text{incompétent} \mid \text{incompatibiliste})$ est élevée (0,91). Mais les intuitions des incompetents, dans la mesure où ils ne comprennent pas bien le déterministe, n'est en fait *ni* une intuition incompatibiliste, *ni* une intuition compatibiliste ; c'est une intuition « hors-sujet ». Pour mettre en évidence nos intuitions compatibilistes ou incompatibilistes, il faut donc *mettre de côté* les incompetents.

³⁴⁰ La quatrième question est un peu plus litigieuse car la notion de contrôle n'est pas absolument claire ; mais dans la mesure où un modèle déterministe permet que nos volontés aient un effet causal sur nos actions, on peut dire qu'il permet sans doute *une certaine forme* de contrôle ; dans ce cas, quelqu'un qui refuserait à l'agent *toute* forme de contrôle dans le scénario déterministe l'aurait sans doute confondu avec un scénario plus radical, fataliste ou de court-circuit.

³⁴¹ Ils montrent également que la proportion d'incompétents parmi les *compatibilistes* est beaucoup plus faible.

La probabilité qui nous intéresserait vraiment pour connaître si nous avons une intuition incompatibiliste ou non, c'est la probabilité d'être incompatibiliste *lorsqu'on est compétent*. Autrement dit $P(\text{incompatibiliste} \mid \text{compétent})$ et non pas $P(\text{incompétent} \mid \text{incompatibiliste})$. Or ces deux probabilités sont indépendantes l'une de l'autre : il se pourrait très bien que la plupart des incompatibilistes soient incompétents, et que pourtant, parmi les compétents, il reste une majorité d'incompatibilistes. Ce serait le cas, par exemple, dans le schéma suivant :



Dans une telle situation, bien que la majorité des incompatibilistes soient incompétents ($50/85=59\%$), on ne pourrait pas conclure que l'intuition majoritairement incompatibiliste (85%) « est due à » la mécompréhension du déterminisme, puisqu'on retrouverait une forte majorité incompatibiliste en se concentrant sur les sujets compétents ($35/45=77\%$). La vraie question, pour connaître nos intuitions communes sur le compatibilisme ou l'incompatibilisme, consiste donc bien à se demander si la majorité incompatibiliste existe toujours lorsqu'on se concentre sur les sujets compétents. C'est pour cette raison que (Deery, Bedke, and Nichols 2013, 133–134) ont mis en place dans leur dispositif expérimental une étape « d'entraînement au déterminisme » qui a pour but de s'assurer que les sujets testés comprennent bien la notion de déterminisme et ne la confondent ni avec le fatalisme ni avec le court-circuit.³⁴² Le résultat de Deery et al. est très clair : lorsqu'on se concentre sur les sujets compétents, sur les sujets qui ne *font pas* la confusion entre déterminisme et court-circuit, on retrouve tout de même le fort taux d'intuitions incompatibilistes. La confusion entre déterminisme et court-circuit n'est donc pas « ce qui explique » la présence d'une forte intuition incompatibiliste.³⁴³

Une deuxième considération peut être avancée pour répondre à Nahmias et Murray. Malgré notre première réponse (le fait que « parmi les compétents », il reste une majorité

³⁴² Le dispositif est le suivant : après avoir lu une description de ce que signifie le déterminisme (appelé en l'occurrence « complétude causale » pour éviter les éventuelles incompréhensions liées au mot « déterminisme »), les sujets doivent répondre à une question dont la réponse manifeste une éventuelle confusion avec le court-circuit. Ceux qui répondent *mal* à cette question reçoivent une explication de leur erreur, et une deuxième chance. S'ils se trompent une deuxième fois, ils sont tout simplement éliminés de l'étude et ne reçoivent même pas les questions sur le libre arbitre. Les sujets qui réussissent l'entraînement (au premier ou au deuxième coup), sont jugés compétents pour pouvoir exprimer une intuition vraiment compatibiliste ou vraiment incompatibiliste.

³⁴³ On pourrait se demander si la meilleure manière « d'éliminer les incompétents » n'est pas tout simplement de restreindre son échantillon... aux philosophes professionnels. Après tout, ce sont eux qui sont censé maîtriser le mieux les concepts en jeu dans la discussion. Le problème, c'est que les philosophes professionnels ne sont pas seulement « compétents », il sont aussi potentiellement biaisés dans leurs intuitions par d'autres facteurs que leur simple compétence (considérations dans d'autres parties de leur système philosophique, le poids d'une habitude de penser philosophique, etc.) Ce qui nous intéresse, c'est d'observer l'effet de la *seule* compétence conceptuelle sur l'intuition de l'incompatibilisme.

d'incompatibilistes), il n'en demeure pas moins que Nahmias et Murray ont mis en évidence une forte proportion de confusion dans la population générale. Par conséquent, il semble toujours pertinent de dire que *pour la population générale* au moins, une grande partie de l'incompatibiliste est fondée sur une confusion entre déterminisme et court-circuit. Mais cela n'est pas évident non plus : une corrélation statistique n'indique pas l'ordre causal ou l'ordre d'explication. Nahmias et Murray expliquent la corrélation entre incompetents et incompatibilistes en disant que c'est parce qu'on est incompetent qu'on est incompatibiliste. Cet ordre d'explication est clairement possible, mais ce n'est pas le seul : il se pourrait en effet que certains sujets soient au contraire « incompetents » parce qu'ils sont incompatibilistes. Qu'est-ce que cela voudrait dire ? Ainsi qu'on vient de l'observer, il y a une forte tendance à être *vraiment* incompatibiliste (à être incompatibiliste à propos du concept bien compris de déterminisme). Quelqu'un qui est un *vrai* incompatibiliste jugera donc que, dans l'univers déterministe A, nous n'avons ni libre arbitre, ni responsabilité morale, ni fondement au blâme etc. Toutes ces conséquences sont généralement tenues pour moralement « déprimantes ». Lorsqu'on pose ensuite à ce sujet la deuxième série de questions (dans cet univers, nos volontés ou décisions ont-elles un effet causal ? etc.), il est très possible que les conséquences déprimantes de la première réponse aient un effet de « contagion » et que le sujet réponde pour cela négativement. Autrement dit, le sujet répondrait qu'on est comme dans un univers de court-circuit non pas parce qu'il a mal compris la description de l'univers déterministe, mais parce que l'univers déterministe et l'univers de court-circuit sont équivalents du point de vue des conséquences morales qu'on vient de lui faire envisager. Des effets de contagion bien plus spectaculaires ont été observés en philosophie expérimentale. Par exemple, dans un scénario célèbre de Joshua Knobe, un chef d'entreprise choisit de mettre en œuvre un programme dont on lui a annoncé qu'il serait positif pour l'environnement, mais le chef d'entreprise ne se soucie pas le moins du monde de ces conséquences positives collatérales ; son choix est entièrement guidé par son intention de faire progresser son entreprise. Dans un second scénario comparatif, le même chef d'entreprise choisit un programme dont on lui a fait savoir qu'il *nuirait* à l'environnement ; mais là encore, le chef d'entreprise ne se soucie que de son entreprise. Dans l'expérience d'origine, ces scénarios servaient à tester les intuitions communes sur l'action *volontaire* : une majorité de gens répondaient que le chef d'entreprise n'avait pas volontairement aidé l'environnement, mais qu'il avait *nui* volontairement. L'asymétrie entre les deux réponses était l'élément intéressant, compte tenu de l'absence d'asymétrie dans l'attitude du chef d'entreprise. Dans certaines expériences plus tardives, Knobe a testé d'autres intuitions, et notamment l'intuition concernant le mot « savoir » : le chef d'entreprise *savait-il* qu'il aiderait/nuirait à l'environnement ? Et étonnamment, la même asymétrie est apparue dans les réponses : bien qu'il soit *stipulé* dans le scénario que le chef d'entreprise était au courant, un certain nombre de personnes ont eu tendance à dire que le chef d'entreprise *ne savait pas* qu'il aiderait l'environnement. Ceci s'explique par effet de contagion : lorsqu'on a à l'esprit l'enjeu moral de la question, le chef d'entreprise est *aussi peu louable* que s'il n'avait absolument rien su des effets sur l'environnement. Les deux situations sont équivalentes. De la même façon, si quelqu'un est un incompatibiliste véritable et qu'on attire d'abord son attention sur les conséquences moralement « déprimantes » de cet incompatibilisme, alors il aura des chances de ne pas *porter attention* à la différence entre déterminisme et court-circuit parce que cette différence n'est pas pertinente dans le contexte. Dans ce cas, la forte proportion d'intuitions de « confusion » dans la population générale ne s'expliquerait pas par le fait que de nombreuses personnes sont incompetentes et confondent effectivement déterminisme et court-circuit, mais par le fait qu'un grand nombre de personnes jugent non pertinente la distinction (qu'ils peuvent faire) entre déterminisme et court-circuit.

La leçon générale concernant la théorie de l'illusion par confusion avec le fatalisme et le court-circuit est donc la suivante : tout d'abord l'intuition incompatibiliste reste majoritaire parmi les sujets dont on s'assure qu'ils ne font aucune confusion, et d'autre part, si la confusion *semble* répandue dans la population générale, ce pourrait très bien être parce que la population générale est constituée (essentiellement) d'incompatibilistes *compétents* qui jugent *non pertinente* la distinction existante entre déterminisme et court-circuit.

Une distinction peut très bien être comprise et n'être pourtant pas pertinente pour le problème qui est en jeu. C'est également l'objection qu'on peut faire aux théories du croque-mitaine : il est tout simplement *évident* que le déterminisme ne fait pas apparaître un *agent* manipulateur, et qu'il y a donc une « distinction » entre déterminisme et scénarios de manipulation, quels qu'il soient. C'est tellement évident qu'on peut fortement douter que quiconque ait jamais « confondu » le déterminisme avec un scénario de manipulation. Mais si les scénarios de manipulation sont si présents dans la littérature philosophique, c'est peut-être pour une autre raison : non pas parce que nous aurions tendance à « oublier » ou à « ne pas remarquer » la différence avec le déterminisme, mais plutôt au contraire parce que nous aurions tendance à « ne pas oublier » ou à « bien remarquer » que la différence existante n'est pas pertinente pour la question du libre arbitre et de la responsabilité morale. C'est là le cœur des arguments incompatibilistes « de la Manipulation », en particulier de « l'argument des quatre cas » de Derk Pereboom. Nous retrouvons ici la même situation dialectique que pour les théories de l'erreur précédentes : là où le compatibiliste voit une confusion entre deux étapes logiques distinctes, on peut voir au contraire l'intuition d'une *inférence* qui constitue un argument très fort. Mais voyons l'argument.

L'argument de la Manipulation présenté par Derk Pereboom (2001, 112 sqq) utilise quatre « cas », quatre situations d'action déterministe, entre lesquels Pereboom défend qu'on ne peut trouver *aucune* différence pertinente pour la question du libre arbitre et de la responsabilité morale. Le premier cas est un scénario de manipulation typique, dans lequel Pereboom prend soin néanmoins de satisfaire toutes les conditions *positives* qu'un compatibiliste peut donner pour définir la liberté :

Cas 1. Le professeur Plum a été créé par des neuroscientifiques, qui peuvent le manipuler directement grâce à une technologie radio, mais il est à part ça aussi semblable que possible à un être humain normal. Supposons que ces neuroscientifiques le manipulent « localement » à mener un processus de raisonnement qui suscite et modifie ses désirs – produisant ainsi chacun de ses états d'instant en instant. Les neuroscientifiques le manipulent, notamment, en poussant une série de boutons juste avant qu'il commence à raisonner à sa situation, et causent ainsi le fait que son processus de raisonnement est rationnellement égoïste. Plum n'est pas contraint à agir dans le sens où il n'agit pas par un désir irrésistible – les neuroscientifiques ne lui fournissent pas un désir irrésistible – et il ne pense ni n'agit contrairement à son caractère, puisqu'il est souvent manipulé à être rationnellement égoïste. Son désir effectif de premier ordre de tuer Mme White est conforme à ses désirs de second ordre. Le processus de raisonnement de Plum a les différents aspects de la sensibilité aux raisons. (ibid., p.112-113)

On pourrait dire que, dans le cas 1, le professeur Plum est « idéalement manipulé » : il est manipulé d'une manière qui respecte entièrement les conditions compatibilistes positives de la liberté, et donc la *seule* raison pour laquelle un compatibiliste pourrait dire que Plum n'est pas libre, c'est à cause de la condition *négative* qu'être libre suppose de n'être pas manipulé. Mais cette condition purement négative est-elle vraiment une condition plausible pour la liberté ?

C'est ce que les cas suivants sont censés mettre en doute. Le deuxième cas est encore un scénario de manipulation, mais un scénario de manipulation *indirecte* :

Cas 2. Plum est semblable à un être humain ordinaire, à ceci près qu'il a été créé par des neuroscientifiques qui, bien qu'ils ne puissent pas le contrôler directement, l'ont programmé à peser ses raisons d'action de telle sorte qu'il soit souvent, mais pas exclusivement, rationnellement égoïste, et le résultat est que dans les circonstances où il se trouve actuellement, il est causalement déterminé à suivre un processus sensible aux raisons et à posséder les désirs de premier et de deuxième ordre qui résulteront dans le meurtre de Mme White. [Par ailleurs], il n'agit pas sous l'effet d'un désir irrépessible. (ibid. p. 113-114)

D'après Pereboom, la distinction qu'il y a entre les cas 1 et 2 – distinction bien réelle, à n'en pas douter – ne *peut pas* être pertinente pour le libre arbitre et la responsabilité morale. Si l'on juge que Plum n'est pas libre dans le premier scénario, il n'y a aucune bonne raison de ne pas porter le même jugement dans le second. En effet, dans les deux cas, il s'agit d'une manipulation par les mêmes agents (les neuroscientifiques), dans la même intention, et qui mène aux mêmes résultats. La différence tient seulement à la manière dont les manipulateurs interviennent, mais un compatibiliste ne devrait pas juger cette différence pertinente s'il juge la manipulation elle-même problématique.

Pourtant le cas 2 commence à devenir problématique pour le compatibiliste, car il est extrêmement proche d'un troisième cas de détermination, un cas de détermination par le milieu et l'éducation :

Cas 3. Plum est semblable à un être humain normal, à ceci près qu'il a été déterminé par les pratiques de dressage rigoureuses de sa famille et de sa communauté de telle sorte qu'il est souvent mais pas exclusivement rationnellement égoïste (exactement aussi égoïste que dans les cas 1 et 2). Son dressage a eu lieu à une époque trop précoce pour qu'il ait pu avoir la capacité d'empêcher ou de modifier les pratiques qui ont déterminé son caractère. Dans les circonstances présentes, Plum est donc causalement déterminé à suivre le processus sensible aux raisons, et à posséder les désirs de premier et de deuxième ordre, qui résulteront dans le meurtre de Mme White. [Par ailleurs], il n'agit pas sous l'effet d'un désir irrépessible. (ibid., p.114)

De nouveau, il est évident qu'il y a une différence entre les cas 2 et 3 : dans le cas 2, le caractère de Plum qui l'amène déterministiquement à tuer a été formé (il y a longtemps) par des neuroscientifiques, tandis que dans le cas 3, ce même caractère qui va l'amener à perpétrer le même acte selon le même processus décisionnel a été formé (il y a longtemps aussi) par un « dressage » éducatif. S'il y a un facteur dans le cas 2 qui enlève la liberté et la responsabilité morale, ça ne peut pas être un facteur positif concernant le processus qui mène à l'action : tous ces facteurs positifs sont respectés en suivant les meilleures théories compatibilistes. Donc le seul facteur qui peut enlever la liberté et la responsabilité, c'est la présence d'un facteur déterminant dans le passé, et ce type de facteur est également présent dans le cas 3 : la différence entre les deux facteurs déterminants tient seulement à la *manière* dont la communauté neuroscientifique d'une part, et la communauté éducative d'autre part, ont produit dans le passé le même effet de détermination du caractère.

Enfin, le cas 3 se rapproche dangereusement du scénario déterministe normal, qui constitue le « cas 4 » de Pereboom :

Cas 4. Le déterminisme physicaliste est vrai, et Plum est un être humain normal, engendré et éduqué dans des circonstances normales, qui est souvent mais pas exclusivement rationnellement égoïste (exactement aussi égoïste que dans les cas 1 à 3). Le meurtre de White par Plum est le résultat du fait qu'il a suivi un processus de délibération sensible aux raisons, qu'il témoigne d'une bonne organisation de ses desirs de premier et de deuxième ordre, et il n'agit pas sous l'effet d'un désir irrépressible. (ibid., p.115)

Certes, les cas 3 et 4 présentent une différence évidente. Dans les deux cas, Plum est causalement déterminé à agir, et il l'est par son caractère, et son caractère a été formé avant l'action considérée, et ultimement par des facteurs qui ne dépendent pas de son contrôle. La seule différence tient aux facteurs qui ont déterminé son caractère : dans le cas 3, il s'agissait d'agents (la communauté qui l'a éduqué), alors que dans le cas 4, il peut s'agir au moins partiellement de facteurs qui ne sont pas des personnes ou des agents. Mais pourquoi cette seule différence serait-elle pertinente pour le libre arbitre ? Pour montrer que la différence agent/non-agent n'est pas pertinente, Pereboom imagine encore un autre scénario qui est identique au cas 1 à ceci près que l'instance manipulatrice ne serait pas un neuroscientifique mais une machine qui aurait été générée de manière spontanée : il est clair que cette machine priverait Plum de liberté tout autant qu'un manipulateur neuroscientifique ; le simple fait de n'être pas conscient ne change rien, intrinsèquement, à la question de la liberté et de la responsabilité tant que la contrainte opérée sur Plum reste la même. Par conséquent, le cas 4 ne présente aucune différence pertinente avec le cas 3.

En réunissant les différentes étapes de cet argument, on arrive à la conclusion que les cas extrêmes, le cas 1 et le cas 4, ne présentent entre eux aucune différence pertinente pour la liberté et la responsabilité morale. Dans les quatre cas, la menace éventuelle sur le libre arbitre est exactement la même : c'est que l'agent Plum n'est pas l'origine ultime de son acte ; dans certains cas, l'origine ultime de l'acte est un autre agent, dans d'autres cas, l'origine ultime de l'acte est un processus non agentif et non intentionnel, mais cette distinction n'est pas pertinente. Si il y a une menace, c'est celle de n'être pas la source ultime de l'acte. Autrement dit, il n'y a que deux possibilités : soit Plum est libre dans tous les cas, même le cas 1, soit il n'est libre dans aucun des cas, pas même le cas 4. La première réponse est ce qu'on appelle la « ligne dure », car elle consiste à admettre qu'on peut être libre et moralement responsable lorsqu'on est totalement manipulé (manipulé localement pour chaque acte), à condition que cette manipulation soit suffisamment ingénieuse. La seconde branche de l'alternative est l'incompatibilisme, puisqu'elle revient à nier la liberté et la responsabilité dans un scénario déterministe normal ; c'est la conclusion visée par l'argument de Pereboom à cause du caractère difficilement acceptable de la « ligne dure ».³⁴⁴

L'argument de la Manipulation, en particulier dans cette version des « quatre cas », est considéré aujourd'hui comme l'un des meilleurs arguments incompatibilistes, et fait l'objet d'une discussion abondante (Voir notamment Vihvelin 2011, sec. 3.2). Comme pour l'argument de la conséquence un peu plus haut, mon intention n'est pas ici d'entrer dans le détail des discussions pour soutenir que l'argument est ultimement un succès. Mon intention ici est seulement de faire observer que si notre intuition incompatibiliste repose sur une assimilation entre déterminisme et situations de manipulation, il est tout sauf clair que cette intuition soit une « simple confusion » ; au contraire, il y a toutes les raisons de penser que cette intuition suive implicitement les lignes d'un des arguments incompatibilistes les plus résistants. Les « croque-mitaine » sont donc peut-être des « pompes à intuition » si l'on entend par là positivement un moyen de faire appel à une intuition que nous pouvons

³⁴⁴ La réponse dite « ligne souple » consiste à soutenir que les scénarios ne sont pas équivalents du point de vue du libre arbitre, que l'une des différences au moins est pertinente.

naturellement avoir (en l'occurrence, l'intuition d'une équivalence entre déterminisme et situations de manipulation), mais ce ne sont sûrement pas des invitations à commettre une simple confusion.

Pour conclure de manière plus générale sur les théories de l'illusion « du grand méchant loup », il faut donc distinguer deux types de cas. Dans le cas de la confusion déterminisme / mécanisme, nous avons des raisons empiriques (en philosophie expérimentale) de penser que notre intuition incompatibiliste ne *repose pas* (factuellement) sur une telle assimilation. Dans le cas de l'assimilation déterminisme / manipulation, il est fort possible que notre intuition incompatibiliste repose en effet sur une telle assimilation, mais loin d'être une simple confusion, due à un manque d'attention ou de précision, cette assimilation constitue au contraire un des plus forts arguments en faveur de l'incompatibilisme lorsqu'on l'étudie avec toute l'attention et toute la précision requises.

III.4. Bilan des théories de l'illusion

Dans les sections précédentes, j'ai étudié successivement trois grandes théories qui ont en commun d'attribuer notre intuition incompatibiliste une origine illusoire. Dans le premier cas, notre intuition viendrait d'une confusion entre « l'acte libre » et la « volonté libre » (ce dernier concept représentant une erreur de catégorie). Dans le second cas, notre intuition reposerait sur une confusion modale entre nécessité de la conséquence (la conséquence passé $\rightarrow$ acte) et nécessité du conséquent (donc la nécessité de l'acte). Dans le troisième cas, notre intuition reposerait principalement sur une confusion ou une fausse assimilation entre déterminisme et manipulation. Ma conclusion, dans chacun de ces trois cas, a consisté à montrer que la « confusion » alléguée pouvait au contraire manifester l'intuition d'une inférence tout à fait légitime, ou en tout cas tout à fait défendable sous la forme d'un argument incompatibiliste très fort. La conclusion générale est donc la suivante : nous ne disposons pas d'une théorie de l'illusion qui soit initialement ou indépendamment plausible. C'est-à-dire qu'à moins d'avoir des raisons positives d'être compatibiliste, nous n'avons aucune raison de supposer que notre intuition incompatibiliste reposerait sur une illusion ; au contraire, nous avons toutes les raisons d'en rester à la théorie la plus simple sur la question, à savoir que notre intuition incompatibiliste provient probablement d'une incompatibilité réelle entre les concepts de libre arbitre et de déterminisme. C'est là la leçon principale à tirer de cette discussion des théories de l'erreur ; il me semble qu'on peut également tirer deux autres leçons, l'une qui vient de la *diversité* des théories de l'illusion envisagées, l'autre qui vient du *contenu* positif du concept de libre arbitre déjà abordé dans ces sections.

Remarquons tout d'abord la *diversité* des théories de l'illusion envisagées : certaines situent l'origine du problème dans une erreur sur le concept de libre arbitre (l'erreur de catégorie de la volonté libre), d'autres dans une erreur sur le concept de déterminisme (les théories du grand méchant loup) ; certaines considèrent qu'un déterminisme par lois psychologiques serait menaçant (Searle), d'autres qu'il est au contraire le meilleur type de déterminisme par rapport à un déterminisme cour-circuitant (Nahmias) ; certaines admettent qu'il faut un certain contrôle des volitions (Vihvelin), d'autres situent l'illusion précisément dans le réquisit de contrôle des volitions (conditionnalisme classique). En entrant d'avantage dans les détails, on aurait pu montrer encore d'autres exemples de théories de l'illusion qui prennent l'exact contre-pied les unes des autres. Mon intention n'est pas ici de suggérer que les théoriciens de l'illusion se décrédibilisent en défendant tout et son contraire : on pourrait aisément trouver, pour n'importe quelle théorie philosophique, les mêmes dissensions de détails entre ses différents partisans. Mais il me semble que ces divergences permettent

néanmoins de supposer que les théoriciens de l'illusion le sont généralement « par ajustement », et non pas de manière « indépendamment motivée ». Je réutilise ici la distinction faite plus haut entre deux manières d'être compatibiliste (et donc deux manières d'adopter une théorie de l'illusion incompatibiliste) : le compatibilisme indépendamment motivé (qui prétend que le compatibilisme est la meilleure théorie de notre concept lorsqu'on examine ce concept en lui-même, *indépendamment* de la question de savoir si ce concept est satisfait dans la réalité), et le compatibilisme d'ajustement (qui prétend que, même si notre concept semble initialement être incompatibiliste, il faut le supposer compatibiliste lorsqu'on s'aperçoit qu'un concept incompatibiliste serait insatisfaisable ou inconnaissable). Si les auteurs compatibilistes l'étaient généralement de manière indépendamment motivée, cela voudrait dire qu'il y a une *bonne* raison indépendante d'accepter une théorie de l'illusion, et on devrait s'attendre à ce que cette bonne raison soit, en général, la même pour tout le monde. Inversement, si les auteurs compatibilistes le sont par ajustement, alors leur conviction qu'il y a une illusion dans notre intuition incompatibiliste sera initialement une conviction générale : il doit bien y avoir une erreur, mais on ne sait pas encore laquelle. Et par conséquent, on devra s'attendre à ce que leurs théories de l'illusion aient un caractère spéculatif ou hypothétique, et présentent des divergences importantes entre elles. Par conséquent, le fait que les théories de l'illusion soient largement divergentes et parfois contradictoires semble indiquer que les auteurs compatibilistes sont souvent plutôt des compatibilistes d'ajustement (cherchant une erreur qu'il doit bien y avoir pour sauver le libre arbitre), que des compatibilistes indépendamment motivés (acceptant le compatibilisme parce que l'erreur incompatibiliste leur aurait sauté aux yeux avec évidence).

Deuxième remarque que l'on peut faire à l'issue de cette partie sur les théories de l'illusion : chacune de ces théories avait à la fois un versant négatif et un versant positif. Le versant négatif consistait à dire ce que le libre arbitre *n'est pas* et avec quoi on pourrait le confondre – par exemple, le libre arbitre n'est pas le contrôle des volitions d'après le conditionnalisme classique, bien qu'on puisse le confondre avec cette condition. Mais pour pouvoir rendre compte de la confusion, le théoricien de l'illusion devait aussi dire ce que le libre arbitre *est* positivement, ou plus précisément ce qui, dans ce concept, peut donner lieu à confusion. Autrement dit, même si notre intention dans cette partie n'était pas de donner une théorie du contenu positif du concept de libre arbitre, nous avons déjà eu quelques éléments de réponse. Dans la partie suivante, nous envisagerons cette question du contenu de manière directe, notamment grâce à la philosophie expérimentale et à la phénoménologie de notre expérience agentive. Ces sources nous apporteront des indications *directes* sur le contenu que peut avoir notre concept de libre arbitre. Mais il peut être bon, auparavant, de résumer les éléments de contenu que nous avons déjà eu l'occasion d'aborder à titre hypothétique dans cette partie.

La théorie de l'erreur de catégorie accepte comme condition nécessaire au libre arbitre le fait d'avoir un contrôle alternatif de nos actes. Mais par ailleurs, elle refuse un contrôle alternatif de nos *volitions* en prétendant que cette condition de contrôle des volitions n'est admise que par confusion entre liberté d'action et liberté de volonté. Sur la base de l'argument de Lehrer, nous avons montré que le passage d'une condition de contrôle de l'action à une condition de contrôle des volitions était plus probablement une inférence légitime qu'une pure et simple confusion. Le résultat positif de cette discussion est donc le suivant : le libre arbitre suppose certainement un contrôle alternatif de nos actions, et si l'on accepte un contrôle alternatif de nos actions, on est sans doute amené à accepter également (du fait de l'argument de Lehrer) un contrôle alternatif de nos volitions.

La théorie de la confusion modale accepte comme condition nécessaire au libre arbitre le fait que nos actes ne soient pas (inconditionnellement) *nécessaires*. Elle refuse en revanche que nos actes doivent être non-nécessités (par le passé et les lois de la nature). Pour cette

théorie, le fait de passer d'une condition de non-nécessité à une condition de non-nécessitation repose sur une simple confusion. Mais nous avons que ce passage était plus probablement une inférence tout à fait défendable, reposant sur la nécessité du passé (et des lois) et la règle modale (β). Le résultat positif est donc le suivant : le libre arbitre suppose certainement que nos actions soient non-nécessaires, et si l'on accepte cette condition, on est très probablement amené à accepter que nos actions ne doivent pas non plus être nécessitées par des conditions elles-mêmes nécessaires, à savoir les lois et le passé.

Enfin, la théorie du grand méchant loup, dans sa version manipulation, accepte comme condition nécessaire au libre arbitre que nous ne devons pas être manipulés pour être libres. Elle refuse en revanche que nous devions être non-déterminés par le passé et attribue cette condition à une simple confusion avec la situation de manipulation. Mais nous avons vu, avec l'argument de Pereboom, que le passage de la situation déterministe à la situation de manipulation avait de bonnes chances de reposer sur une intuition légitime d'absence de différence pertinente. Par conséquent, le résultat positif est le suivant : le libre arbitre requiert très certainement de n'être pas manipulé par des agents, si lointains soient-ils, et si l'on admet cette contrainte, on a toutes les raisons d'admettre une contrainte plus générale de « source ultime » : la source ultime de notre action doit se trouver en nous-mêmes, et pas dans un processus passé, qu'il s'agisse d'un processus intentionnel et volontaire (un manipulateur) ou d'un processus inconscient et aveugle.

On peut résumer toute cette discussion dans le tableau suivant :

	Erreur de catégorie	Confusion NCNC	Grand Méchant Loup
condition nécessaire concédée au LA	contrôle alternatif de l' <i>action</i>	non nécessité de l'action	non manipulation
confusion alléguée	confusion <i>action</i> / <i>volition</i>	confusion <i>nécessité</i> / <i>nécessitation</i>	confusion <i>manipulation</i> / <i>déterminisme</i>
argument sous-jacent probable	argument de Lehrer	argument de la conséquence	argument des quatre cas
condition nécessaire probable au LA	contrôle alternatif des <i>volitions</i>	non nécessitation par le passé (<i>leeway</i>)	<i>source</i> ultime de nos actions

Ces trois théories permettent donc de faire ressortir clairement trois conditions qui ont une très grande importance dans la littérature sur le libre arbitre : la condition de contrôle des volitions, la condition des possibilités alternatives laissées ouvertes par le passé et les lois de la nature (ou condition du *leeway*), et enfin la condition de source ultime (*sourcehood*). En droit (et en fait), ces conditions peuvent être séparées les unes des autres. En particulier, un certain nombre d'auteurs, comme Derk Pereboom, rejettent la condition des alternatives (*leeway*) et fondent leur incompatibilisme uniquement sur la condition de source ultime (qui, par l'argument des quatre cas, est *suffisante* pour établir l'incompatibilité du libre arbitre et du déterminisme). C'est la stratégie qu'on appelle « incompatibilisme de la source ». J'exposerai, un peu plus loin dans ce chapitre (sec. IV.2.c.) mes raisons de conserver la condition des alternatives, et pas seulement celle de la source. Mais pour l'instant, je remarquerai simplement que, à première vue, ces trois conditions ont une certaine probabilité d'être des conditions nécessaires du libre arbitre, et qu'à ce titre il peut être intéressant d'essayer de les *combiner* pour obtenir un premier « portrait robot » de ce que pourrait être le libre arbitre. La manière la plus naturelle de combiner ces trois conditions est la suivante : pour être libre, il faut que nos actions dépendent de nos volitions, mais il faut que nos volitions elles-mêmes soient contrôlées par nous dans le double sens où nous devons en être la source ultime et où nous avons la possibilité, au moment de la volition, de déterminer cette volition ou la volition contraire, les deux alternatives étant compatibles avec le passé et les lois de la nature. Ce

portrait du libre arbitre est, au vocabulaire près, celui donné par Robert Kane (1996, 74 sqq). Pour qu'une action soit libre, d'après Kane, nous devons en être « ultimement responsables » dans le sens où l'origine de notre action remonte ultimement à une « volition auto-formante » (self-forming willing). Et dire qu'une volition est auto-formante, c'est dire qu'elle n'est pas la conséquence logique de ses antécédents (croyances et désirs) et des lois de la nature, mais que tout était compatible, au moment où elle a été formée, avec la formation d'une autre volition.

Notre concept de libre arbitre est-il *réellement* le concept d'une volition indéterminée et auto-formée comme source ultime de nos actions ? Nous ne pouvons pas l'affirmer de manière directe pour le moment, mais les concessions mêmes des théoriciens de l'erreur, et le peu de plausibilité de leurs théories de l'erreur, peuvent déjà nous amener à le penser.

IV. Analyse phénoménologique de notre expérience du libre arbitre

Comment étudier de manière directe le contenu de notre expérience du libre arbitre ? Comment déterminer positivement le contenu de cette expérience qui, en nous, génère naturellement un conflit intuitif avec l'idée de déterminisme ? Pour pouvoir faire cela, il y a une étape préliminaire fondamentale : il faut d'abord *identifier* ou *localiser* cette expérience. Au chapitre 1 (p.74 sqq), j'ai distingué trois théories possibles pour identifier ou localiser l'expérience du libre arbitre. En bref, ces trois modèles étaient les suivants :

Modèle phénoménal :

L'expérience qui génère un conflit avec l'idée de déterminisme est l'expérience phénoménale agentive, i.e. l'expérience que nous faisons de nous-mêmes comme agents au moment même où nous sommes en train d'agir.

Modèle moral :

L'expérience qui génère un conflit avec l'idée de déterminisme est une « expérience doxastique » qui provient de notre entraînement éducatif aux concepts moraux, i.e. le fait qu'il nous « semble vrai » que nous avons telles et telles capacités comme fondement à l'application de concepts tels que « je dois faire X ».

Modèle « scotiste » :

L'expérience qui génère un conflit avec l'idée de déterminisme est une expérience phénoménale des autres agents, i.e. la manière dont les *autres* nous apparaissent au moment où nous les percevons en train d'agir.

Pour l'instant, dans le présent chapitre, je n'ai pas cherché à favoriser tel modèle plutôt que tel autre. Il me semble en effet que la présomption incompatibiliste résiste aux théories de l'erreur sans qu'on ait à s'engager en faveur d'une théorie particulière sur l'origine de concept de libre arbitre. C'est-à-dire que quelle que soit l'origine de ce concept, on n'a pas de bonnes raisons de penser que le conflit qu'il génère est un conflit purement illusoire au lieu d'un conflit conceptuel véritable avec la notion de déterminisme. Cependant, si l'on désire investiguer de manière *directe* la manière dont ce concept entre en conflit avec la notion de déterminisme, il est clair qu'il nous faudra localiser ce concept pour pouvoir l'étudier.

Dans cette section, je commencerai par défendre le modèle phénoménal : on peut aisément argumenter que l'origine la plus probable de notre expérience du libre arbitre est notre expérience agentive en première personne. Ayant ainsi identifié et localisé notre

expérience du libre arbitre, je montrerai qu'on peut étudier le contenu de cette expérience de manière directe, au moyen de méthodes « phénoménologiques ». Ces méthodes révéleront un contenu conceptuel riche, dont on peut défendre de manière convaincante qu'il est incompatible avec le déterminisme causal. Ainsi la tentative pour accéder de manière directe à notre concept de libre arbitre ne fera que renforcer la présomption incompatibiliste établie dans les sections précédentes de ce chapitre.

IV.1. L'expérience phénoménale agentive comme origine du concept de libre arbitre

IV.1.a. Caractéristiques de l'expérience phénoménale agentive

Avant de défendre le modèle phénoménal par rapport aux autres modèles, il faut clarifier la nature de cette expérience phénoménale agentive sur laquelle il repose. Au chapitre 1 (p.70 sqq), j'ai eu l'occasion de présenter trois caractéristiques que *devra* posséder notre expérience du libre arbitre, quel que soit le modèle retenu. J'ai présenté cette expérience comme étant *intentionnelle* (et non pas purement *qualitative*), *véridicale* (et non pas *télique*), et *engageante* (plutôt que sans engagement véritatif). Je rappelle ici brièvement le sens et la raison de ces trois premières caractéristiques.

Lorsque nous parlons de « l'expérience agentive », nous nous intéressons non seulement à l'expérience de l'action au sens objectif (le vécu que nous avons lorsque, objectivement, nous agissons), mais surtout à l'expérience subjective de l'action (un vécu dont le *contenu* intentionnel est de nous représenter à nous-mêmes comme étant en train d'agir). La distinction est importante, car parmi les vécus que nous avons au moment où nous agissons, il se peut très bien qu'il y en ait certains (ou certains aspects) qui ne possèdent pas du tout de contenu intentionnel ; il se peut qu'il y ait de purs *qualia* dans notre expérience (objective) de l'action. Mais ces éventuels *qualia* de l'expérience agentive ne seront clairement pas pertinents pour notre problème, puisque nous recherchons un aspect de l'expérience agentive qui est à l'origine d'un conflit *conceptuel* (avec le concept de déterminisme). Il est donc à peu près certain que l'expérience recherchée devra être une expérience dotée de contenu conceptuel.

Parmi les vécus intentionnels que nous expérimentons au moment de l'action, il est à peu près indéniable que certains sont *téliques*, c'est-à-dire ayant une direction d'ajustement Esprit → Monde. Par exemple, lorsque je suis en train de construire une maison, j'ai à l'esprit une certaine représentation de la maison achevée. Cette représentation n'est clairement pas un vécu dont le contenu prétend être vrai et correspondre à la réalité extérieure ; c'est au contraire un contenu qui guide l'action. Dans la mesure où un vécu télique ne prétend pas représenter le monde, il n'est même pas susceptible d'être vrai ou faux ; il n'a pas de « conditions de véridicalité » selon l'expression de (Horgan, Tienson, and Graham 2003, 333). De nouveau, on ne voit pas comment une expérience sans véridicalité pourrait entrer en conflit avec la thèse déterministe : le vécu télique d'une maison achevée n'entre aucunement en conflit avec le fait que la maison ne soit pas achevée. Si donc notre expérience du libre arbitre entre en conflit avec l'idée de déterminisme, elle doit être véridicale, et non télique.

Plus précisément, elle doit même être *véritative*, c'est-à-dire qu'elle doit comporter une forme « d'engagement à la vérité ». Cette précision est importante pour distinguer l'expérience du libre arbitre d'une expérience telle que l'imagination ou la pure spéculation : une pure spéculation est un vécu qui possède un contenu intentionnel, et ce contenu intentionnel a des conditions de véridicalité. Par exemple, si je choisis, au milieu d'une

conférence ennuyeuse, de m'imaginer à la plage, le vécu que j'ai à ce moment-là possède un contenu intentionnel, et ce contenu intentionnel a des conditions de véridicalité – qui en l'occurrence ne sont pas satisfaites, ce vécu n'étant par conséquent pas véridique. Mais ce vécu n'a aucune *prétention* à être véridique, il ne présente pas son contenu *comme* étant vrai (bien que ce contenu soit *susceptible* d'être vrai ou faux). De nouveau, un vécu sans engagement véridatif n'a pas de raison de susciter un conflit avec l'idée de déterminisme, puisqu'il n'a aucune prétention à la vérité. Parmi les vécus que nous expérimentons au moment de l'action, il est possible que certains soient sans engagement véridatif, mais il est clair que ce ne sont pas ces aspects de l'expérience agentive qui nous intéresseront ici.

Les aspects de l'expérience agentive qui nous intéresseront seront donc uniquement ceux qui ont un contenu intentionnel susceptible d'être vrai et qui présentent ce contenu intentionnel *comme* étant vrai. D'une manière générale, j'ai appelé ce type d'expérience une expérience doxastique. Et même si l'on refuse le modèle phénoménal, il est clair que notre expérience du libre arbitre devra être, d'une manière ou d'une autre, une expérience doxastique. Dans le modèle moral, c'est l'expérience morale qui nous représente comme étant libres et nous engage à croire que nous le sommes. Dans le modèle « scotiste », c'est l'expérience perceptive de l'action d'autrui qui représente autrui comme étant libre et nous engage à croire qu'il l'est. Ces trois premières caractéristiques de l'expérience du libre arbitre constituent donc des réquisits pour tout modèle.

Est-ce que l'expérience agentive en première personne peut satisfaire ces réquisits ? C'est-à-dire, y a-t-il réellement parmi les vécus que nous expérimentons au moment de l'action certains vécus intentionnels qui nous représentent nous-mêmes *comme* étant libres (et nous engageant à croire cela) ? On pourrait en douter, en soutenant par exemple que dans l'action notre conscience est entièrement absorbée par la fin visée, de telle sorte que notre vécu serait entièrement téléique. Autrement dit, lorsque je suis en train de construire une maison, ce que j'ai à l'esprit, c'est là maison à construire ; je ne suis pas en train de penser à moi-même en tant qu'agent. Il me semble que cette difficulté repose essentiellement sur une conception « étroite » de l'expérience consciente. En suivant (Hurlburt and Schwitzgebel 2007, 229 sqq), on peut en effet distinguer une conception « riche » et une conception « étroite » de l'expérience. Selon la conception « riche », notre expérience à un instant *t* contient à la fois « un centre focal de l'expérience et une périphérie qui est d'une certaine manière plus vague ou moins vif » (p.233). La conception « étroite », à l'inverse, considère que le vécu conscient à un instant *t* se réduit à l'objet de notre attention. La distinction entre conception riche et conception étroite porte en particulier sur la question « intermodale » de la perception : lorsque je suis en train de prêter toute mon attention à un morceau de musique, mais que je garde les yeux ouverts, n'ai-je pas néanmoins une expérience visuelle, bien que cette expérience visuelle ne se situe pas au centre focal de mon attention ? Hurlburt et Schwitzgebel ont réuni des données expérimentales qui suggèrent fortement l'existence d'une expérience visuelle périphérique dans ce genre de situation. Mais si l'on se situe au niveau intra-modal, par exemple au niveau de la seule perception visuelle, il semble difficile de refuser totalement l'existence d'une périphérie de l'expérience à côté du centre focal d'attention : lorsque je fixe le centre d'une page, il est évident que mon expérience visuelle ne se limite pas au point ou au mot que je fixe, mais qu'il y a en même temps une expérience visuelle atténuée des autres parties de mon champ visuel sur lesquelles je ne centre pas mon attention. A partir de cette distinction entre centre focal et périphérie de l'expérience consciente, il me semble qu'on peut aisément repousser l'idée que l'expérience agentive serait entièrement téléique. Il est fort possible que le centre focal de notre attention, au moment où nous agissons, soit constitué d'un vécu téléique, au sens où notre attention serait fixée sur notre intention, ce que nous cherchons à réaliser. Mais il paraît clair que notre expérience agentive

contient également, au moins à titre périphérique, certaines expériences non téléiques, ne serait que pour nous informer du contexte ou de la situation dans laquelle nous agissons. Est-ce que nous avons en particulier une expérience qui nous représente nous-mêmes en train d'agir ? Il n'y a pas de raison solide d'en douter une fois que l'on a écarté l'objection de « l'attention téléique » et qu'on s'attache à décrire phénoménologiquement l'ensemble des aspects de notre expérience agentive (ce que nous ferons dans la section suivante).

Selon le modèle phénoménal, notre expérience du libre arbitre sera donc un vécu intentionnel, à engagement véridatif, et qui se situe probablement à la périphérie de notre expérience subjective au moment où nous agissons.

Il reste encore deux caractéristiques importantes à souligner pour bien distinguer le modèle phénoménal des deux autres modèles : selon le modèle phénoménal, l'expérience du libre arbitre sera une expérience *de se* et une expérience *occurrentielle*.

J'appelle *occurrentielle* une expérience doxastique qui se présente comme simultanée (ou temporellement corrélée) avec l'occurrence de l'événement qui constitue le contenu de cette expérience. Par exemple, lorsque j'ai l'expérience visuelle d'un homme en train de lever le bras, cette expérience n'est véridique que si l'homme lève le bras au moment même où j'ai cette expérience. De manière comparable, lorsque j'ai l'expérience auditive du tonnerre qui éclate, l'occurrence de mon expérience est corrélée à l'occurrence du tonnerre (bien qu'il y ait typiquement un léger retard). Toutes les expériences doxastiques ne présentent pas une telle corrélation temporelle entre l'occurrence de l'expérience et l'occurrence du contenu intentionnel. Par exemple, lorsque je considère la proposition « $7+5=12$ » et qu'elle me semble vraie, cette expérience doxastique ne présente pas son contenu comme se produisant au moment de l'expérience. L'occurrence de l'expérience doxastique est simplement déterminée par le moment où je considère la proposition : j'ai cette intuition maintenant parce que j'y pense maintenant, et non pas parce que $7+5=12$ viendrait de devenir vrai. Le cas des vérités mathématiques est particulier, car le contenu intentionnel n'a pas du tout de dimension temporelle, et qu'il est donc *impossible* d'établir une corrélation temporelle entre expérience et contenu intentionnel. Mais on peut donner également l'exemple des expériences doxastiques mémorielles : lorsqu'on me demande ce que j'ai fait ces vacances, et qu'il me semble vrai que je suis allé à la plage, cette expérience doxastique n'est pas temporellement corrélée au moment où je suis allé à la plage (si ce n'est qu'elle doit lui être postérieure, bien sûr). Le caractère *occurrentiel* ou non se remarque en particulier dans la possibilité de réitérer une même expérience doxastique. Lorsque j'ai l'impression que $7+5=12$, ou que je suis allé à la plage cet été à *t*, je peux renouveler cette expérience doxastique un peu plus tard, en repensant à ces propositions. En revanche, lorsque j'ai l'impression visuelle qu'un homme lève le bras, je ne peux pas avoir un peu plus tard une expérience de même contenu ; si j'ai un peu plus tard l'impression visuelle d'un homme qui lève le bras, ce sera l'impression d'un *deuxième* événement de lever de bras.

Notre expérience originelle du libre arbitre est-elle *occurrentielle* ou non ? Cela dépend du modèle que l'on retient. D'après le modèle moral, notre expérience doxastique d'être libres est fondamentalement fondée sur notre capacité générale à utiliser les concepts moraux (« je dois faire *x* », « il m'est permis de faire *y* »), une capacité que nous acquérons dès notre plus jeune âge, bien avant de nous demander réflexivement si nous sommes libres ou non. Ainsi, à chaque fois que nous considérons la proposition « nous sommes libres », notre schème conceptuel moral génère en nous l'impression que cette proposition doit être vraie, comme condition de possibilité à un tel schème conceptuel. Notre expérience d'être libre n'est donc pas *occurrentielle* : notre impression d'être libre n'a aucune corrélation particulière avec tel ou tel événement où se manifeste notre liberté, et elle peut se répéter à volonté (avec le même contenu) simplement en reconsidérant la proposition pertinente. A l'inverse, le modèle

phénoménal et le modèle scotiste font de l'expérience du libre arbitre une expérience occurrentielle. Pour le modèle phénoménal, mon expérience agentive me représente comme libre (plus précisément, comme exerçant ma liberté) au moment même où (d'après cette expérience) je l'exerce. Pour le modèle scotiste, mon expérience représente l'agent A comme exerçant sa liberté au moment même où (d'après cette expérience) il l'exerce.³⁴⁵ Dans les deux cas, le contenu de notre expérience du libre arbitre est indexé au temps : c'est l'expérience que « X (moi ou autrui) exerce *maintenant* sa liberté ». Et c'est seulement par généralisation que l'on peut ensuite passer à la proposition générale dispositionnelle « X a la faculté 'liberté' » puis « les hommes en général ont la faculté 'liberté' ». Dans le modèle moral, le rapport est inversé : l'expérience doxastique du libre arbitre a *initialement* un contenu général dispositionnel.

La dernière caractéristique du modèle phénoménal est d'être un modèle d'expérience *de se*, par opposition au modèle scotiste qu'on pourrait appeler *de altero*. Selon le modèle phénoménal, mon expérience du libre arbitre est d'abord l'expérience que « *je suis libre (maintenant)* ». Ceci n'empêche pas, évidemment, que je puisse ensuite, de manière dérivée, arriver à l'expérience doxastique que « mes semblables sont doués de libre arbitre », par exemple par un raisonnement d'analogie qui peut être extrêmement rapide et inconscient. Dans le modèle scotiste en revanche, mon expérience du libre arbitre est d'abord une expérience en troisième personne « telle personne que j'observe est libre (maintenant) » - ce qui n'empêche pas que cette expérience puisse être traduite ensuite en première personne si la personne que j'observe se trouve être moi-même (dans un miroir, par exemple).

Pour résumer, les trois grands modèles de l'expérience du libre arbitre sont : un modèle général dispositionnel (le modèle moral), un modèle occurrentiel *de altero* (le modèle scotiste), et un modèle occurrentiel *de se* (le modèle phénoménal). On pourrait essayer de raffiner ou de combiner ces modèles. Par exemple, (Nida-Rümelin 2007) défend un modèle dans lequel nous avons *à la fois* une expérience phénoménale agentive *de se* et une expérience perceptive *de altero* de même contenu.³⁴⁶ On pourrait également soutenir que nous avons d'une part une expérience phénoménale *de se*, qui nous donne la conviction que « je suis libre (maintenant) », et d'autre part une expérience doxastique générale à propos d'autrui, qui reposerait sur notre schème conceptuel moral (et non pas sur une analogie à partir du cas en première personne). Mais ces subtilités n'auront pas une grande importance pour notre propos : ce qui nous intéresse ici, c'est de déterminer l'origine du concept qui génère un conflit avec le déterminisme. Si nous avons deux expériences doxastiques hétérogènes, de contenus conceptuels différents, alors l'un d'eux seulement sera le concept de libre arbitre recherché, et donc l'autre expérience doxastique ne nous concernera pas ; si en revanche les deux expériences doxastiques partagent le même concept, on pourra sans danger se concentrer sur l'une des deux expériences, car elle sera *suffisante* pour localiser le concept recherché. Dans la suite de cette section, je soutiendrai que notre expérience phénoménale agentive a toutes les chances d'être la source de notre concept de libre arbitre, mais je n'exclurai pas qu'il puisse y avoir d'autres expériences doxastiques intéressantes, soit comme source « surdéterminante » du même concept, soit comme source d'un autre concept de liberté, différent de celui qui nous occupe pour le problème du libre arbitre.³⁴⁷

³⁴⁵ L'idée du modèle scotiste est celle d'une sorte de « perception » de la liberté d'autrui au moment où elle s'exerce.

³⁴⁶ « The phenomenological claim is this: it is normally part of the content of our experience when we do something that we are active in that doing; and it is part of the content of our perceptual experience when we observe others in their doings that they are active in what they do. » (Nida-Rümelin 2007, 258).

³⁴⁷ Une autre possibilité, suggérée par Cyrille Michon, serait la suivante : il se pourrait que notre expérience agentive nous donne le concept de la source ultime, donc un concept susceptible de générer un conflit avec le déterminisme, et que notre schème conceptuel nous donne un *autre* concept, celui des alternatives, également susceptible de générer un conflit avec le déterminisme, mais d'après un autre argument, ainsi que nous l'avons

IV.1.b. Arguments en faveur du modèle phénoménal

Pourquoi adopter le modèle phénoménal plutôt que le modèle moral ou le modèle scotiste ? Je vois trois arguments en faveur du modèle phénoménal : l'argument de la force, l'argument de la disponibilité, et l'argument d'intégration.

L'argument de la force

L'enjeu est donc de déterminer d'où provient notre concept de libre arbitre, c'est-à-dire notre concept qui génère un conflit avec l'idée de déterminisme. De ce concept, nous ne savons pas grand-chose à ce stade de la réflexion, si ce n'est justement qu'il génère un conflit avec l'idée de déterminisme. Mais nous savons également que ce conflit intuitif avec l'idée de déterminisme est un conflit largement répandu dans l'histoire de l'humanité, et un conflit qui est *fort* ou puissant. Dans son autobiographie, Stuart Mill décrit ce conflit entre déterminisme et libre arbitre dans des termes très forts – comme un « accès d'accablement » et un « cauchemar » (Mill 1981, 175). Autrement dit, l'expérience qui nous fournit le concept de libre arbitre, quelle qu'elle soit, doit être une expérience très forte, difficile voire impossible à éliminer. Et cette force de l'expérience du libre arbitre peut constituer un premier indice. Galen Strawson en tire argument pour soutenir que notre expérience du libre arbitre *doit* être une expérience *de se* et non pas une expérience *de altero* :

Si notre engagement à la croyance à la liberté et à la responsabilité était entièrement (ou même principalement) fondé sur notre expérience des autres personnes, alors je pense que nous n'aurions pas d'explication satisfaisante de sa force. Une telle explication est évidente lorsqu'on réalise qu'il est fondé principalement sur notre expérience de notre propre action, et seulement de manière secondaire sur notre expérience des autres personnes comme objets légitimes des attitudes réactives. Le vrai centre de notre engagement à la notion de liberté humaine se trouve dans notre attitude à l'égard de nous-mêmes et notre expérience de nous-mêmes. (G. Strawson 1986, 94–95)

Il me semble que cet argument est tout à fait suffisant pour préférer un modèle phénoménal à un modèle strictement scotiste : il est clair qu'une expérience de *mon* action a toutes les chances d'être plus forte et inéliminable que l'expérience perceptive que j'ai de l'action

vu dans la partie III. de ce chapitre. Dans ce cas, il y aurait potentiellement deux sources distinctes du « conflit originaire », et donc le « problème traditionnel » du libre arbitre serait fondamentalement un problème ambigu ou en tout « bicéphale ». La suite de cette étude permettra de voir clairement comment je traiterais cette éventualité. Ainsi que je l'annoncerai à la section IV.1.b., je suis très dubitatif quant à la possibilité de construire une épistémologie satisfaisante pour aux l'expérience doxastique morale ; c'est-à-dire que *si* notre concept des alternatives venait de notre schème conceptuel moral, je conclurais sans doute que notre conviction d'avoir des alternatives possibles est réfutée par l'objection épistémologique. Mais je continuerais à penser que *l'autre* source de concept incompatibiliste, l'expérience agentive, nous donne une vraie connaissance de l'indéterminisme des actions (j'argumenterai en effet au chapitre 6 que l'expérience agentive *peut* constituer une source de connaissance). Si donc « le libre arbitre », le concept à l'origine du conflit originaire, était double (source ultime *via* l'expérience agentive, et alternatives *via* le schème moral), ma réponse serait partiellement un déterminisme dur (nous ne pouvons pas savoir que nous avons les alternatives) et partiellement un libertarianisme (nous pouvons savoir que nous sommes source ultime). Cette position ne serait pas très confortable. Mais je doute fort que cette éventualité soit réelle, parce qu'elle repose sur l'idée que nous n'avons *aucune* expérience agentive des alternatives ; or si l'on peut douter qu'il y ait une expérience des alternatives dans *toute* action, il me semble très difficile de contester que nous avons une expérience des alternatives à *l'occasion des choix difficiles*, ainsi que je le défendrai à la section IV.2.c. Je ne suis donc pas prêt à penser que notre croyance aux alternatives reposerait *uniquement* sur une source « inintégrable » telle que la source morale ; elle repose certainement, au moins en partie, sur l'expérience du choix, pour laquelle il existe une solution d'intégration (cf. chapitre 6).

d'autrui. Est-ce que cet argument permet également de préférer le modèle phénoménal au modèle moral ? Théoriquement, oui : en effet, le modèle moral, bien qu'il ne soit pas occurrence, doit être considéré comme un modèle *de altero*. Notre expérience doxastique, dans ce modèle, a initialement un contenu en troisième personne (« les hommes sont libres »), ou à la rigueur un contenu en première personne du pluriel (« nous sommes libres »), mais pas un contenu en première personne du singulier. Cela dit, le modèle moral pourrait posséder d'autres ressources pour rendre compte de la force de notre expérience doxastique : en effet, si notre expérience doxastique repose sur notre schème conceptuel le plus fondamental, en constituant une condition de possibilité de ce schème, on peut comprendre également que cette expérience soit très forte et quasi inéliminable.

Pour ma part, si l'argument de la force était le seul argument pertinent pour cette discussion, je resterais prudemment agnostique entre le modèle phénoménal et le modèle moral – en acceptant, encore une fois, la possibilité d'un modèle qui combinerait ces deux sources. Pour aller plus loin, il est donc important à mon sens de considérer les arguments suivants.

L'argument de la disponibilité

Puisque nous cherchons une expérience dont le contenu génère un conflit avec l'idée de déterminisme, une autre stratégie consiste à se demander quelle expérience *pourrait* générer un tel conflit : y a-t-il une expérience qui soit disponible pour jouer ce rôle ? L'expérience phénoménale agentive a-t-elle une telle potentialité ? Et qu'en est-il de l'expérience doxastique liée à notre schème moral conceptuel ? Pour répondre à ces questions, il faut entrer dans l'analyse de ces expériences pour elles-mêmes ; pour ce qui est de l'expérience agentive, l'analyse de son contenu occupera la section suivante, et l'argument de disponibilité suppose donc d'anticiper un peu sur les résultats de cette section. Dans la section suivante, je défendrai que le contenu de l'expérience phénoménale agentive a toutes les chances d'être incompatible avec le déterminisme causal. Pour établir ce point, je m'appuierai sur des analyses « phénoménologiques », mais aussi sur certains résultats de X-Phi, notamment (Deery, Bedke, and Nichols 2013) qui se sont efforcés d'étudier *spécifiquement* notre expérience agentive (et non pas nos jugements sur le mot « libre arbitre »), et ont mis en évidence une forte présomption incompatibiliste. Si l'on accepte ce résultat, cela signifie qu'elle a aussi toutes les chances de générer un conflit intuitif avec l'idée de déterminisme, et elle est donc disponible pour jouer le rôle de source du concept de libre arbitre. Pour résumer, une bonne raison de penser que notre conflit intuitif avec le déterminisme provient de l'expérience agentive, et non pas d'autres expériences doxastiques plus complexes, c'est que l'expérience agentive fournit tout ce qu'il faut pour générer un tel conflit ; il n'y a donc pas lieu d'aller chercher ailleurs.

Qu'en est-il de l'expérience doxastique liée à notre schème moral conceptuel ? Rien n'empêche que cette expérience fournisse également un concept incompatibiliste. Dans ce cas, il y aurait une sorte de « surdétermination » incompatibiliste. Mais, comme je l'ai défendu un peu plus haut, on pourra dans ce cas se concentrer sur une des deux sources de concept incompatibiliste, puisque ce qui nous intéresse présentement est seulement de savoir si notre expérience de nous-mêmes et de nos actions est compatible ou non avec le déterminisme causal ; une seule expérience incompatibiliste suffit à trancher la question.

Aussi, la conclusion de l'argument de la disponibilité est la suivante : quoi qu'il en soit de notre schème moral conceptuel, notre expérience phénoménale agentive fournit un concept incompatibiliste, et elle a donc toutes les chances de constituer la source (ou éventuellement *une* source surdéterminée) de notre conflit intuitif avec l'idée de déterminisme.

L'argument d'intégration

Il y a une dernière raison, plus fondamentale, de préférer le modèle phénoménal au modèle moral (ou au modèle scotiste). Comme l'argument précédent, cette raison suppose d'anticiper sur des discussions qui sont encore à venir, et plus précisément sur la discussion du problème d'intégration lui-même au chapitre 6. L'argument le plus fort, à mon sens, en faveur du modèle phénoménal, c'est qu'il est le meilleur modèle pour relever le défi de l'intégration. Ou pour le dire autrement, le modèle phénoménal présente des avantages épistémologiques considérables par rapport aux autres modèles. Dans cette section, je ne ferai que mentionner brièvement ces avantages épistémologiques, sans chercher à montrer qu'ils permettent de résoudre parfaitement le problème d'intégration. Pour la solution complète au problème d'intégration (dans le modèle phénoménal), il faudra attendre le chapitre 6 qui y est consacré.

Le grand avantage du modèle phénoménal, d'un point de vue épistémologique, réside dans les deux caractéristiques que nous avons mentionnées plus haut, c'est-à-dire le fait qu'il s'agit d'un modèle *occurrentiel* et d'un modèle *de se*. Chacune de ces deux caractéristiques se révélera extrêmement avantageuse lorsqu'il s'agira de montrer que notre expérience du libre arbitre peut éviter le hasard épistémique, et par conséquent constituer une source de *connaissance* et pas seulement une source de justification. D'une manière générale, éviter le hasard épistémique suppose de mettre en évidence une forme de « dépendance » entre le fait qui est cru et la croyance, ou entre le fait qui semble être le cas et l'expérience doxastique qui le présente comme étant le cas. On peut voir assez facilement qu'une expérience *occurrentielle* et *de se* aura de meilleures chances d'assurer une forme de dépendance fait-expérience qu'une expérience non *occurrentielle* ou qu'une croyance *de altero*.

Une expérience *occurrentielle*, comme on vient de le voir, se caractérise par une forme de *corrélation* temporelle entre l'occurrence de l'expérience et l'occurrence de son contenu intentionnel. C'est au moment même où le lion marche que j'ai l'expérience visuelle d'un lion qui marche, et lorsqu'il ne marche pas, je n'ai pas cette expérience visuelle. Ou en tout cas, c'est ainsi que se *présente* mon expérience doxastique. Il pourrait bien sûr s'agir d'une illusion, mais en tout cas, mon expérience doxastique se présente elle-même comme corrélée, ou comme temporellement *dépendante*, du fait dont elle est l'expérience. Cette corrélation fait-expérience, bien qu'elle s'exprime en termes temporels plutôt que modaux, est clairement très proche de la condition de *traçage* (ou de *sensibilité*) dont nous avons vu au chapitre 3 (p.225 sqq) qu'elle constituait une solution importante au problème du hasard épistémique : quand le lion est en train de marcher, j'ai l'expérience visuelle qu'il est en train de marcher, et quand il ne l'est pas, je n'ai pas cette expérience visuelle. On pourrait dire que les expériences *occurrentielles*, bien qu'elles puissent être hasardeuses en fait, se présentent elles-mêmes comme épistémiquement « traçantes », et par conséquent, elles suggèrent par leur nature même une possible théorie anti-hasard. Il n'en va pas de même des expériences non *occurrentielles* : par exemple, le souvenir d'avoir été à la plage cet été n'a pas besoin d'être corrélé temporellement au fait d'avoir été à la plage cet été, et il peut se reproduire à différents moments pour des raisons diverses. On peut bien sûr trouver une théorie pour expliquer que cette expérience mémorielle est sensible d'un point de vue modal (si je n'avais pas été à la plage, je n'aurais pas dans le cerveau telle trace mémorielle qui est à l'origine de mes diverses expériences mémorielles d'avoir été à la plage), mais dans ce cas, la dépendance modale n'est pas immédiatement suggérée par la nature même de l'expérience, ou par la répartition temporelle de ses occurrences.

Ainsi, dans le modèle phénoménal du libre arbitre (ou dans le modèle scotiste), l'expérience du libre arbitre se présente comme ayant lieu quand, et seulement quand, cette liberté est mise en œuvre. Elle se présente donc elle-même comme « traçante », et suggère ainsi une possible théorie anti-hasard. Dans le modèle moral, en revanche, notre expérience

doxastique n'offre pas le début d'une théorie anti-hasard : notre expérience doxastique du libre arbitre peut avoir lieu n'importe quand, sans aucune corrélation avec les occurrences éventuelles d'actes libres. Et contrairement au cas de la mémoire, il ne semble pas facile d'imaginer une autre théorie de la dépendance fait-expérience. Le modèle moral dit simplement que l'expérience du libre arbitre repose sur notre schème conceptuel moral, mais on ne voit pas clairement comment notre schème conceptuel pourrait *dépendre* du fait que nous avons telle ou telle faculté. En particulier, si l'on imagine que notre expérience du libre arbitre suppose *l'indéterminisme* causal, on ne voit pas comment notre schème conceptuel moral pourrait *dépendre* d'aucune façon de l'indéterminisme de la nature, alors que le modèle phénoménal ou le modèle scotiste suggèrent au moins une forme de corrélation temporelle entre certains événements indéterministes (les actes libres que j'accomplis, ou que quelqu'un d'autre accomplit) et notre expérience du libre arbitre. Dans une métaphysique théiste, on pourrait sans doute défendre que l'auteur de la nature nous a donné des schèmes conceptuels adaptés à la réalité physique qu'il a créée ; dans ce cas, on pourrait défendre qu'il y a bien une dépendance réelle entre notre schème conceptuel moral et la réalité de l'indéterminisme causal : si Dieu crée un monde indéterministe, il nous crée avec un schème conceptuel qui génère une intuition du libre arbitre ; et si Dieu avait créé un monde déterministe, il nous aurait créé avec un autre schème conceptuel, qui ne génère pas cette intuition. On résoudrait ainsi le problème de la corrélation fait-croyance par une forme « d'harmonie préétablie » entre le monde et nos facultés de connaissance. Cette sorte d'épistémologie est défendable, mais elle a deux inconvénients. D'une part, elle reviendrait à dire que les athées ne peuvent pas résoudre le problème de l'intégration pour le libre arbitre.³⁴⁸ D'autre part, elle a un caractère trop général, qui s'applique indifféremment à toutes les expériences épistémiques ; en effet, dans une telle épistémologie, même nos expériences perceptives ne sont fiables ultimement que parce que Dieu nous a donné une faculté de perception fiable... mais cela n'empêche pas qu'il y ait plus à dire sur leur fiabilité (et leur traçage) que la simple harmonie contrefactuelle liée à la bonté divine. De même, si l'on adopte une épistémologie de type phénoménal et introspectif pour le libre arbitre, on pourra dire aussi que la fiabilité de notre introspection du libre arbitre est ultimement due au caractère non trompeur de Dieu ; mais on aura tout de même à disposition une théorie plus riche du traçage fait-expérience que dans le modèle moral qui *réduit* le traçage fait-expérience à la simple bonté divine. Pour le dire autrement, même dans une épistémologie théiste de ce genre, les modèles occurrentiels présentent un avantage épistémologique par rapport aux modèles non occurrentiels qui doivent réduire leur épistémologie à la simple harmonie préétablie.

D'une manière générale, il me semble donc évident qu'un modèle d'expérience *occurrentielle* offre des ressources bien plus précieuses que le modèle moral lorsqu'on s'intéresse spécifiquement au problème de l'intégration, ou au problème anti-hasard. De fait, les auteurs qui ont défendu une version ou une autre du modèle moral, l'ont fait d'une manière qui, soit ne manifestait aucun intérêt particulier pour la question épistémologique, soit manifestait un intérêt simplement pour la question de la *justification*. C'est le cas en particulier de Peter van Inwagen. Pour ce dernier, nous pouvons « savoir » que nous sommes libres parce que nous pouvons « savoir » que nous sommes moralement responsables et que l'un implique l'autre. Mais lorsqu'il s'agit de défendre que notre conviction de la responsabilité morale peut constituer une « connaissance », voici ce que dit van Inwagen :

³⁴⁸ C'est la position défendue, comme on sait, par Descartes à propos des mathématiques : les athées ne peuvent pas être ultimement justifiés dans leurs croyances mathématiques parce qu'ils ne peuvent pas concevoir que leurs intuitions mathématiques dépendent fiablement des « faits » (ou des vérités) mathématiques. Seuls les théistes peuvent croire qu'il y a une telle dépendance fiable entre intuitions mathématiques et vérités mathématiques, grâce à la bienveillance du Dieu non trompeur qui a créé à la fois notre entendement et les vérités mathématiques.

Je crois que, si nous examinons nos convictions avec honnêteté, avec sérieux, et avec attention, [...] nous découvrirons que nous ne pouvons pas faire autrement que de considérer notre croyance en la responsabilité morale comme une croyance justifiée, une croyance qui n'est simplement pas ouverte au doute raisonnable.
(van Inwagen 1983, 209)

Van Inwagen a peut-être raison de considérer notre croyance en la responsabilité morale comme justifiée, mais cela n'en fait pas pour autant une connaissance (tout ce que dit ici van Inwagen est compatible avec une position qui refuserait le principe de méta-cohérence et considérerait que notre croyance en la responsabilité reste justifiée bien qu'elle ne puisse constituer une connaissance). Ce qui manque ici, c'est une réflexion sur le type de *dépendance* qui pourrait exister entre le fait que nous soyons moralement responsables et le fait que nous croyions l'être. Et il n'est pas étonnant que la question soit éludée par van Inwagen car c'est un défi particulièrement difficile à résoudre dans l'épistémologie qu'il adopte. Il me semble évident que, dès qu'on s'intéresse au problème de la connaissance proprement dit, ou au problème anti-hasard, on aperçoit immédiatement les avantages d'une épistémologie occurrentielle par rapport à une épistémologie fondée sur la moralité.

Deuxièmement, l'intérêt pour la question anti-hasard fait aussi apparaître un avantage d'une épistémologie *de se* par rapport à une épistémologie occurrentielle *de altero* telle que l'épistémologie scotiste. Pour quelle raison ? En termes relativement vagues, la réponse est la suivante : parce que dans un modèle *de se*, l'expérience est plus « proche » du fait expérimenté que dans un modèle *de altero*. Dans un modèle *de se*, le fait expérimenté (le fait que « j'agis de manière libre à *t* ») est un fait qui se situe, peut-on supposer, quelque part dans mon corps et/ou dans mon esprit ; et l'expérience phénoménale elle-même (l'expérience d'être en train d'agir librement) est également un fait qui se situe, peut-on supposer, quelque part dans mon corps et/ou dans mon esprit. Les deux faits sont donc « localement » proches. Cela ne suffit pas à établir que l'un dépend de l'autre, mais les conditions pour une telle dépendance sont clairement meilleures que dans un modèle *de altero*. En effet, dans ce second modèle, le fait expérimenté est réalisé quelque part dans le corps et/ou l'esprit d'un homme qui se trouve en face de moi, tandis que l'expérience elle-même se trouve toujours dans mon corps et/ou dans mon esprit. Cela n'exclut pas qu'on puisse trouver une forme de dépendance entre ces deux faits, mais de nouveau, si on peut trouver une telle dépendance, il semble évident qu'on pourra beaucoup plus facilement, et plus immédiatement, trouver une dépendance entre *ma* liberté et mon expérience. Au chapitre 6, nous verrons de manière plus précise le genre de dépendance (causale) que cette proximité locale rend possible dans le cas de l'expérience *de se*. Pour l'instant, il est suffisant de remarquer, de manière générale et simplement comparative, que l'épistémologie *de se* a toutes les chances d'être plus simple et plus directe que l'épistémologie *de altero*.

Pourquoi trouve-t-on alors des défenses, au moins partielles, du modèle *de altero* ? De nouveau, la réponse tient, me semble-t-il, à un manque d'attention au problème de l'Intégration. Si l'on pense, par exemple, à la défense par Martine Nida-Rümelin d'un modèle mixte *de se, de altero*, on peut voir que l'épistémologie sous-jacente est essentiellement une motivation de type *justificatiionnelle* :

Nous nous intéressons aux théories philosophiques qui n'attribuent pas une erreur massive à nos expériences normales de nous-mêmes et des autres dans l'action.
(Nida-Rümelin 2007, 263)

Sans doute sommes-nous *justifiés*, au moins *prima facie*, à rejeter les théories qui nous attribuent des erreurs massives, ou à accepter celles qui n'impliquent pas de telles erreurs ;

cela semble un bon principe pour déterminer quelles théories sont justifiées ou non, mais cela ne donne pas le début d'une indication sur la manière dont ces théories pourraient constituer des *connaissances* ou des croyances non hasardeuses du point de vue externaliste. Et dès qu'on s'intéresse à la contrainte de dépendance externaliste, il me semble évident qu'on doit reconnaître le grand avantage épistémologique de l'expérience *de se* par rapport à l'expérience *de altero*.

Pour conclure, il y a principalement trois raisons de penser que le conflit intuitif avec le déterminisme, et le concept de libre arbitre sur lequel il repose, provient de notre expérience phénoménale en première personne. Mais ces raisons sont de force et de portée différentes. L'argument de la force de l'expérience permet sans doute de rejeter une source de type « scotiste », mais il n'est pas clair qu'il soit assez fort pour rejeter le modèle moral. L'argument de la disponibilité permet de montrer que l'expérience phénoménale a de très bonnes chances d'être au minimum une des sources du conflit, mais là encore, cela n'exclut pas totalement qu'il y ait une autre source surdéterminante, par exemple dans notre schème conceptuel moral. Ultimement, la raison principale de préférer le modèle phénoménal se situe pour moi dans le problème d'intégration : l'expérience agentive phénoménale est le seul candidat capable de constituer réellement une source de connaissance non hasardeuse de son contenu. Dans la mesure où nous n'avons pas encore mené l'analyse conceptuelle du contenu de notre expérience phénoménale, ni abordé les résolutions du problème d'intégration, les deux derniers arguments (qui sont les plus importants) sont présentés ici de manière anticipée. Si l'on voulait être parfaitement exhaustif, il faudrait peut-être envisager pour chacun des modèles d'abord son analyse conceptuelle puis son rapport au problème d'intégration, mais les éléments que je viens d'indiquer suffiront, je pense, à se convaincre que le modèle phénoménal est le meilleur candidat pour mener jusqu'au bout le problème, et c'est donc sur ce modèle que nous nous concentrerons à présent, pour montrer qu'il *peut* effectivement aller jusqu'au bout d'une défense épistémologique satisfaisante.

Pour toutes ces raisons, je tiendrai désormais pour une hypothèse acquise que notre concept de libre arbitre provient de notre expérience phénoménale agentive.

IV.2. Le contenu de l'expérience phénoménale agentive

Si notre concept de libre arbitre provient de notre expérience phénoménale agentive, reste à savoir si cette expérience fournit un concept *réellement* incompatibiliste, ou si elle fournit un concept qui génère seulement une *illusion* incompatibiliste. Pour cela, nous allons mener une enquête qu'on peut appeler « phénoménologique » dans le sens où l'on cherchera à décrire le contenu conceptuel d'une expérience phénoménale donnée. Dans ce sens général, la phénoménologie de l'action est une entreprise qui a suscité de plus en plus d'attention durant ces dix dernières années.³⁴⁹ Je m'appuierai largement sur les résultats de ces études récentes, qui permettent de se convaincre que la tâche phénoménologique, malgré ses difficultés, n'est pas impossible. Les deux grandes difficultés que l'on rencontre sont les suivantes : il y a d'une part une difficulté de principe à vouloir décrire de manière scientifique et objective un objet (l'expérience) qui est par définition irréductiblement subjectif et non partageable. Deuxième difficulté, la phénoménologie agentive est extrêmement riche et contient une quantité de composantes parfois difficiles à isoler les unes des autres et parmi lesquelles il est

³⁴⁹ Il faut signaler ici un article qui a anticipé d'une quarantaine d'année ce développement récent de la discussion phénoménologique : dans « The Feeling of Freedom » (1964), Douglas Browning a mené exactement ce type d'enquête (c'est-à-dire une enquête strictement phénoménologique destinée à servir à la discussion entre libertariens et compatibilistes), et ses classifications aussi bien que ses résultats restent tout à fait pertinents.

aisé de se perdre. Je répondrai donc brièvement à ces deux objections avant d'entrer dans le détail de notre phénoménologie du libre arbitre. Enfin, je soutiendrai que notre phénoménologie du libre arbitre suggère l'existence d'une expérience agentive réellement incompatibiliste.

IV.2.a. Préliminaires méthodologiques

Comment déterminer de manière scientifique ou en tout cas « objective » une expérience en première personne ? En particulier, dans le cas du libre arbitre et de la phénoménologie agentive, comment trancher un désaccord entre un philosophe qui prétendrait avoir une expérience agentive incompatibiliste et un philosophe qui prétendrait avoir une expérience agentive compatibiliste ? N'arrivera-t-on pas rapidement à une situation dialectique « bloquée », parole contre parole ? Même en supposant que chacun des deux philosophes soit sincère, ne peut-on pas imaginer que l'un ou l'autre (ou les deux) soit *biaisé* dans son appréciation du contenu de son expérience subjective ? Le problème qui se pose ici est le suivant : pour connaître une expérience irréductiblement subjective, notre seul moyen d'accès sera *l'introspection* (au sens étroit et non plus au sens large), mais l'introspection est-elle toujours absolument fiable, susceptible de constituer une source d'information scientifique ? La meilleure réponse à ce type d'objection est, à mon sens, de « prouver le mouvement en marchant » : les études publiées depuis une dizaine d'années ont permis de faire apparaître des points d'accord et des critères d'objectivité dans la description phénoménologique de l'action. Mais il sera utile néanmoins de présenter brièvement la méthodologie (parfois implicite) qui a permis d'arriver à de tels accords.

Toute méthode phénoménologique, on vient de le voir, reposera ultimement sur l'introspection, c'est-à-dire sur l'attention portée par un sujet à sa propre expérience subjective (et à son contenu). Mais on peut encore distinguer plusieurs méthodes selon les réponses apportées à trois questions : qui ? quand ? comment ? Qui est le sujet d'expérience ? Est-ce le phénoménologue lui-même (le « scientifique, ou le théoricien) ou bien un sujet « cobaye » étudié par le phénoménologue ? A quel moment l'expérience est-elle observée et décrite ? Au moment même où l'expérience se produit, ou instantanément après, ou bien à un moment postérieur indéterminé en ayant recours à notre mémoire de l'expérience passée ? Enfin, comment l'expérience est-elle observée ? De manière totalement neutre ou en portant une attention spécifique à certains aspects en fonction des questions auxquelles on cherche à répondre ?

Ces trois critères permettent de distinguer différentes méthodologies phénoménologiques. Par exemple, la méthode de « Descriptive Experience Sampling » (DES) de (Hurlburt and Akhter 2006) est une méthode qui consiste à interviewer un sujet cobaye, qui aura pris soin de décrire son expérience de manière absolument libre (et non guidée par des questions), et ce au moment même où il a l'expérience, ou immédiatement après (la méthode consiste à donner au sujet un « beeper », et à lui demander d'écrire son expérience interne dès que le beeper sonne, pour pouvoir ensuite présenter ce résultat à l'expérimentateur lors d'un entretien). La motivation de Hurlburt et Schwitzgebel est assez claire : il s'agit pour eux d'avoir une description la moins « biaisée » possible de la description phénoménologique. Si l'expérimentateur décrivait sa propre expérience, il pourrait être biaisé dans sa description par ses attentes scientifiques ou par ses compétences conceptuelles. De même, si on posait des questions au sujet, son observation pourrait être biaisée par ces questions mêmes et leur conceptualité propre, qui ne correspond peut-être pas à la conceptualité de l'expérience en elle-même. Enfin, si l'expérience est rapportée longtemps après, il y a un risque de souvenir erroné ou imprécis. Il faut voir cependant que ce type de méthode a aussi ses limites. En

particulier, le fait de demander une description absolument neutre a pour conséquence de ne délivrer, en général, que les éléments de phénoménologie qui font le « centre focal » de l'expérience subjective (voir plus haut p.309) ; les éléments qui se trouvent à la périphérie du champ d'expérience auront peu de chance d'être décrits dans la méthode DES, et par conséquent, si l'on s'intéresse à des aspects de la phénoménologie qui se trouvent typiquement en périphérie, il semble qu'on soit obligé d'avoir recours à une enquête guidée plutôt qu'à une enquête libre. Comme on le verra dans la section suivante, ce type de problème se pose particulièrement pour un phénoménologue qui s'intéresse à l'expérience de l'action *dans le but* d'y trouver des aspects phénoménologiques qui génèrent un conflit avec le déterminisme. Il existe alors d'autres méthodes d'interview guidée, notamment la « description en deuxième personne » de (Petitmengin 2006). A ma connaissance, aucune méthode d'interview (ni libre, ni guidée) n'a été menée à ce jour dans le domaine de la phénoménologie du libre arbitre. Le propre d'une méthode d'interview est de se concentrer sur *un* sujet d'expérience cobaye, afin de pouvoir obtenir de lui des descriptions plus approfondies. Mais l'inconvénient corrélatif est la possibilité que le cobaye retenu n'ait pas une phénoménologie typique ou représentative de la population.

Pour obtenir des résultats qui soient à la fois sensibles aux aspects fins de l'expérience et représentatifs d'une population, on peut avoir recours à un questionnaire systématique posé à un grand échantillon de sujets. En ce sens, certaines enquêtes que l'on classe aujourd'hui dans la famille de la « philosophie expérimentale » sont clairement des enquêtes de phénoménologie guidée, qui ont seulement la particularité d'être menées sur un grand nombre de cobayes à la fois. C'est le cas notamment de (Deery, Bedke, and Nichols 2013) pour le problème du libre arbitre. Dans cette étude, Deery et al. demandent explicitement aux sujets interrogés de « considérer la manière dont les choses [leur] semblent / apparaissent »³⁵⁰ dans une situation d'action donnée (en l'occurrence, il s'agit d'une situation de choix entre deux directions à prendre dans un parcours de luge, et on demande aux sujets d'imaginer qu'ils se trouvent dans cette situation). Là où la plupart des dispositifs de philosophie expérimentale demandent seulement au sujet de rapporter une opinion ou un jugement, le dispositif de Deery et al. demande de prêter attention à une expérience phénoménale, et de tâcher de la décrire.³⁵¹ Le premier type de résultats dont nous disposons aujourd'hui concernant la phénoménologie de l'action repose donc sur cette méthode de description en deuxième personne, mais guidée, et effectuée sur un large échantillon. Le large échantillon de sujets permet d'avoir un résultat solide, à l'abri des éventuels biais subjectifs individuels.

Mais l'investigation en deuxième personne, quel que soit l'échantillon, aura toujours des limites quant à sa qualité descriptive. C'est pourquoi de plusieurs auteurs, notamment (Horgan, Tienson, and Graham 2003; Holton 2006; Nida-Rümelin 2007), ont mené une enquête en première personne, c'est-à-dire où le sujet d'expérience est l'investigateur lui-même. Malgré ses risques de biais théorique, il n'est en effet pas légitime de refuser absolument cette méthode qui a le grand intérêt de donner des résultats plus précis et plus ordonnés. De fait, la meilleure méthode consiste probablement à *combiner* les différentes méthodes pour contrebalancer leurs inconvénients respectifs. Dans le cas du libre arbitre, il me semble que l'on peut aujourd'hui bénéficier à la fois de résultats assez précis grâce à la

³⁵⁰ « Consider how things seem to you as you approach the fork in the path » (Deery, Bedke, and Nichols 2013, 132).

³⁵¹ On pourrait objecter que le scénario de la luge, n'étant pas véritablement expérimenté par les sujets, mais seulement imaginé, le dispositif expérimental ne permet pas de tester l'expérience phénoménale que les sujets ont réellement mais seulement celle qu'ils imaginent qu'ils auraient. Pour éviter ce problème, Deery et al. ont développé une deuxième étude, dans laquelle ils ont placé les sujets devant un véritable choix à faire (donner 50 cents à une association ou à une autre) avant de leur demander de décrire leur expérience de ce choix. Les résultats, que nous verrons plus loin, ne différaient pas sensiblement de la description phénoménologique du choix simplement « imaginé » (Deery, Bedke, and Nichols 2013, 135–140).

méthodologie en première personne, et d'une confirmation des grandes lignes de ces premiers résultats grâce à la méthode de X-Phi, qui permet de montrer que les données de la phénoménologie directe ne concernent pas simplement les expériences biaisées des philosophes professionnels. Comment fait-on de la phénoménologie en première personne ? Cela consiste-t-il simplement à donner une description de son expérience subjective en espérant que les autres philosophes accepteront cette description sans argument ? Pas uniquement. Il y a au moins un élément de méthode repérable qui permet d'argumenter en vue d'un accord intersubjectif. C'est ce que l'on pourrait appeler la « méthode différentielle ». Par exemple, si je dis que l'expérience agentive de lever le bras contient une expérience du mouvement de bras « comme volontaire », je peux argumenter en faveur de l'existence de ce contenu en montrant ce que serait une expérience de contenu contraire, en l'occurrence une expérience du mouvement de bras « comme se produisant simplement ». Chacun peut comprendre ce que serait l'expérience d'un mouvement de bras « comme se produisant simplement », et chacun peut reconnaître que notre expérience agentive de lever le bras n'est *pas* cette expérience-là. Par différentiation, chacun peut donc se convaincre que notre expérience agentive du lever de bras a un contenu plus riche qu'un simple contenu d'événement de mouvement.

Dans cette section, j'utiliserai donc les résultats de deux méthodologies phénoménologiques distinctes : une méthodologie en première personne et une méthodologie en deuxième personne (un certain type de philosophie expérimentale).

IV.2.b. Les différentes composantes de la phénoménologie agentive

La phénoménologie agentive est incontestablement une phénoménologie assez riche, avec des composantes diverses et variées. Toutes ces composantes, ou tous ces aspects, ne seront pas pertinentes pour notre propos qui concerne seulement le libre arbitre (c'est-à-dire cet aspect de notre expérience de nous-mêmes et de nos actions qui a une propension à entrer en conflit avec l'idée de déterminisme). Mais pour pouvoir mettre en évidence les composantes pertinentes, il sera utile d'avoir un panel de l'ensemble de la phénoménologie agentive. Les auteurs qui ont mené une enquête phénoménologique en première personne ont proposé diverses classifications pour ordonner toutes les composantes de cette phénoménologie. Je retiendrai ici deux classifications importantes.

La première classification importante est la distinction entre expérience de l'action volontaire et expérience du choix. Lorsqu'on pense à l'action volontaire dans un contexte philosophique, on a tendance à considérer le cas idéal dans lequel un mouvement (appuyer sur la gâchette) réalise une intention (tuer Kennedy) que j'ai formée à la suite d'une délibération et d'un choix (tel jour, à telle heure, pour telles raisons). Mais ainsi que le rappelle notamment (Holton 2006, 2), toute action n'est pas nécessairement précédée d'un choix (la plupart de nos actions quotidiennes sont faites d'après des intentions qui ne requièrent de notre part aucune délibération), et même pour les actions qui *sont* précédées d'une décision ou d'un choix, on peut clairement distinguer le moment de la décision et le moment de l'action. Et il n'est pas clair du tout que ces deux moments suscitent des phénoménologies comparables. Il faut donc distinguer clairement entre l'expérience que nous faisons au moment du choix et l'expérience que nous faisons au moment de l'action volontaire. Notez que cette première classification n'est pas une classification de différents *contenus* de notre expérience agentive, mais plutôt une classification entre deux expériences agentives qui sont objectivement distinctes.³⁵²

³⁵² Voir la distinction au chapitre 1 (p. 70) entre sens objectif et sens subjectif de l'expérience

La seconde classification est plus complexe, c'est celle qui concerne les différents contenus de nos expériences agentives. Quel est le contenu intentionnel de notre expérience lorsque je choisis de tourner à gauche plutôt qu'à droite, ou lorsque je peins volontairement un tableau ?

La première chose à remarquer, c'est que notre expérience agentive contient sans nul doute possible des contenus intentionnels *non véridicaux* : quand je peins volontairement un tableau, un contenu important de mon expérience à ce moment-là est l'idée du tableau réalisé (sinon de manière totalement détaillée, du moins dans les grandes lignes). On peut penser également au skieur qui, au moment où il descend la piste, a une représentation des virages qu'il va devoir prendre bientôt. Ce genre de contenu qui représente l'ensemble de l'action, ou du moins les parties à venir immédiatement, ne prétend pas représenter le monde tel qu'il est, mais tel qu'il doit être pour que l'action soit réalisée. C'est ce que nous avons appelé plus haut un contenu *télique*. Il est important de noter que notre expérience de l'action non seulement a un contenu télique, mais qu'en outre ce contenu télique constitue probablement le centre focal de notre attention au moment de l'action. Ce qui occupe toute l'attention du skieur, ce sont les virages à venir. On peut faire le même genre de remarque pour l'expérience du choix : lorsque j'opère un choix entre tourner à droite et tourner à gauche, mon esprit est principalement occupé par deux contenus, « je tourne à droite » et « je tourne à gauche ». Mais aucun de ces deux contenus ne se présente comme correspondant au monde tel qu'il est. Au contraire, ces contenus se présentent comme « optionnels » ou simplement possibles. Il faut bien distinguer ici entre le contenu « je tourne à droite » et le contenu « je pourrais tourner à droite » : le premier est un contenu non véridical (qui ne prétend pas décrire un état de fait), tandis que le second prétend décrire un état de fait (il peut être faux). Il est probable que notre expérience du choix contient l'un et l'autre contenus, mais il est aussi très probable que le centre de notre attention au moment du choix est le premier, non pas le second : quand j'opère un choix entre deux scénarios, je considère normalement que ces deux scénarios sont *possibles*, mais c'est là une présupposition non problématique à laquelle je porte à peine attention ; toute mon attention est portée sur les *détails* des deux scénarios (ce qui se passerait si je tournais à droite, ou à gauche), et non pas sur *le fait que tel scénario soit possible*. Pour résumer, l'expérience du choix et l'expérience de l'action volontaire ont toutes deux des composantes non véridicales (l'expérience télique et l'expérience des options) en plus de leurs composantes véridicales. Les seules composantes importantes pour nous (comme on l'a vu plus haut) seront les composantes véridicales... mais il faut noter que ces composantes véridicales sont probablement des aspects *périphériques* de nos expériences agentives et non pas leur *centre focal*. C'est à cause de cela que la méthodologie non guidée de Hurlburt et Schwitzgebel ne sera pas adaptée : pour obtenir une bonne description des composantes véridicales de notre phénoménologie agentive, il est presque indispensable que le sujet qui d'expérience soit guidé dans son attention introspective, ou qu'il soit l'investigateur lui-même sachant ce qu'il recherche.

Passons à présent aux composantes véridicales de notre expérience de l'action. On rencontre pour cela une certaine variété de classifications dans la littérature.³⁵³ On peut cependant remarquer un intérêt récurrent pour deux aspects importants, sous diverses terminologies : d'une part l'aspect des possibilités alternatives, et d'autre part l'aspect du soi

³⁵³ Par exemple, (Horgan, Tienson, and Graham 2003) distinguent (i) l'aspect du soi-comme-initiateur, (ii) l'aspect de finalité (*purposiveness*), et (iii) l'aspect d'être volontaire (*voluntariness*, par où ils entendent en fait l'aspect de « pouvoir faire autrement »). (Bayne and Levy 2006) distinguent (i) la phénoménologie de l'action mentale, (ii) la phénoménologie d'être l'auteur de l'action (*authorship*), et (iii) la phénoménologie de l'effort. (Pacherie 2008) distingue (i) le sens de la causalité intentionnelle, (ii) le sens de l'initiation, et (iii) le sens du contrôle.

comme source. De fait, ce sont ces deux aspects qui s'avèreront pertinents pour la question du libre arbitre. Mais essayons de mettre un peu d'ordre dans ces classifications de manière générale. On peut ordonner les différents aspects de la phénoménologie en fonction des *éléments* de l'action qu'ils mettent en jeu. Ces éléments essentiels sont au nombre de quatre : l'agent, l'intention, le mouvement, et les effets. L'expérience du choix concerne surtout les deux premiers éléments (c'est une expérience de formation de l'intention), tandis que l'expérience de l'action volontaire engage les quatre. Dans l'expérience du choix, on peut reconnaître essentiellement les deux aspects que nous venons de mentionner :

* L'expérience du soi comme source (terme de Horgan 2011a), comme expérience du lien particulier entre l'agent et l'intention formée. Cet aspect du soi comme source a été étudié sous divers termes : (Horgan, Tienson, and Graham 2003) l'appellent l'aspect du « soi comme initiateur », (Bayne and Levy 2006) parlent du fait de se sentir « l'auteur » de l'action (*authorship*). Cette idée correspond également au « sens de l'initiation » chez (Pacherie 2008), et surtout à l'expérience de la causalité agentive pour (O'Connor 1995) ou de la causalité du sujet (*subject causation*) pour (Nida-Rümelin 2007).

* L'expérience des possibilités alternatives peut aussi être conçue comme expérience du lien entre l'agent et l'intention qu'il a formée (puisque c'est l'expérience que l'agent aurait pu former une *autre* intention à la place), ou bien avec (Holton 2006) comme expérience du lien non déterministe entre nos désirs et l'intention formée (le fait que notre intention formée ou notre choix n'est pas initialement *déterminé dans un sens ou un autre* par nos désirs et nos croyances). Cet aspect a également reçu différents noms, « capacité de faire autrement » chez (Nahmias et al. 2004), *voluntariness* chez (Horgan, Tienson, and Graham 2003), *core optionality* chez (Horgan 2011a).³⁵⁴

Dans l'expérience de l'action volontaire, on pourra faire l'expérience non seulement du lien agent-intention ou du lien désirs-intention, mais également du lien avec les deux autres éléments que sont le mouvement et les conséquences. En particulier, on trouvera :

* L'expérience de la finalité est une expérience du lien entre intention et mouvement, à savoir que le mouvement est produit avec une certaine intention qui lui correspond. Cette expérience, que (Horgan, Tienson, and Graham 2003) appellent *purposiveness*, peut être constituée de plusieurs niveaux : je peux par exemple avoir l'expérience que mon mouvement de bras correspond à mon intention de lever le bras pour faire stopper le bus, pour monter dans le bus, pour aller à la bibliothèque, pour consulter un ouvrage, etc. L'expérience de la finalité peut donc être parfois très riche dans certaines actions volontaires. Mais elle n'aura pas d'enjeu dans notre discussion. En effet, l'expérience d'une hiérarchie de finalités auxquelles correspond telle action n'est pas susceptible de produire un conflit avec l'idée de déterminisme. On peut sans aucune difficulté concevoir des systèmes robotiques totalement déterministes capables de calculer des moyens en vue de fins.

* L'expérience de la causalité mentale peut aussi être conçue comme une expérience du lien entre intention (mentale) et mouvement (physique), mais c'est cette fois l'expérience qu'il y a entre les deux un lien *causal* : le mental cause le physique (Voir notamment Pacherie 2008, sec. 6.1; Bayne and Levy 2006, sec. 2). Nous verrons plus

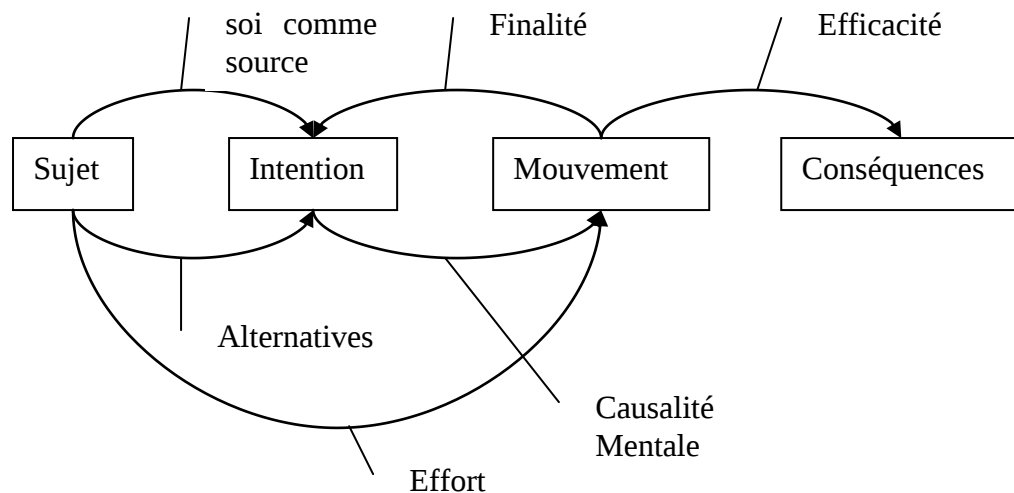
³⁵⁴ On pourrait également le ranger dans la catégorie générale du « sens du contrôle », selon l'expression de (Pacherie 2008), dans la mesure où le « contrôle des alternatives » est clairement un sens possible du contrôle de l'action.

loin qu'il n'est pas évident que nous ayons véritablement une expérience de nos actions comme *causées par* nos intentions (plutôt que comme causées par nous *en vue de* telle intention). Dans ce cas, il faudrait abandonner l'idée d'une expérience de la causalité mentale au sens étroit. On pourrait préserver l'idée d'une phénoménologie de la causalité mentale dans un sens plus large, à savoir dans le sens où le « mental » inclut non seulement les événements mentaux mais aussi le *sujet* lui-même. Si notre expérience de la causalité mentale est une expérience de nos actions comme causées *par nous* (comme agents), alors elle est équivalente à l'expérience de la causalité agentive dont nous avons parlé plus haut sous l'intitulé « expérience du soi comme source ». On verra les problèmes que peut poser notre expérience de la causalité agentive pour le déterminisme, mais il est clair que si l'on s'en tient au sens étroit de la causalité mentale (la causalité d'un mouvement physique par un événement mental comme l'intention), il n'y a pas le début d'un conflit possible avec le déterminisme : les intentions peuvent tout à fait causer de manière *déterministe* nos mouvements. Et à vrai dire, il vaut mieux (pour être libres) que nos intentions soient immanquablement suivies du mouvement correspondant ; autrement, on aurait un cas d'échec de l'action, une intention frustrée, et non pas un exercice de la liberté. C'est sur ce genre de considérations que se sont appuyés certains compatibilistes traditionnels, tout spécialement (Hobart 1934), pour argumenter que le libre arbitre, loin d'être incompatible avec le déterminisme, *requerrait* au contraire le déterminisme. A mon sens, ce type d'argument repose simplement sur une confusion entre deux moments de la phénoménologie de l'action : le moment qui *précède* la formation de l'intention (le moment du choix) et le moment qui *suit* la formation de l'intention (le moment de l'application de l'intention). Il est clair que notre phénoménologie agentive entre en conflit avec le déterminisme essentiellement à cause du premier moment, et non pas du second : si notre liberté fait apparaître de l'incertitude, c'est au moment où elle se décide, et non pas une fois qu'elle s'est décidée. La causalité mentale n'entrera donc pas directement en ligne de compte pour notre discussion à venir.

* L'expérience de l'efficacité est une expérience du lien entre un mouvement volontaire et ses conséquences (voulues) ; c'est l'expérience du fait que les conséquences extérieures ont été efficacement causées par nos mouvements. Cette expérience, évoquée dans (Pacherie 2008, sec. 6.1), a été étudiée pour elle-même par (Siegel 2008). Lorsque nous faisons l'expérience que nos mouvements non seulement ont causé le bon effet, mais que nous aurions pu causer le bon effet par d'autres mouvements dans des conditions différentes, l'expérience de l'efficacité peut être appelée une « expérience du contrôle ». Je préfère cependant éviter la terminologie de l'expérience du contrôle (adoptée notamment par Pacherie 2008, sec. 6.3), à cause de la grande variété de concepts de « contrôle » : le concept dont il s'agit ici est celui du « contrôle de guidage », mais il y a aussi un concept de « contrôle des alternatives » qui fait l'objet d'une expérience phénoménologique à part, comme nous l'avons vu. Là encore, il est évident que l'expérience de l'efficacité ne présente pas le début d'un conflit possible avec l'idée de déterminisme : l'expérience de l'efficacité n'est en fait qu'une partie de notre expérience de la causalité dans le monde extérieur, et cette expérience pourrait fort bien être déterministe. Si la liberté de l'agent introduit un indéterminisme (ou une apparence d'indéterminisme), ce doit être au moment où il intervient (par son choix ou son action) ; ça ne peut pas être dans la suite des conséquences qui se produisent une fois que toute intervention de sa part est finie.

* L'expérience de l'effort, enfin, est plus difficile à classer. On pourrait dire qu'il s'agit d'une expérience du rapport directement entre le mouvement volontaire et le sujet. C'est un type de phénoménologie agentive qui a été fréquemment étudié dans la tradition – on pense notamment à James ou à Maine de Biran. Les phénoménologues actuels de l'action n'ont pas totalement abandonné ce type d'études (Voir notamment Bayne and Levy 2006, sec. 4). Mais là encore, il est probable que cet aspect de notre expérience agentive n'est pas central pour la discussion du libre arbitre. Il ne semble pas que la notion d'effort en tant que telle puisse générer un conflit avec l'idée de déterminisme.³⁵⁵

On pourrait résumer l'ensemble des aspects de la phénoménologie agentive par le schéma suivant :



Les quatre derniers aspects que nous venons de voir sont propres à l'expérience de l'action volontaire par opposition à l'expérience du choix, parce qu'un simple événement de choix ou de décision n'est qu'une action mentale, qui n'engage ni mouvement physique ni conséquences extérieures. En revanche, on peut se demander si les deux premiers aspects, l'expérience des alternatives et l'expérience du soi comme source sont présents seulement dans l'expérience du choix ou s'ils sont présents également comme composantes de l'expérience de l'action volontaire, y compris lorsque celle-ci n'est pas précédée d'un processus délibératif et d'un choix explicites. Pour répondre de manière satisfaisante à cette question, il faudra entrer plus en détail dans l'analyse de ces deux aspects. Pour l'instant, notons simplement que ces deux aspects de la phénoménologie agentive sont assez fréquemment attribués dans la littérature à l'expérience de l'action volontaire et pas seulement à l'expérience du choix. C'est vrai de l'expérience des alternatives : l'idée est alors que nous aurions, même dans les mouvements que nous faisons sans aucune hésitation, l'expérience que nous *aurions pu* ne pas faire ce mouvement. C'est encore plus clair pour l'expérience du soi comme source dans sa version de causalité agentive : chez (Nida-Rümelin 2007) en particulier, l'expérience de la causalité agentive est d'abord l'expérience du fait qu'une substance est source d'un de ses mouvements, que ce mouvement ait requis ou non une forme de choix ou de décision. Il n'en reste pas moins que ces deux aspects de l'expérience agentive sont beaucoup plus nets et saillants dans l'expérience du choix que dans l'expérience de

³⁵⁵ Nous verrons cependant au chapitre 6, sec. II.4. que l'expérience de l'effort peut être étroitement liée à la notion de « pouvoir causal », qui constitue un ingrédient important d'une des théories libertariennes, à savoir la théorie de la causalité agentive.

l'action volontaire. Puisque ce sont précisément ces deux aspects qui sont importants pour la question du libre arbitre, on ne sera donc pas étonné de voir l'expérience du choix prendre une place importante dans nos analyses à venir, plutôt que l'expérience de l'action à proprement parler.³⁵⁶

IV.2.c. L'expérience de la Source et des Possibilités Alternatives dans le Choix

Les auteurs libertariens ont souvent appuyé leur conception du libre arbitre sur une analyse phénoménologique en première personne qui rapportait leur propre introspection. C.A. Campbell, par exemple, dit ceci :

Chacun est obligé de faire l'expérience introspective pour lui-même ; mais je peux peut-être m'aventurer à rapporter [...] que je ne peux pas m'empêcher de croire qu'il m'appartient, ici et maintenant, de manière absolue, d'adopter l'une ou l'autre de deux possibilités ouvertes. (Campbell 1951, 463)

De même, Keith Lehrer et John Searle rapportent leur propre introspection et en appellent de manière confiante à une introspection similaire chez tout un chacun :

[L'expérience des alternatives] décrit de manière adéquate ce que je trouve par introspection, et je ne peux pas croire que les autres ne trouvent pas la même chose. (Lehrer 1960, 150)

Réfléchissez très attentivement au caractère des expériences que vous avez lorsque vous vous engagez dans des actions humaines ordinaires. Vous sentirez que la possibilité d'actions alternatives est intégrée dans ces expériences. (Searle 1984, 95)

Horgan et Timmons (qui par ailleurs ne négligent d'argumenter) estiment que les deux aspects de phénoménologie importants sont tout simplement évidents :

En ce qui nous concerne, nous pensons qu'il est *tout simplement évident* introspectivement que les actions sont normalement vécues comme émanant de soi-même comme source, et également qu'elles sont normalement vécues comme étant telles que l'on aurait pu faire autrement dans les circonstances actuelles. (Horgan and Timmons 2011, 188)

Le problème de ces déclarations confiantes, c'est que certains compatibilistes au moins ont rapporté des expériences introspectives qui ne concordent pas avec les descriptions phénoménologiques des incompatibilistes. Il nous faudra donc tâcher d'argumenter en faveur de telle ou telle description phénoménologique, notamment à l'aide de la méthode différentielle présentée plus haut.

Commençons par l'aspect du soi comme source. Avons-nous réellement une telle phénoménologie, et si oui, comment la décrire précisément ? On peut poser la question d'une part pour l'expérience de l'action volontaire, et d'autre part pour l'expérience du choix.

³⁵⁶ Plusieurs auteurs ont souligné le caractère central de l'expérience du choix, en particulier (Browning 1964; Holton 2006). Pour Browning, l'expérience du libre arbitre n'est même rien d'autre que l'expérience du choix. En revanche, ce parti pris s'éloigne des auteurs qui donnent plus d'importance à l'expérience de la causalité agentive des mouvements corporels, notamment (Nida-Rümelin 2007).

Dans l'action volontaire, on peut mettre en évidence une expérience du soi comme source du mouvement, en considérant ce que serait l'expérience contraire. Si notre action volontaire n'était pas vécue comme étant produite ou causée par nous, comment serait-elle vécue ? Une possibilité serait que nous avons l'expérience d'une volition (de lever le bras, par exemple), suivie d'une expérience (perceptive) du mouvement concordant, sans aucune autre expérience particulière que celle de la concordance entre volition et mouvement. Ce serait une *phénoménologie du mouvement corporel fortuitement approprié*, pour reprendre l'expression de (Horgan 2011a, 78). Est-ce là notre phénoménologie habituelle de l'action volontaire ? Chacun peut se convaincre que notre expérience de l'action volontaire n'est *pas* celle du mouvement corporel fortuitement approprié, simplement en considérant que nous pouvons *aussi* avoir parfois cette autre phénoménologie, et que c'est une expérience très différente de celle de l'action. Par exemple, un tireur peut avoir l'expérience suivante : il se concentre longuement sur sa cible, ajuste son arme, prépare son doigt sur la gâchette, forme la volition de tirer à un moment ou à un autre, ... puis soudainement il voit son doigt bouger et le coup part tout seul ; cette expérience n'est clairement celle d'appuyer volontairement sur la gâchette. Si l'expérience de l'action n'est pas celle d'un rapport fortuit entre volition et mouvement, ce pourrait être l'expérience d'un lien *causal* entre volition et mouvement. Peut-être avons-nous l'expérience que notre volition (comme événement mental) cause notre mouvement sans intervention de notre part. C'est ce que (Horgan 2011a, 79) appelle la *phénoménologie de la causalité-événementielle psychologique du mouvement corporel*. Là encore, on peut se convaincre rapidement que notre expérience normale de l'action n'est *pas* ce genre d'expérience, que nous pouvons par ailleurs avoir dans d'autres circonstances. On peut, par exemple, faire l'expérience qu'un mouvement de tremblement est causé par un événement mental de peur (voir Horgan, *ibid.*). Certes un événement mental de peur n'est pas une volition, mais en réutilisant les exemples célèbres de séries causales déviantes, (Bayne and Levy 2006, 52) ont montré qu'on pouvait avoir une phénoménologie de la causalité volition-mouvement qui n'est *pas* notre phénoménologie de l'action volontaire : un alpiniste peut par exemple faire l'expérience d'une volition de relâcher la corde par laquelle il retient un coéquipier en danger, et faire également l'expérience que cette volition le terrifie et que sa peur cause un tremblement qui lui fait relâcher la corde (voir Davidson 1973). Dans une action volontaire normale, notre phénoménologie n'est donc pas celle d'un mouvement *fortuitement* adapté à l'intention, ni celle d'un mouvement adapté à l'intention *par lien causal événementiel*. Qu'est-elle alors ? Elle est l'expérience que *je* produis (ou cause) le mouvement *d'après* mon intention. Telle est l'expérience de la *causalité agentive* dans le cas du mouvement volontaire : c'est l'expérience d'un type de causalité qui a lieu non pas entre un événement A et un événement B, mais entre un *agent* et un mouvement physique.

Ce qui est délicat avec l'expérience de la causalité agentive du mouvement, c'est qu'il n'est pas absolument évident qu'elle soit logiquement incompatible avec une causalité événementielle sous-jacente. Certes, nous faisons l'expérience de *nous-mêmes* comme causes du mouvement au sens agentif, et nous ne *faisons pas* l'expérience que notre volition est une cause du mouvement au sens événementiel, mais cela ne veut pas dire que nous fassions l'expérience que notre volition *n'est pas* cause événementielle du mouvement. Et par conséquent, les auteurs qui défendent la causalité événementielle ont souvent cherché à montrer que la *terminologie* de la causalité agentive était légitime, mais qu'elle pouvait être *interprétée* comme une manière particulière de décrire de la causalité événementielle sous-jacente, à savoir de la causalité par des événements étroitement associés à l'agent (voir notamment Balaguer 2010, 87). Or si l'on réduit ainsi la causalité agentive à la causalité événementielle, il n'est pas clair du tout que la causalité « agentive » du *mouvement* présente le moindre problème pour le déterminisme puisque le lien causal entre volition et mouvement a toutes les chances d'être un lien causal déterministe, dans toutes les conceptions de l'action

libre. Le lieu éventuel de l'indéterminisme pour qu'il y ait liberté se situe en amont de la volition, ou plutôt au moment même où celle-ci est formée. Si donc la causalité agentive pose problème pour le déterminisme, ce sera plutôt la causalité agentive *du choix* que celle du mouvement volontaire à proprement parler.

Au moment du choix, on peut retrouver le même genre d'éléments phénoménologiques. C'est-à-dire que le choix n'est *pas* expérimenté comme se produisant de manière fortuite à la suite d'une délibération (*phénoménologie du choix fortuit*), et il n'est *pas* expérimenté non plus comme étant l'effet causal de nos raisons et de nos croyances (*phénoménologie de la causalité-événementielle psychologique du choix*). Dans le cas normal, notre choix est expérimenté comme étant produit ou causé *par moi-même* comme sujet, et ce *en vue* de telle ou telle raison. Tim O'Connor, un des principaux défenseurs contemporains de la causalité agentive, soutient cette conception précisément à partir d'une telle analyse phénoménologique du choix (elle-même étayée par un argument différentiel)³⁵⁷ :

[La causalité agentive] est attrayante parce qu'elle capture la manière dont nous expérimentons ordinairement notre propre activité. Il ne me semble (au moins ordinairement) que je suis causé à agir par les raisons qui favorisent de le faire ; il semble, plutôt, que *je produis ma propre décision en vue de ces raisons*.
(O'Connor 1995, 196)

Le fait que l'expérience de la causalité agentive apparaisse également dans le choix est peut-être une conséquence du fait que, comme le souligne (Holton 2006, 3), le choix est lui-même une sorte d'action. Mais c'est un type d'action très particulier qui donne lieu à une expérience particulièrement intéressante du soi comme source, comme on va le voir bientôt. Avant cela, essayons de nous assurer de la solidité du résultat présent : est-il vraiment évident que nous avons une expérience d'être, comme sujet, la source ou la cause de nos choix ? Nous avons apporté en faveur de cette conclusion un argument différentiel, mais cet argument est-il absolument irréfutable ? Un auteur ne pourrait-il pas embrasser tout bonnement l'une de nos deux phénoménologies différentielles, la phénoménologie du choix fortuit ou la phénoménologie de la causalité-événementielle psychologique du choix ? C'est ce qu'a fait malheureusement Daniel Dennett, qui décrit ainsi son expérience du choix :

Nous devons attendre pour voir ce que nous allons décider sur tel sujet, et lorsque nous décidons effectivement, cela vient exploser comme une bulle à la surface de notre conscience sans que l'on sache d'où cela vient. Nous n'assistons pas à quelque chose que nous *faisons* ; nous assistons à son *arrivée* (Dennett 1984a, 78)

Que penser d'une telle description phénoménologique, que Dennett s'attribue non seulement à lui-même, mais à un « nous » qui semble inclure tout sujet humain normal pour toute décision normale ? Il me semble que bien peu de gens accepteraient pour eux-mêmes la description phénoménologique que donne Dennett. Ou en tout cas, il me semble que peu de gens l'accepteraient pour leurs décisions normales. Bien sûr, nous *pouvons* aussi avoir cette phénoménologie qu'on pourrait appeler la *phénoménologie de l'attente du tirage au sort* (une forme de *phénoménologie du choix fortuit*), mais ce n'est *pas* la phénoménologie habituelle de la délibération conclue par un choix volontaire, et il m'est difficile d'imaginer que Dennett puisse totalement manquer de cette seconde phénoménologie. On est ici très proche des limites de possibilité de l'argumentation en phénoménologie.

³⁵⁷ Voir également, pour le même argument, (Horgan, Tienson, and Graham 2003, 225).

Le dernier argument qu'on pourrait apporter pour montrer, contre Dennett, la normalité de la phénoménologie du soi comme source, est l'argument clinique. Non seulement il semble que la majorité des gens refuseraient la description phénoménologique de l'action ou du choix fortuits (sans soi comme source), mais en outre, on constate que certains patients atteints de pathologies psychologiques rapportent précisément ce type de phénoménologie. Ainsi, dans le fameux « syndrome de la main anarchique », les patients font l'expérience de leurs mouvements comme n'étant pas « les leurs » ou n'étant pas produits par eux. Par exemple, la main anarchique s'efforce de déboutonner la chemise tandis que le patient, avec sa main valide s'efforce de l'en empêcher ; ou bien la main anarchique prend de la nourriture dans une assiette contre la volonté du sujet. La manière dont ces expériences sont décrites est étroitement associée au vocabulaire du soi ou du sujet : l'action n'est plus mienne, je n'en suis plus l'auteur ou l'origine. Ce genre de pathologie a divers intérêts théoriques en théorie de l'action. Mais en phénoménologie du libre arbitre, elle est fréquemment utilisée (par exemple, par Bayne and Levy 2006, 56–57) pour mettre en évidence le caractère nécessaire d'une phénoménologie du soi comme auteur (*authorship*) ou du soi comme source, pour qu'il y ait une phénoménologie agentive normale. De fait, la manière la plus évidente de décrire ces pathologies, c'est de dire qu'il leur *manque* la phénoménologie du soi comme source. Pour cette raison, on peut raisonnablement estimer que la phénoménologie du choix que donne Dennett ne correspond certainement pas à notre phénoménologie normale du choix.

Nous avons donc mis en évidence l'existence d'une phénoménologie du soi comme source dans l'expérience du choix. Reste à décrire plus précisément cet aspect. Si l'expérience du choix est plus intéressante que l'expérience du mouvement volontaire, c'est parce dans la seconde, il semble exister, en plus de l'agent comme cause, une cause événementielle suffisante pour le mouvement, à savoir la volition. Dans le cas du choix, il est beaucoup moins clair qu'il y ait une telle cause événementielle suffisante à notre disposition : en effet, les antécédents évidents du choix (lors du processus délibératif) sont nos croyances et nos désirs (ou nos raisons), mais comme le fait remarquer (Holton 2006, 4), notre expérience normale du choix est telle que « le choix n'est pas déterminé par nos croyances et désirs antécédents ». Cela est-il toujours vrai ? N'y a-t-il pas certains choix que nous faisons de manière absolument confiante en considération de raisons qui nous paraissent l'emporter absolument sur toute autre raison ? Ce genre de *choix confiant* existe en effet. Aussi, ce n'est pas toute expérience de choix qui nous intéressera particulièrement, mais plutôt l'expérience de ce que (Nahmias et al. 2004, 168) appellent des *close calls* (cas serrés ou cas limites), et (Balaguer 2010, 21 sqq) des *torn decisions* (décisions perplexes ou tiraillées). Il y a incontestablement certains choix dans lesquels nous faisons l'expérience phénoménologique que nos désirs (et nos croyances) sont tout à fait insuffisants pour déterminer telle ou telle décision. Je les appellerai personnellement des *choix difficiles*.

Dans l'expérience du choix difficile, par opposition au choix confiant, notre phénoménologie est donc non seulement celle d'être la source ou la cause de l'acte de choix, mais *en outre* l'expérience qu'il n'y a pas dans ma psychologie (dans mes raisons et mes croyances) de cause antécédente suffisante de mon choix. On peut aller plus loin et montrer que, dans le choix difficile, je fais l'expérience d'un choix qui n'a *aucune* cause antécédente suffisante, ni psychologique ni d'un autre ordre, c'est-à-dire l'expérience d'un choix qui est le *point de départ* d'une série causale, ou encore l'expérience d'être moi-même *l'initiateur* du choix (et de ses conséquences). Comment faire ce passage d'une absence de cause psychologique suffisante à une expérience de l'initiation ? Ce passage n'est-il pas le genre de généralisation illégitime que critiquait Spinoza : nous ne voyons pas (dans notre psychologie) les causes de nos actions, et nous en déduisons illégitimement qu'elles n'ont aucune cause. Mais ce qui nous intéresse ici n'est pas de savoir si tel ou tel raisonnement est légitime ; nous

n'étudions pas un raisonnement philosophique mais seulement notre *expérience* agentive, qu'elle soit vraie ou fausse. L'objection pertinente ici consisterait alors à dire que nous ne *pouvons pas* avoir une expérience agentive avec ce contenu : nous ne pourrions avoir l'expérience que *absolument rien* ne cause notre choix de manière suffisante. Notre expérience serait alors telle qu'elle ne représente pas de cause, et c'est seulement un raisonnement postérieur qui ajouterait qu'il n'y a pas de cause. A cette objection, une première réaction serait de demander tout simplement « pourquoi » : pourquoi une expérience phénoménale ne pourrait-elle pas avoir un contenu du type « il n'y a aucune cause de notre choix » ? Certes, il est fondamental de distinguer entre une expérience qui a ce contenu et une absence d'expérience de la cause de notre choix. Cela ne fait pas de doute. Mais une fois qu'on a fait cette distinction, pourquoi serait-il *impossible* que nous ayons, en outre d'une absence d'expérience de la cause, une expérience de l'absence de cause ? N'est-il pas possible d'avoir l'expérience phénoménale que « il n'y a rien dans le frigo » ? Je ne vois rien dans le contenu « il n'y a pas de cause de mon choix » qui empêche ce contenu de faire l'objet d'une expérience phénoménale. Mais il est possible de renforcer cette première réponse en montrant *comment* notre expérience phénoménale pourrait avoir un tel contenu général.

L'expérience du soi comme initiateur des choix difficiles, c'est-à-dire l'expérience de ces choix comme n'ayant pas cause antécédente suffisante, repose à mon sens sur deux autres expériences plus fondamentales : d'une part, l'expérience de la non-détermination par les raisons (et les croyances), et d'autre part l'expérience de la causalité mentale dans un sens particulier. Pour la première, l'expérience de la non-détermination par les raisons (et les croyances), nous venons de voir qu'elle était caractéristique de certains de nos choix, à savoir précisément nos choix difficiles. Quant à l'expérience de la causalité mentale, j'entends ici l'expérience phénoménale que ma décision ou mon choix est un événement strictement mental quant à son origine causale, c'est-à-dire l'expérience que mon choix repose sur mes croyances, mes désirs, mes raisons, ou ce genre de choses, *et sur rien d'autre* (d'extérieur au mental). Initialement, l'expérience de la causalité mentale n'est pas une expérience dont le contenu est quantifié (il n'y a aucune cause non-mentale), mais plutôt une expérience sur la nature de cet événement (un événement strictement mental), et c'est cette expérience sur la nature du choix qui entraîne un contenu dont Spinoza ou d'autres ont estimé qu'il était « trop universel » pour être expérimentable. Si nous faisons l'expérience que nos choix sont causés de manière strictement mentale (pas de cause non mentale) et que nos choix difficiles ne sont *pas* déterminés par nos désirs et croyances (pas de cause suffisante mentale) que nous sommes bien placés pour connaître, alors, nous pouvons faire l'expérience que nos choix difficiles n'ont pas du tout de cause antécédente suffisante, c'est-à-dire que ces choix (ou le sujet lui-même comme cause agentive) est au point de départ d'une nouvelle série causale.

Avons-nous réellement une telle expérience de la causalité (exclusivement) mentale des choix, et sa conséquence, une expérience de l'initiation absolue pour les choix difficiles ? Il y a plusieurs remarques à faire pour établir ce point.

Tout d'abord, il faut comprendre que la causalité mentale telle que je l'entends ici est compatible avec l'idée que le mental est entièrement réalisé dans le physique (dans les états du cerveau). En ce sens, notre expérience de la causalité mentale permet que nos choix confiants aient des causes physiques dans le sens précis où ces causes physiques seraient les réalisateurs neuronaux de nos désirs et croyances qui ont causé ce choix. L'idée de causalité mentale suppose simplement qu'il n'y ait aucune cause physique « court-circuitant » le mental, pour reprendre l'expression de Nahmias (voir plus haut p.295). A vrai dire, l'expérience de la causalité mentale n'est rien d'autre que ceci : l'expérience qu'il n'y a pas de cause court-circuitante à l'œuvre dans notre choix. L'introspection peut-elle réellement nous donner une telle information ? Certains auteurs en doutent, inspirés parfois par Spinoza. Mais notre

question présente est seulement de savoir si notre expérience peut donner une telle *impression* – que cette impression constitue ensuite une illusion ou une source d'information. Autrement dit, notre expérience du choix contient-elle un élément qui serait faux (ou illusoire) si on nous révélait qu'il y a une cause court-circuitante de nos choix ? Il me semble qu'une fois qu'on a posé cette question de manière claire, la réponse qui s'impose est oui : si on nous révélait l'existence de causes court-circuitantes de nos choix, cela s'opposerait à « quelque chose » dans notre expérience normale du choix. Cette expérience est donc l'expérience qu'il *n'y a pas* de cause court-circuitante, c'est-à-dire l'expérience d'une causalité strictement mentale.

Deuxièmement, on peut essayer d'établir l'existence d'une phénoménologie de la causalité mentale par un argument différentiel, en imaginant une *phénoménologie spinoziste*, c'est-à-dire une *phénoménologie de la cause ignorée*. Si mon expérience agentive était telle que Spinoza l'imagine, alors au moment de faire un choix je devrais faire l'expérience que mon choix a une cause, sans expérimenter quelle est cette cause. J'aurais donc une phénoménologie d'ignorance ou de manque d'information, comparable au cas suivant : je vois devant moi un homme debout, et subitement la tête de cet homme fait un mouvement vers le bas, comme s'il avait été frappé d'un grand coup par un bâton, mais je ne vois aucun bâton au-dessus de lui. L'homme se gratte la tête, se retourne, cherche, et ne voit rien (ni moi non plus). Si l'on accepte que les relations causales puissent être le contenu de nos vécus perceptifs, on pourrait décrire cette expérience de la manière suivante : je fais l'expérience que le mouvement de la tête de cet homme a une cause, mais que j'ignore cette cause. C'est un mouvement causé d'une manière que j'ignore. Notez que cette ignorance attise normalement ma curiosité. Lorsque je prends une décision, il me semble évident que je n'ai pas une expérience d'un événement causé par une cause ignorée (laquelle ignorance pourrait éveiller mon étonnement et ma curiosité, et susciter une *recherche* de la cause). J'ai au contraire l'expérience de très bien connaître les tenants et les aboutissants de cet événement, qui se situent dans ma vie mentale. Bien sûr, je *pourrais* me tromper, et il *pourrait* y avoir des causes profondes de mon choix qui m'échappent ; mais ces causes profondes, si un jour on me les révèle, *m'étonneront*, ce qui suffit à montrer que mon *expérience* initiale était l'expérience que ces causes sont absentes, et non pas l'expérience que j'ignore leur nature.

Troisièmement, l'expérience de la causalité mentale n'est ici mise en évidence que pour montrer comment nous pouvons avoir une expérience de l'initiation absolue, ou du soi comme source ultime. Mais certains lecteurs estimeront peut-être que l'existence d'une expérience de source ultime (pour les choix difficiles) est suffisamment évidente en elle-même, ou plus évidente que l'expérience de la causalité mentale. On pourrait en effet essayer de mettre en évidence cette expérience de source ultime par un argument différentiel direct. Lorsque je prends des décisions confiantes, on peut dire qu'en un sens je fais une expérience de « continuité », l'expérience de « poursuivre un scénario » qui a précédé et qui va se continuer. Par exemple, je décide de tourner à droite en fonction de la carte routière parce que l'itinéraire que je suis depuis un quart d'heure et que je vais suivre encore quelque temps suppose de tourner maintenant à droite. La décision est confiante parce qu'elle entre dans une logique ou dans un scénario où je me trouve déjà et dont je n'ai pas de raison de dévier. Inversement, lorsque je prends une décision difficile, je n'ai pas cette expérience de poursuivre un scénario déjà initié ou une logique déjà en œuvre. Je choisis par exemple d'aller à la piscine plutôt qu'à la bibliothèque sans avoir de raisons plus fortes en faveur de la piscine.

Quatrièmement, l'existence d'une phénoménologie de la source ultime peut enfin être mise en évidence de manière indirecte par les *conséquences* qu'elle a, et en particulier par ses conséquences pour la phénoménologie des possibilités alternatives. Nous allons parler dans un instant de cette phénoménologie des possibilités alternatives comme deuxième aspect important de la phénoménologie du libre arbitre, mais il est important de noter que les deux aspects ne sont pas absolument indépendants l'un de l'autre. Je souscris en effet à l'idée,

défendue par plusieurs auteurs, que l'expérience des possibilités alternatives *repose* sur (ou provient de) notre expérience du soi comme source :

Le sentiment que « je fais arriver ceci » entraîne avec lui le sentiment que « je pourrais faire autrement ». (Searle 1984, 95)

Lorsque nous expérimentons notre comportement comme émanant de nous-même comme de sa source, nous nous expérimentons comme capable de nous retenir d'agir ainsi. [...] La phénoménologie de l'agent est d'être un soi-source à part entière du comportement, et le fait d'être source « à part entière » inclut ici le fait d'être tel que l'on aurait pu agir autrement. (Horgan 2011a, 79–82)

Pourquoi la phénoménologie du soi comme source implique-t-elle une phénoménologie des possibilités alternatives ? C'est particulièrement clair d'après ce qu'on vient de voir pour l'expérience des choix difficiles. Pour le choix difficile, dire que je suis la source véritable de mon choix, c'est dire qu'il n'avait pas de cause suffisante en dehors de la sphère du « soi », et puisque le choix est difficile, il n'avait pas non plus de cause suffisante dans mes raisons ou dans mes désirs ; c'est donc dire qu'il n'avait aucune cause suffisante dans les événements antécédents. Si tel est le cas, alors au moment où le choix a été posé, rien n'empêchait qu'un autre choix soit posé à la place, et donc l'auteur du choix aurait pu être l'auteur d'un autre choix. La phénoménologie des possibilités alternatives peut donc se « déduire » de la phénoménologie de la source ultime. Si j'insiste sur le lien entre phénoménologie de la source et phénoménologie des possibilités alternatives, c'est qu'à mon sens ce lien renforce l'une et l'autre. En particulier, les arguments à venir en faveur d'une phénoménologie des possibilités alternatives renforceront la présomption d'une phénoménologie de la source ultime. Or à mes yeux l'existence de la première est sans doute plus incontestable encore que l'existence de la seconde.

Passons donc à présent à la phénoménologie des possibilités alternatives. Là encore, on pourrait s'intéresser à l'expérience des possibilités alternatives *pour l'action volontaire*, mais il sera beaucoup plus fructueux de s'intéresser directement à l'expérience du *choix*, et en particulier du choix difficile. En effet, si l'on peut douter que l'action volontaire suppose toujours la considération de possibilités alternatives, en revanche le choix est par définition un processus de sélection, qui inclut donc analytiquement l'idée d'options alternatives entre lesquelles le choix va trancher. L'expérience du choix est donc nécessairement une expérience qui inclut des alternatives ; reste à savoir si cette expérience présente les alternatives comme *possibles*, et *en quel sens* ces alternatives sont présentées comme possibles. On verra en effet que peu de gens contestent le premier point (l'existence d'une phénoménologie des alternatives *possibles*) mais que les compatibilistes ont tendance à nier que nous ayons une expérience des alternatives comme *absolument possibles* ou *possibles dans les circonstances actuelles*. Plusieurs auteurs, libertariens ou non, ont défendu introspectivement l'existence d'une expérience des alternatives *absolument possibles*. Par exemple :

(Au moins ordinairement) [...] il me semble que je produis ma décision en vue de ces raisons, et que j'aurais pu, dans un sens inconditionnel, décider différemment. (O'Connor 1995, 196)

Les actions sont normalement vécues [...] comme étant telles que l'on aurait pu faire autrement dans les circonstances actuelles. (Horgan and Timmons 2011, 188)

Et Searle précise que notre sentiment de pouvoir faire autrement est tel que

nous pourrions faire autrement juste ici et maintenant, c'est-à-dire, toutes les autres conditions restant les mêmes. (Searle 1984, 95)

Mais encore une fois, il ne suffit pas de rapporter sa propre introspection pour trancher les désaccords phénoménologiques qui sont apparus dans le débat sur le libre arbitre.

Commençons donc par la question de savoir si les options du choix sont expérimentées ou non comme des alternatives *possibles*. La réponse ici n'est pas vraiment contestée. Mais si l'on veut s'en convaincre, on peut utiliser deux types d'arguments. Premièrement, un argument conceptuel. Le choix est une décision qui fait suite à une délibération. Or une délibération entre plusieurs options suppose analytiquement que je tiens ces options pour possibles. On le sait depuis Aristote : « nous délibérons sur les choses qui dépendent de nous et que nous pouvons réaliser » (Aristote 1959, bk. III1112a30). Il n'y a donc pas de délibération sans expérience de la possibilité des options, et dans la mesure où le choix est le terme ultime de la délibération, l'expérience de la possibilité doit durer jusqu'à l'instant du choix, ce dernier étant justement, par définition, l'action par lequel une des deux possibilités est éliminée. Par conséquent, à l'instant où nous choisissons, il faut bien que nous ayons l'expérience des deux alternatives, celle que je choisis et celle que je rejette, comme *possibles*.

Si cet argument conceptuel ne suffit pas, on peut l'étayer par un argument différentiel phénoménologique. Que serait une expérience des alternatives *sans* expérience de leur possibilité ? Ce serait l'expérience de la délibération comme d'une procédure qui va nécessairement amener à un résultat déterminé, même si on ne le connaît pas encore ; la procédure de délibération ne servirait en fait qu'à *découvrir* un résultat déjà déterminé, l'autre option étant dès le départ fictive. On pourrait appeler cela la *phénoménologie de la procédure de découverte d'une information déterminée*. Nous connaissons bien ce type de phénoménologie dans d'autres circonstances. Par exemple, si l'on me demande « quel est le 18^{ème} mot de la Marseillaise », je suis incapable de le dire instantanément, mais je peux chanter intérieurement la Marseillaise en comptant les mots sur mes doigts ; j'aurai alors l'expérience d'une procédure que je conduis pour *découvrir* un résultat déjà déterminé. Mais le processus mental de délibération ne ressemble pas du tout à cela phénoménologiquement. Je n'ai pas l'expérience de lancer une procédure qui va me révéler quelle décision j'ai déjà prise ; j'ai l'expérience de lancer une procédure qui va me permettre de prendre une décision à la fin. Bien sûr, cette expérience *pourrait* être erronée ; il se pourrait par exemple que lorsqu'on délibère, on ne fait que chercher des justifications *a posteriori* à une décision qui est déjà prise dans notre subconscient. Mais les auteurs qui proposent ce genre de théories (qu'ils soient philosophes, comme Sartre, ou psychologues) prétendent démasquer par là nos *illusions* spontanées quant à la délibération. Autrement dit, ces théories elles-mêmes révèlent que notre *expérience* naturelle de la délibération (véridique ou erronée) ne ressemble pas à une procédure qui aboutirait simplement à découvrir quelle option est déjà déterminée. Au contraire, notre expérience naturelle de la délibération, et du choix qui la conclut, est une expérience des *deux* alternatives comme possibles.

Aucun n'auteur, à ma connaissance, n'a cherché à contester cette phénoménologie des alternatives possibles dans l'expérience du choix. Ce qui a fait l'objet de débat en revanche, c'est en quel sens ces alternatives sont possibles : sont-elles *absolument* possibles, ou possibles à certaines conditions ? Cette question est évidemment d'une grande importance pour la question du compatibilisme ou de l'incompatibilisme de notre phénoménologie agentive. En effet, si notre expérience des alternatives était simplement que l'option non choisie était possible *dans d'autres circonstances*, ou *si nous avions été dans un état mental*

différent, alors on obtiendrait une description compatible avec le déterminisme. De fait, on obtiendrait un concept de libre arbitre qui correspond à l'analyse conditionnaliste étudiée plus haut : si j'avais eu d'autres désirs (ou d'autres croyances), j'aurais pu faire un autre choix, former une autre volition ; mais étant donné mes désirs actuels (et mes croyances actuelles), j'étais causalement déterminé à faire le choix que j'ai fait. J'ai rapporté un peu plus haut les déclarations de différents auteurs qui s'attribuent une phénoménologie des alternatives comme possibles *absolument*, ou *dans un sens inconditionnel*, ou encore *toutes les autres conditions restant les mêmes*. Mais ces auteurs auraient pu se tromper dans leur enquête introspective, leur introspection aurait pu être biaisée. C'est ce que l'on peut craindre, en particulier, quand on découvre des déclarations exactement inverses, par exemple chez John Stuart Mill ou chez Adolf Grünbaum :

Cette conviction, qu'on l'appelle conscience ou seulement croyance, que notre volonté est libre, quelle est-elle ? De quoi sommes-nous convaincus ? On me dit que lorsque je décide d'agir ou de m'abstenir, je sens que j'aurais pu prendre la décision contraire. Je demande à ma conscience ce que je ressens, et je découvre qu'en fait je sens (ou suis convaincu) que j'aurais pu choisir l'autre option *si je l'avais préféré* ; mais pas que j'aurais pu choisir une option en préférant l'autre. [...] Lorsque nous pensons à nous-mêmes de manière hypothétique comme ayant pu agir autrement que nous l'avons fait, nous supposons toujours une différence dans les antécédents : nous nous représentons nous-mêmes comme ayant connu autre chose que ce que nous connaissions [...] ou comme ayant désiré quelque chose [...] plus ou moins que nous l'avons désiré. (Mill 1865, 504–505)

Examinons attentivement le contenu de ce sentiment qu'en certaine occasion nous aurions pu agir autrement que de la manière dont nous avons effectivement agi. Que trouvons-nous ? Le sentiment que nous avons nous informe-t-il que nous aurions pu agir autrement *exactement dans les mêmes conditions motivationnelles externes et internes* ? Non, répond le déterministe, ce sentiment révèle seulement que nous étions capables d'agir en accord avec notre désir le plus fort à ce moment, et que nous aurions pu en effet agir autrement si un motif différent avait prévalu à ce moment. (Grünbaum 1952, 672)

Il semble que nous soyons ici face à un désaccord introspectif irréductible : les uns font état d'une expérience d'alternatives absolues, les autres d'une expérience d'alternatives hypothétiques. Une possibilité est que les uns et les autres n'aient effectivement pas la même expérience. Cela tendrait à dire que le problème du libre arbitre apparaît pour certaines personnes et pas pour d'autres, en fonction de nos différentes expériences phénoménales. Je pense pourtant qu'un tel pessimisme n'est pas justifié, et qu'on peut non seulement estimer qu'il y a bien une expérience commune des alternatives, mais en outre qu'on peut raisonnablement déterminer quelle est cette expérience malgré ce désaccord introspectif. En effet, si l'on prête attention aux descriptions données ici par Grünbaum et Mill, on peut aisément s'apercevoir qu'ils se sont tout simplement trompés de cible : ils n'ont pas décrit la *bonne* expérience, et c'est pourquoi ils ne sont pas arrivés au bon contenu.

L'erreur de Mill peut être repérée en observant l'argument qu'il donne en faveur de sa description phénoménologique conditionnaliste. Son argument est en fait un argument différentiel : notre expérience, dit-il, doit être une expérience que « j'aurais pu choisir l'autre option *si je l'avais préférée* » parce qu'elle n'est *pas* une autre expérience, à savoir l'expérience que « j'aurais pu choisir une option *en préférant l'autre* ». Mais on peut remarquer que cet argument différentiel repose sur un faux dilemme : lorsque je choisis une

option A, il semble que pour Mill je la choisisse soit (i) en la préférant, soit (ii) en préférant l'autre. Mais n'est-il pas possible que je n'aie aucune préférence ? Ou plus précisément que je n'aie aucun désir *prévalent* ? Autrement dit, Mill ne considère que les situations où nous avons un désir prévalent : si nous préférons A, alors nécessairement nous choisissons A, et nous n'aurions pu choisir B qu'en ayant un état de désir différent. Ces situations de désir prévalent sont celles que nous avons appelées plus haut les situations de choix confiant. Et Mill a peut-être raison de dire que *dans les situations de choix confiant*, nous n'avons pas l'expérience que l'autre choix était *absolument* possible. Par exemple, si je fais le choix confiant de donner les 20 € de mon portefeuille plutôt que de me faire tuer par le bandit qui me demande mon argent, je ne fais peut-être pas l'expérience qu'en un sens absolu je pouvais faire autrement, parce qu'il serait absurde pour moi de faire autrement et que je ne suis pas assez fou pour cela. Mais tout cela n'a aucune incidence sur notre discussion présente puisque nous nous concentrons ici sur les *choix difficiles*, ceux pour lesquels je n'ai justement pas de préférence ou de désir prévalent. Dans un cas de choix difficile, je n'ai certes pas l'expérience que « je pourrais choisir une option en préférant l'autre », ce qui serait absurde, mais il n'y a pas à choisir entre cette description absurde et la description conditionnelle sur laquelle Mill se rabat. Dans un choix difficile, je choisis une option pour laquelle je n'ai pas de préférence décisive, et j'ai le sentiment que je pourrais choisir l'autre option pour laquelle je n'ai pas non plus de préférence décisive. Contrairement à ce que Mill semble penser, l'expérience des alternatives *absolument* possibles n'est pas l'expérience d'une possibilité d'agir *contre* ses préférences, mais plutôt d'une possibilité d'agir *en l'absence* de préférence (ou, en tout cas, en l'absence de préférence *décisive*). S'il avait porté son attention aux choix difficiles plutôt qu'aux choix qui s'ensuivent sereinement de nos préférences, il y a toutes les raisons de penser que Mill lui-même aurait pu constater sa propre phénoménologie des alternatives absolument possibles. Mill a donc simplement porté son introspection sur la mauvaise expérience.

L'erreur de Grünbaum est encore plus simple à repérer et plus révélatrice, car elle va nous permettre de faire voir l'argument positif en faveur d'une phénoménologie des alternatives absolument possibles. Dans l'extrait que j'ai cité plus haut, on peut remarquer que l'expérience est rapportée au passé, c'est l'expérience que « *j'aurais pu faire autrement* ». Ce point est explicité clairement par Grünbaum à un autre endroit : le fait phénoménologique qu'il cherche à décrire est « le fait que lorsque *nous regardons en arrière vers notre conduite passée*, nous avons fréquemment le sentiment très fort que nous aurions pu faire autrement » (Grünbaum 1952, 672). Mais l'expérience dont nous parlons ici n'est pas une conviction rétrospective. L'expérience du libre arbitre est l'expérience que nous faisons *au moment* où nous agissons, et plus précisément au moment où nous choisissons. C'est l'expérience que *maintenant* je pourrais faire autrement. Il ne faut donc pas confondre l'expérience rétrospective des alternatives et l'expérience présente des alternatives. À dire vrai, je doute que la description conditionnelle de Grünbaum soit véritablement adaptée pour notre expérience rétrospective, pour autant que cette expérience est une expérience rétrospective d'un choix difficile. Il y a tout lieu de penser que Grünbaum, comme Mill, a malencontreusement focalisé son attention sur les expériences (rétrospectives) de choix confiants, puisqu'il parle du « désir le plus fort ». Mais à supposer qu'il ait raison pour l'expérience rétrospective, cela n'aurait pas d'incidence sur notre discussion qui concerne l'expérience présente des alternatives. Or lorsqu'on prend conscience que notre expérience des alternatives est une expérience *présente*, on s'aperçoit que ce caractère même mène naturellement à une expérience des possibilités *absolues*.

L'argument positif en faveur des possibilités absolues est le suivant. Notre expérience des alternatives possibles est l'expérience que « je pourrais *maintenant* choisir autrement ». Ce n'est pas une expérience rétrospective ou prospective. Ce n'est pas non plus seulement

l'expérience que « je pourrais [choisir autrement maintenant] », mais bien l'expérience que « [je pourrais maintenant] choisir autrement » ; c'est-à-dire que le maintenant n'est pas seulement le temps de l'acte de choix alternatif, mais le temps où les deux alternatives sont possibles. Notre expérience n'est pas, par exemple, l'expérience de la possibilité passée d'un état de chose présente. Ce genre d'expérience existe : si je m'ennuie dans une salle de cinéma, je peux faire l'expérience que je pourrais être maintenant dans une autre salle, c'est-à-dire que je pouvais (dans le passé, au moment d'acheter mon billet) faire en sorte que maintenant je sois dans une autre salle. Le maintenant ne concerne alors que l'état de chose et pas la possibilité. Mais dans l'expérience du choix, nous faisons incontestablement l'expérience (à tort ou à raison) que la possibilité même des alternatives est une possibilité présente. Si les deux alternatives sont vécues comme possibles « maintenant », il est évident qu'elles sont aussi vécues comme possibles « ici ». Pour parler comme les linguistes, les alternatives sont vécues comme possibles « à l'index » du sujet qui fait l'expérience du choix, ou « dans son contexte ». Plus simplement, elles sont vécues comme possibles *pour lui, ici et maintenant*. Ce caractère « indexical » de l'expérience des alternatives est ce qui amène naturellement à la considérer comme une expérience des alternatives absolues et non pas conditionnelles. On peut remarquer cette transition logique chez Campbell ou Searle :

[Nous] pourrions faire autrement juste ici et maintenant, *c'est-à-dire*, toutes les autres conditions restant les mêmes. (Searle 1984, 95)

[Je] ne peux pas m'empêcher de croire qu'il m'appartient, *ici et maintenant, de manière absolue*, d'adopter l'une ou l'autre de deux possibilités ouvertes.
(Campbell 1951, 463)

Inversement, si notre expérience des alternatives était seulement l'expérience conditionnelle que « je pourrais choisir autrement *dans d'autres circonstances* », nous n'aurions donc pas l'expérience d'une alternative réellement possible. On peut ici comparer notre expérience du choix avec d'autres expériences d'alternatives conditionnellement possibles : je suis poursuivi par des ennemis qui veulent me tuer et je me retrouve acculé à un précipice ; j'ai à ce moment là la pensée et le sentiment que *si j'avais pris un parachute*, je pourrais sauter et leur échapper vivant ; j'ai donc l'expérience que je pourrais m'en sortir vivant dans d'autres circonstances, mais cette expérience n'est pas l'expérience qu'il m'est réellement possible, ici et maintenant, de m'en sortir vivant. Il est clair que notre phénoménologie des alternatives du choix n'est pas cette phénoménologie des alternatives possibles dans d'autres circonstances, mais bien une phénoménologie des possibilités ouvertes *dans le contexte d'expérience*, et donc de manière absolue.

Résumons rapidement les différents éléments que nous avons découvert lors de cette enquête phénoménologique. Nous avons vu que notre expérience agentive contenait deux aspects importants, l'aspect du Soi comme Source et l'aspect des possibilités alternatives. Nous avons vu que ces expériences étaient plus saillantes dans l'expérience du choix que dans l'expérience de l'action volontaire, et plus précisément dans l'expérience du choix difficile par opposition au choix confiant. En effet, dans le choix difficile, notre expérience du Soi comme Source est une expérience du Soi comme Source ultime ou comme Initiateur d'une série causale. C'est-à-dire que nous expérimentons notre choix comme n'ayant pas de cause antécédente suffisante. Quant à l'expérience des possibilités alternatives, c'est également lors des choix difficiles qu'elle est la plus incontestable et la plus forte : elle est alors l'expérience des alternatives comme *absolument* (ou inconditionnellement) *possibles*. Avons-nous également ces deux expériences (Source Ultime et alternatives absolument possibles) dans

nos actions volontaires et dans nos choix confiants ? Ce n'est pas évident, et je ne chercherai certainement pas à l'établir ici. Il se pourrait très bien que notre phénoménologie de la liberté se concentre sur quelques moments cruciaux où nous prenons des décisions difficiles, les autres choix et les autres actions n'étant considérés comme libres que dans la mesure où ils découlent de ces choix « formateurs ».

Un point que nous avons laissé en suspend concerne l'idée de causalité agentive : notre expérience du choix difficile nous présente-t-elle comme étant, dans un sens irréductible, la cause agentive de nos choix ? Ou bien nous présente-t-elle comme étant certes les auteurs de nos choix, mais seulement dans le sens où *nos* raisons (nos désirs et nos croyances) sont à l'origine de nos choix (sans en être des causes *déterminantes* pour les raisons que nous venons de voir) ? Selon la réponse qu'on apporte à cette question, on aura deux modèles différents de la conception libertarienne du libre arbitre : le modèle de la causalité agentive, dont (O'Connor 2000) soutient qu'il est la meilleure description phénoménologique de notre expérience du libre arbitre, ou le modèle « événementiel indéterministe » qui a été développé notamment par (Kane 1996) (sans que ce dernier ne s'engage très clairement sur l'exactitude phénoménologique de ce modèle). Pour ma part, je ne pense pas être en mesure de favoriser tel ou tel modèle particulier ; je m'en tiendrai donc à un modèle « général » qui ne contient que des éléments acceptables par (à peu près) toute phénoménologie libertarienne : à savoir que nous expérimentons au moins certains choix difficiles dont nous nous sentons la Source Ultime, de telle sorte que nous sentons que nous aurions pu faire le choix inverse.

Il est intéressant de noter qu'au terme de cette enquête phénoménologique, nous retrouvons les trois éléments fondamentaux qui étaient apparus lors de la discussion des théories de l'illusion comme des composantes probables de notre concept de libre arbitre (voir p.306) : la condition de Source Ultime, la condition des Possibilités Alternatives, et la condition de contrôle de nos volitions (qui correspond à la notion de choix). Nous avons vu que les compatibilistes avaient toutes les peines à refuser ces conditions et à se rabattre sur des conditions plus faibles, parce que de solides arguments nous menaient des conditions faibles aux conditions fortes (argument de la manipulation, argument de la conséquence, argument de l'empêchement psychologique).

Ces deux sections confirment donc le « portrait robot » minimal de notre concept naturel du libre arbitre : nous nous concevons naturellement comme capables de faire des choix difficiles, c'est-à-dire de déterminer nos intentions ou nos volitions en étant la Source Ultime de cette détermination, et être la Source Ultime de ces déterminations suppose également qu'il nous soit possible (dans un sens absolu) de choisir la détermination contraire. Reste à déterminer si ce contenu de notre concept de libre arbitre est réellement incompatible avec l'idée de déterminisme.

IV.2.d. La portée incompatibiliste

Nous avons pour l'instant étudié les données phénoménologiques sans poser directement la question de leur compatibilité avec le déterminisme. Mais, pour reprendre la citation de Taylor placée en exergue de ce chapitre, « il est assez clair qu'elles vont présenter des difficultés pour cette théorie ». (Taylor 1991, 42) Ce qui est évident en tout cas, c'est que notre phénoménologie de la Source Ultime, et notre phénoménologie des Possibilités (absolues) Alternatives, ont pu générer au moins une apparence de conflit avec l'idée de déterminisme. Reste à déterminer si ce conflit est seulement apparent ou si l'incompatibilité est réelle.

Commençons par notre phénoménologie de la source ultime. Cette phénoménologie pourrait-elle être véridique si nous vivions dans un monde déterministe ? Telle que je l'ai décrite plus haut, cette phénoménologie représente nos choix difficiles comme n'étant déterminés par aucune cause antécédente suffisante. Mais si le déterminisme causal est vrai, alors tout événement du monde (y compris nos choix difficiles) est tel qu'il existe une cause antécédente suffisante qui détermine que *cet* événement (et aucun autre) doit avoir lieu. On pourra objecter que le déterminisme physique n'a pas nécessairement à être un déterminisme *causal*. La définition la plus générale du déterminisme physique n'a en effet pas besoin d'avoir recours à la notion de causalité mais seulement à la notion de lois naturelles : l'univers est déterministe si et seulement si la conjonction de la description d'un instant donné du passé (par exemple du passé lointain) et des lois de la nature implique logiquement l'état complet de l'univers (et donc tout événement particulier) à tout instant futur de l'univers. Si l'on admet cette définition, n'est-il pas possible que nos choix difficiles soient déterminés sans pour autant être *causalement* déterminés, et que par conséquent ils puissent être sans *cause* antécédente déterminante ? Dans ce cas, notre phénoménologie de la source ultime ne serait-elle pas véridique malgré le caractère déterministe de l'univers ? Certes, si notre phénoménologie de la source ultime était une phénoménologie de l'absence de *cause* au sens strict, alors elle pourrait être compatible avec une certaine forme de déterminisme naturel. Mais dans cette forme de déterminisme, il y aurait néanmoins certaines *conditions* antécédentes qui, en conjonction avec les lois de la nature, impliqueraient logiquement nos choix, et ces conditions antécédentes seraient manifestement en un certain sens (non causal) la source antécédente de nos choix. Or il est peu probable que notre phénoménologie de la source contienne spécifiquement le concept de source *causale* ; au contraire, il semble probable que notre phénoménologie contienne seulement un concept général d'initiation dont les conditions de vérités sont essentiellement temporelles ; c'est le concept d'un choix sans antécédents, que ces antécédents soient causaux ou non. Comment établir que notre expérience de la source ultime ne se limite pas à une absence de source *causale* ? Il me semble qu'il serait difficile de l'établir par un argument phénoménologique direct. L'argument le plus convaincant à mon sens, c'est qu'une phénoménologie de l'absence de source *causale* au sens strict n'impliquerait *pas* une phénoménologie des possibilités alternatives (puisque les possibilités alternatives pourraient être « fermées » par des conditions antécédentes non causales) et que la phénoménologie des possibilités alternatives est solidement établie. Ainsi, bien que la phénoménologie de la source ultime soit probablement le cœur conceptuel de notre expérience du libre arbitre, néanmoins d'un point de vue argumentatif c'est sa conséquence, la phénoménologie des alternatives, qui permet le mieux d'établir son caractère incompatibiliste.

Passons donc à la phénoménologie des possibilités alternatives elles-mêmes. Est-elle véritablement incompatible avec le déterminisme causal ou naturel ? A première vue, la phénoménologie des possibilités alternatives est *directement* contradictoire avec le déterminisme causal puisque ce dernier consiste à dire que la conjonction du passé et des lois de la nature n'est logiquement compatible qu'avec un unique futur. Mais il existe, à ma connaissance, trois stratégies possibles pour accepter la phénoménologie des possibilités alternatives et refuser son caractère incompatibiliste.

La première stratégie, que nous avons déjà considérée, consiste à dire que nous avons seulement une expérience des possibilités *conditionnelles* et non pas des possibilités *absolues*. Nous avons vu que cette stratégie avait peu de chances de marcher : nous avons l'expérience que les deux choix sont possibles « ici et maintenant, réellement » et donc « dans les circonstances présentes ». Si nous avions seulement l'expérience que les alternatives *seraient* possibles dans d'autres circonstances, nous n'aurions pas l'expérience qu'elles *sont* possibles.

La deuxième stratégie consisterait à dire que notre expérience contient un concept de « pouvoir » très particulier, différent du concept de « possibilité » qui permet de définir le déterminisme. Par exemple, j'aurais l'expérience que « je pourrais faire le choix A et le choix B (dans les circonstances présentes) » dans le sens faible où « j'ai les capacités générales, physiques et psychologiques (ou la base causale) requises pour faire le choix A et le choix B (dans les circonstances présentes) » ; cela est compatible avec l'idée qu'il y aurait des causes déterminantes en faveur du choix A. L'important est donc ici d'avoir un concept de « je peux faire A ou B (dans les circonstances présentes) » qui n'implique pas logiquement « il est possible que je fasse A ou B (dans les circonstances présentes) ». Là encore, il me semble que cette stratégie est phénoménologiquement inadéquate. C'est ce qu'on pourrait appeler la *phénoménologie du pouvoir sans possibilité*. Rappelons-nous que l'origine de la phénoménologie des alternatives *possibles* se trouve essentiellement dans notre expérience de la délibération : pour délibérer entre deux options (de manière rationnelle), il faut que j'aie l'expérience que je pourrais faire l'une ou l'autre... mais il faut « aussi » que j'aie l'expérience que l'une et l'autre sont *possibles*. Imaginons que j'aie l'expérience que l'une des deux options est *impossible* bien que j'aie aussi l'expérience d'avoir les capacités ou la base causale qui permette de dire que « je peux choisir cette option ». Serait-il rationnel, dans ces circonstances, de délibérer ? Il est évident que non. Par conséquent, notre expérience de la délibération et du choix contient bien une expérience que les alternatives sont *en elles-mêmes* réellement possibles, et non pas seulement l'expérience que « je peux » réaliser l'une ou l'autre, dans un sens faible compatible avec l'*impossibilité* d'une des deux alternatives.

La troisième stratégie consiste à accepter que notre expérience des alternatives est bien une expérience que « il est possible que je choisisse A et il est possible que je choisisse B dans les circonstances actuelles », mais considère que le domaine de ce qui compte comme « possible dans les circonstances actuelles » est susceptible de variations contextuelles. C'est la stratégie développée par Terry Horgan depuis plusieurs années, en collaboration avec divers auteurs (Horgan 1979; Horgan, Tienson, and Graham 2003; Horgan and Timmons 2011; Horgan 2011a). (Horgan 1979) s'appuie en particulier sur la logique des opérateurs modaux (je pourrais faire que *p*, il est possible que *p*) dans la logique et la linguistique contemporaine. Ces opérateurs sont classiquement interprétés comme des quantificateurs sur les mondes possibles : il existe un monde possible dans lequel *p*. Or les quantificateurs sont notoirement susceptibles de variations contextuelles, comme nous l'avons vu à propos du contextualisme épistémologique (voir ch.2 note 169). Par exemple, la phrase « il n'y a rien dans le frigo » peut être vraie dans un contexte quotidien où l'on quantifie uniquement sur la nourriture, mais elle sera fausse dans un contexte limite où l'on quantifie sur tous les objets matériels, de quelque grandeur que ce soit (plateaux du frigo, atomes d'oxygène, etc.). De même, la phrase « j'aurais pu faire autrement dans les mêmes circonstances » sera compatible avec le déterminisme dans les contextes courants où l'on envisage des possibilités larges, et ne deviendra incompatible avec le déterminisme que dans des contextes limites où l'on prend en compte *toute forme* d'empêchement à la possibilité. Notre expérience commune du libre arbitre serait donc compatible avec le déterminisme, et c'est seulement l'expression linguistique de cette expérience qui, dans certains contextes (en particulier les contextes où l'on *parle* du déterminisme) prendrait un sens limite incompatibiliste.

La théorie de Horgan est riche et complexe, mais il convient de s'y arrêter car c'est sans doute l'auteur compatibiliste le plus conscient des difficultés du compatibilisme. En particulier, Horgan a conscience qu'un vrai compatibilisme doit être un compatibilisme *à propos* de notre concept agentif phénoménal et que ce concept agentif phénoménal a des aspects introspectivement clairs qui peuvent poser problème avec l'idée de déterminisme (le

soi comme source, et les possibilités alternatives).³⁵⁸ Il défend pourtant que ces aspects phénoménaux ne sont *pas* incompatibilistes. Pour bien évaluer sa théorie, il faut distinguer deux niveaux : d'une part sa conception compatibiliste elle-même (le modèle qu'il donne d'une phénoménologie compatibiliste des alternatives), et d'autre part ses *raisons* d'adopter une telle conception compatibiliste de notre phénoménologie. On pourrait discuter longuement des détails de sa théorie compatibiliste elle-même, mais d'un point de vue dialectique il sera suffisant de s'intéresser ici à ses raisons.³⁵⁹ La difficulté vient de ce que Horgan adopte une conception compatibiliste de notre phénoménologie pour des raisons *non phénoménologiques*, ou non introspectives. En effet, Horgan considère que l'analyse introspective a ses limites : bien qu'il accepte de la mener jusqu'à des aspects aussi précis que le soi comme source ou les alternatives possibles, il estime que la question du compatibilisme n'est quant à elle pas accessible à l'introspection. Non pas parce que notre expérience en elle-même ne serait ni compatibiliste ni incompatibiliste, mais parce qu'il y a certains aspects de nos expériences qui ne sont pas « manifestes », pas accessibles (pas aisément accessibles ou pas accessibles du tout) à l'attention introspective :

Lorsque que j'examine introspectivement ma phénoménologie en me posant cette question assez complexe [du compatibilisme], il me semble que *aucune réponse n'est introspectivement manifeste* pour moi. (De fait, il me semble introspectivement manifeste qu'aucune réponse à cette question n'est introspectivement manifeste.) (Horgan 2011b, sec. 3.2)

Si donc Horgan estime que notre concept phénoménologique de libre arbitre est contextuellement variable (et qu'il est compatibiliste dans le contexte de notre expérience agentive courante), ce n'est pas parce qu'il aurait observé introspectivement ce compatibilisme, mais seulement parce qu'il estime avoir des raisons indirectes et abductives d'adopter le compatibilisme, et qu'il se rabat donc sur la théorie phénoménologique compatibiliste qui lui semble la plus plausible. Mais Horgan a-t-il raison de dire que l'introspection ne permet pas de déterminer le caractère compatibiliste ou non de notre phénoménologie des alternatives ? A-t-il raison en outre d'affirmer qu'il est *manifeste* que l'introspection est ici incompétente, et qu'on est donc obligé d'avoir recours à des raisons indirectes abductives ? Il me semble que la réponse est clairement non : l'introspection est *peut-être* incompétente à trancher la question de compatibilisme de notre expérience, mais il n'est certainement pas « manifeste qu'aucune réponse à cette question n'est introspectivement manifeste », pour la simple raison que lorsque nous examinons introspectivement notre phénoménologie des alternatives en envisageant la question du déterminisme, *il y a* une réponse qui vient naturellement à l'esprit, à savoir une réponse incompatibiliste. Et à vrai dire, Horgan lui-même a reconnu ce point :

Je reconnais que lorsque nous examinons introspectivement notre phénoménologie de l'action libre, avec son aspect présentationnel du soi comme source qui inclut l'aspect de la liberté [des alternatives] comme composant essentiel, et lorsque nous nous

³⁵⁸ « We ourselves think that it is introspectively *just obvious* that actions are normally experienced as emanating from oneself as source, and also are normally experienced as being such that one could have done otherwise in the actual circumstances » (Horgan and Timmons 2011, 188)

³⁵⁹ Une des difficultés de la théorie en elle-même semble être la suivante : il n'est pas aisé de déterminer dans quelle mesure elle est compatibiliste. En effet, si le *simple fait d'évoquer l'idée de déterminisme* est suffisant pour créer un contexte dans lequel le contenu de notre expérience est incompatible avec le déterminisme, alors il semble que (i) notre concept de libre arbitre est *au moins parfois* incompatibiliste, et plus radicalement encore que (ii) la réponse à la question de compatibilité est incompatibiliste « toutes les fois que je la prononce ou que je la conçois en mon esprit »..

demandons en même temps si la véridicalité de cette phénoménologie requiert [...] le libertarianisme, nous sentons une *certaine* tendance à juger que la réponse à cette question est Oui. (Horgan 2011a, 94)

Donc si Horgan estime que l'introspection est incompétente sur cette question, ce n'est *pas* parce que l'introspection ne délivrerait aucune réponse, mais parce qu'il estime avoir des raisons indépendantes de *rejeter* la réponse qu'apporte *prima facie* l'introspection. Dans la fin de cette section, je ferai deux choses : premièrement, je vais montrer que les raisons indépendantes de Horgan (pour rejeter les données introspectives incompatibilistes) ne sont pas convaincantes ou pas pertinentes ; deuxièmement, j'apporterai une confirmation empirique du fait que l'introspection apporte bel et bien une réponse (au moins *prima facie*) à la question du compatibilisme ; c'est-à-dire que je montrerai que la « tendance » introspective qu'éprouve Horgan (ou que j'éprouve moi-même) à considérer notre expérience agentive comme incompatibiliste n'est pas une simple idiosyncrasie, totalement variable d'une personne à l'autre, mais qu'elle est une tendance « normale », en un sens statistique.

Pour quelles raisons Horgan rejette-t-il la réponse que semble lui donner l'introspection à la question du compatibilisme ? Après tout, lorsque l'introspection semblait lui indiquer que notre phénoménologie contient l'aspect du soi comme source ou l'aspect des alternatives, Horgan se fiait à ces données de l'introspection. Pourquoi cesse-t-il de faire confiance à l'introspection lorsque celle-ci, à l'examen de la question compatibiliste, fait naître une « certaine tendance » à répondre de manière incompatibiliste ? Horgan (2011b, sec. 3.5) apporte cinq types de considérations évidentielles qui, collectivement, constituent pour lui un fort argument abductif en faveur du compatibilisme. On peut résumer ces cinq raisons à deux arguments fondamentaux : (i) l'argument de non prise en compte, et (ii) l'argument scientifique d'insatisfiabilité.

L'argument de non prise en compte (raisons 1, 3 et 4 de Horgan) consiste à dire que la question du déterminisme ou de l'indéterminisme n'est *pas prise en compte* dans les contextes dans lesquels nous cherchons à appliquer notre concept de libre arbitre. En particulier, lorsque l'on nous décrit un scénario d'action qui ne contient aucune référence à la question du déterminisme ou de l'indéterminisme, nous sommes généralement capables de classer cette action libre ou non libre. De même, dans la vie quotidienne, nous opérons une telle classification indépendamment de toute information ou de toutes données concernant le déterminisme ou l'indéterminisme. Enfin, les pratiques d'attribution de responsabilité morale sont généralement associées par nous à la liberté, sans que nous cherchions, là encore, à déterminer si l'action de l'agent était déterministe ou non. Pour Horgan, la non prise en compte du déterminisme ou de l'indéterminisme s'explique beaucoup mieux si l'on admet l'hypothèse compatibiliste : nous ne prenons pas en compte la question du déterminisme tout simplement parce que notre concept de libre arbitre est compatible avec l'une ou l'autre réponse à la question déterministe. Que penser de cet argument ? Cet argument me semble extrêmement faible parce que l'incompatibilisme peut apporter une explication tout aussi forte (quoiqu'un peu plus longue) de la non prise en compte. Cette explication a deux aspects (qui peuvent être pris ensemble ou séparément) : la présomption des conditions nécessaires et l'intuition de résilience. Pourquoi ne « cherchons-nous pas » à savoir si le monde est indéterministe lorsque nous attribuons la liberté ou la responsabilité ? Tout d'abord parce que, dans tout contexte normal, nous *présumons* naturellement que les conditions nécessaires d'application de notre concept sont satisfaites. En effet, l'expérience du libre arbitre dont nous parlons est une expérience de sens commun, qui appartient fortement à notre conception naturelle de nous-mêmes et de nos actions ; il est donc évident que ces conditions nécessaires seront normalement présumées satisfaites *quelles que soient ses conditions nécessaires d'application*, qu'elles soient compatibilistes ou incompatibilistes. L'existence d'une telle

présomption devrait être admise aussi bien par les compatibilistes que par les incompatibilistes, du simple fait de la naturalité de l'expérience du libre arbitre. Mais s'il y a une telle présomption, alors l'incompatibilisme implique directement que nous avons une présomption naturelle d'indéterminisme (puisque dans l'incompatibilisme l'indéterminisme est une condition nécessaire d'applicabilité de notre concept). Et si nous *présumons* naturellement l'indéterminisme, nous n'avons pas besoin de soulever la question à chaque attribution de liberté ou de responsabilité morale. Avons-nous réellement une telle présomption naturelle d'indéterminisme ? Il semble bien que oui si l'on se réfère à la « question d'actualité » en philosophie expérimentale (voir chapitre 1 p.66) : nous avons une tendance naturelle à juger que notre univers « ressemble » plus à un univers où certains indéterminismes existent au moment des choix. Deuxièmement, si nous ne « cherchons pas » à savoir si le monde est déterministe ou non lorsque nous attribuons la liberté, c'est peut-être également parce que nous sentons vaguement que la réponse déterministe ne *changerait pas* notre jugement (ou notre « classification » entre actes libres et non libres). Cela veut-il dire que nous sommes compatibilistes ? Non, selon l'argument développé dans la section II.4. de ce chapitre, cela montre simplement que notre croyance au libre arbitre (et nos critères de distinction entre actes libres et non libres) est *résiliente* à l'éventuelle découverte du déterminisme. C'est-à-dire que si le déterminisme était prouvé, *dans ce cas* nous enlèverions l'indéterminisme de la liste des conditions constitutives de l'action libre ; mais cela n'empêche pas que *pour l'instant* il se trouve dans cette liste. Ainsi, lorsque nous cherchons à classer un acte comme libre ou non, nous ne faisons pas d'enquête scientifique sur le déterminisme parce que (i) nous présumons que la réponse est indéterministe, et (ii) nous sentons que si cette présomption était fautive nous changerions nos critères d'applicabilité. Je reconnais volontiers que l'explication de la « non prise en compte » par le compatibiliste est plus « simple » que celle-ci dans le sens où elle est moins longue et moins composée. Je ne suis pas sûr en revanche qu'elle soit « meilleure » dans la mesure où cette explication ne suppose aucune hypothèse que le compatibiliste n'accepte normalement. A supposer que l'explication compatibiliste soit meilleure, son bénéfice abductif me semble extrêmement ténu et incapable de contrebalancer à elle seule l'intuition forte et naturelle que notre expérience des alternatives est incompatible avec un monde déterministe.

Le deuxième argument de Horgan en faveur du compatibilisme est l'argument scientifique d'insatisfiabilité. Pour Horgan, les données de la science actuelle (en physique quantique, et en neuro-biologie) ne permettent aucunement de corroborer l'idée qu'il puisse y avoir des indéterminismes « bien placés » qui nous permettraient d'être libres dans un sens libertarien. Horgan présente cette considération comme une « raison » d'être compatibiliste. Mais c'est un type de raison dont nous avons déjà eu l'occasion de parler abondamment : il s'agit d'un argument « d'intégration », un argument qui soutient une certaine théorie du contenu de notre concept de libre arbitre *pour sauver* la véridicalité de notre conviction naturelle que nous sommes libres. Nous aurons l'occasion de discuter au chapitre suivant du cœur de l'argument – de la question de savoir si un concept libertarien est scientifiquement satisfiable ou non. Mais dans le présent chapitre, cet argument est simplement non pertinent, puisque nous n'envisageons ici la question du contenu (et donc la question phénoménologique) *indépendamment* de toute motivation d'intégration. Etant donné la force et l'importance épistémologique de notre conviction du libre arbitre, il est possible que nous soyons ultimement obligés de *réviser* ces premières conclusions phénoménologiques. Il est possible que nous soyons ultimement amenés à dire ceci : la phénoménologie, en elle-même, aurait toutes les raisons d'être incompatibiliste, mais une telle phénoménologie nous amènerait à tenir pour illusoire notre conviction du libre arbitre, donc notre phénoménologie doit *finalement* être tenue pour compatibiliste. Tout cela est possible, mais ne nous concerne pas pour l'instant. Or si l'on s'en tient aux considérations « indépendamment motivées » pour

aborder le contenu de notre phénoménologie, il me semble qu'il n'y a que deux considérations pertinentes : d'une part notre « tendance » naturelle à considérer par introspection que notre expérience des alternatives serait fausse dans un monde déterministe ; et d'autre part, éventuellement, le fait de la « non prise en compte ». Comme le fait de la non prise en compte peut très bien s'expliquer dans un modèle incompatibiliste, il me semble donc que Horgan n'apporte finalement aucune bonne raison de rejeter les données phénoménologiques directes.

Terminons par ces données phénoménologiques directes en faveur de l'incompatibilisme. Jusqu'à présent, nous avons mené l'enquête phénoménologique presque exclusivement au moyen d'une méthodologie « en première personne », la méthodologie « différentielle ». Cette méthodologie est sans doute suffisante pour établir l'existence des aspects de soi comme source et des possibilités alternatives. Mais pour aller jusqu'au bout de la question qui nous intéresse vraiment, c'est-à-dire pour trancher entre compatibilisme et incompatibilisme de notre phénoménologie, il sera bon de pouvoir combiner l'introspection en première personne avec une méthode en deuxième personne, et en particulier avec une méthode de philosophie expérimentale qui permettra de se convaincre que la « tendance » éprouvée par Horgan (ou par moi-même) à juger la phénoménologie des alternatives comme incompatibiliste est bel et bien une tendance « naturelle » ou normale, et non pas un biais de spécialiste ou une idiosyncrasie.

L'étude menée par (Deery, Bedke, and Nichols 2013) permet d'apporter exactement la confirmation dont nous avons besoin ici. Tout d'abord, il s'agit d'une étude proprement *phénoménologique* : elle demande aux participants de prêter attention à leurs expériences : « considérez la manière dont les choses vous semblent / vous apparaissent » (Deery, Bedke, and Nichols 2013, 132). Deuxièmement, l'enquête se concentre sur la phénoménologie du *choix*, et en particulier d'un choix « difficile », c'est-à-dire sans raison contrastive évidente : dans la première étude, il s'agit d'un choix entre aller à droite ou à gauche dans un parcours de luge, dans la seconde, d'un choix entre deux associations de bienfaisance à qui donner de l'argent, dans la troisième, d'un choix entre deux boutons sur lesquels appuyer. Troisièmement, l'enquête se concentrait spécifiquement sur l'aspect phénoménologique des possibilités alternatives : après avoir lu la description du choix à faire, les participants devaient dire s'ils éprouvaient ou non une expérience de « pouvoir faire l'un ou l'autre choix ». Les participants qui ne manifestaient *pas* une conscience introspective de cet aspect phénoménologique étaient éliminés de l'étude. Ces candidats éliminés étaient très peu nombreux, ce qui confirme, s'il en était besoin, que l'existence d'une phénoménologie des alternatives est un fait statistiquement « normal » et non pas idiosyncrasique ou un biais de spécialiste.³⁶⁰ Reste à étudier les détails de cet aspect phénoménologique, et les candidats restant étaient donc invités à concentrer leur attention sur cet aspect particulier de leur phénoménologie. Enfin, un dernier aspect particulièrement intéressant de cette enquête, c'est que ses auteurs ont prêté un grand soin à éviter le risque d'une « confusion » conceptuelle du déterminisme avec d'autres notions plus fortes. Pour cela, ils ont tout d'abord évité le terme de « déterminisme » dont plusieurs auteurs ont estimé qu'il était susceptible de provoquer des incompréhensions (le terme utilisé était celui de « complétude causale »). Et surtout, ils ont utilisé la technique de « l'entraînement à un critère », qui consiste à s'assurer par des questions préalables que le concept en question est bien compris et qu'il n'est pas confondu avec tel ou tel autre concept. Grâce à ces questions préalables, les auteurs ont pu repérer les candidats qui confondaient le déterminisme avec le fatalisme ou avec le « court-circuit » et ils

³⁶⁰ Les proportions sont les suivantes : 4 sur 82 dans la première étude, 7 sur 84 dans la seconde, 7 sur 91 dans la troisième.

ont pu éliminer ces candidats de l'étude.³⁶¹ En fin de compte, les résultats manifestaient donc : un jugement d'introspection concernant la compatibilité entre le déterminisme *bien compris*, et la phénoménologie agentive dans *son aspect des possibilités alternatives*. La question de compatibilité était posée de la manière suivante (pour l'exemple de la troisième étude) :

Maintenant, rappelez-vous la situation du choix du bouton. Un peu plus tôt, vous avez donné votre accord au jugement suivant :

« Lorsque je décide quelle option choisir, il me semble que je peux soit choisir [le bouton] H soit choisir [le bouton] V. »

En considérant cette déclaration sur la manière dont les choses vous semblaient avant votre choix, et votre compréhension de la complétude causale, veuillez indiquer votre degré d'accord avec la proposition suivante :

« Si la complétude causale est vraie, alors je n'aurais pas réellement pu choisir autrement que je ne l'ai fait. »

Le degré d'accord devait être indiqué sur une échelle de 1 à 7 (1 = complètement en désaccord, 7 = complètement d'accord). Les résultats de l'étude sont sans ambiguïté : les participants, en prêtant attention à leur expérience des alternatives et à l'idée (bien comprise) de déterminisme, ont manifesté un très fort degré d'accord (plus de 5 en moyenne) avec la description incompatibiliste de leur expérience de pouvoir faire autrement.³⁶²

Les expériences de (Deery, Bedke, and Nichols 2013) permettent donc d'établir le point suivant :

Nous avons une forte tendance naturelle (ou en tout cas répandue) à considérer, lorsque nous prêtons attention à notre expérience phénoménale agentive, qu'un de ses aspects au moins (l'aspect des alternatives) est incompatible avec le déterminisme (bien compris).

Cela veut-il dire que notre expérience phénoménale agentive est *réellement* incompatibiliste ? Il n'y a pas ici d'implication stricte : notre jugement introspectif *pourrait* être erroné. Mais notre jugement introspectif, compte tenu de sa force et de sa représentativité, constitue clairement une très forte raison *prima facie* pour dire que notre phénoménologie des alternatives est réellement incompatibiliste. Pour renverser cette raison, il faudrait une raison au moins aussi forte en sens contraire. Or il n'y a pas, à mon sens, de raison aussi forte en sens contraire, pas en tout cas si l'on s'en tient aux raisons indépendantes du problème d'intégration, c'est-à-dire si l'on s'en tient aux considérations qui concernent la phénoménologie *pour elle-même* indépendamment de sa véridicalité ou de son apport épistémologique.

Nous verrons plus loin si les considérations d'intégration nous obligent à réviser notre jugement sur la phénoménologie et à considérer que, finalement, ce jugement répandu manifeste une erreur commune d'introspection. Mais la conclusion de notre enquête strictement phénoménologique est claire : si l'on étudie notre expérience phénoménale agentive pour elle-même, sans se soucier de savoir si son contenu est vrai ou connaissable,

³⁶¹ Là encore, les candidats éliminés étaient peu nombreux, ce qui suggère que la tendance à confondre le déterminisme avec d'autres concepts n'est pas si répandue que pouvaient le penser (Nahmias et al. 2006) : 6 sur 78 dans la première étude, 5 sur 77 dans la seconde, 4 sur 84 dans la troisième.

³⁶² Dans le détail, les degrés d'assentiment sont les suivants : 4.93 et 5.24 pour la première étude (divisée en deux « conditions », l'une passée, l'autre future), 5.60 et 5.78 pour la deuxième étude (divisée, là aussi, en deux « conditions », l'une de choix imaginé, l'autre de choix actuel), 5.34 pour la troisième étude.

nous avons des raisons extrêmement fortes de penser que ce contenu est *réellement* (et non pas illusoirement) incompatibiliste.

V. Conclusion

Pour simplifier la dialectique de ce chapitre, on peut considérer qu'il contient essentiellement deux arguments incompatibilistes, un argument d'inférence à la meilleure explication (défaisable mais non défaite), et un argument d'analyse phénoménologique directe.

Le premier argument consiste à dire ceci : puisque nous savons, par définition, que notre concept de libre arbitre est un concept tel qu'il génère naturellement un conflit avec l'idée de déterminisme, l'explication la plus simple, à première vue, de ce conflit, c'est qu'il y a incompatibilité réelle entre ce concept et le concept de déterminisme. Il *pourrait* y avoir d'autres explications, mais celle-ci est clairement, de prime abord, la plus simple. Par conséquent, si quelqu'un défend le compatibilisme à propos de notre expérience naturelle du libre arbitre, il devra proposer une meilleure explication du conflit originaire. C'est-à-dire qu'il devra *non seulement* nous présenter un concept compatibiliste (comme analyse du contenu de cette expérience), mais il devra *surtout* nous convaincre que ce concept a de bonnes chances de provoquer un conflit avec l'idée de déterminisme (bien qu'il ne soit pas *en fait* conceptuellement incompatible avec le déterminisme). S'il nous présente un concept qui est clairement compatibiliste, mais si clairement compatibiliste que nous n'avons aucune raison indépendante de supposer qu'il génèrera un conflit avec l'idée de déterminisme, alors nous aurons toutes les raisons de penser qu'il ne s'agit *pas* de notre concept de libre arbitre (mais seulement, éventuellement, d'une condition *nécessaire et non suffisante* pour être un agent libre). Autrement dit, le compatibiliste devra nous présenter une « théorie de l'illusion » qui soit indépendamment plausible, autrement, il manquera le « défi de la pertinence » : le concept qu'il aura mis en avant ne sera tout simplement pas « pertinent » pour la discussion du libre arbitre. J'ai défendu dans ce chapitre qu'aucune « théorie de l'illusion » offerte par les compatibilistes n'était indépendamment plausible, de sorte que tous les concepts compatibilistes correspondants ont toutes les chances d'être des conditions seulement nécessaires de l'action libre. La meilleure explication du conflit originaire reste donc qu'il y a un conflit conceptuel réel et non pas un conflit illusoire. Cette discussion nous a même permis d'aller un peu plus loin ; en effet, j'ai essayé de montrer que les « raisonnements illusoires » ou fallacieux auxquels ont généralement recours dans leurs théories de l'illusion avaient de bonnes chances d'être en fait des raisonnements par lesquels nous apercevons *validement* (ou *légitimement*) certaines conditions incompatibilistes du libre arbitre. Cette discussion a donc permis de donner un premier « portrait robot » (possible) des conditions nécessaires caractérisant notre concept de libre arbitre.

La deuxième démarche est plus directe : pour voir si notre expérience du libre arbitre a un contenu incompatibiliste ou non, la démarche consiste à localiser cette expérience et à l'analyser directement. Pour ce faire, j'ai défendu que l'origine de notre concept de libre arbitre se trouvait (très probablement) dans notre expérience phénoménale agentive, et en particulier notre expérience phénoménale des choix difficiles (et non pas dans notre expérience perceptive des autres agents, ni dans notre expérience doxastique reposant sur notre schème conceptuel moral, et *peut-être pas* dans l'expérience des actions volontaires ou celle des choix confiants). Il ne restait plus ensuite qu'à analyser, phénoménologiquement, le contenu de cette expérience phénoménale des choix difficiles. Et j'ai cherché à montrer que cette expérience contenait deux aspects susceptibles de générer un conflit avec le déterminisme (l'aspect de soi comme source ultime et l'aspect des alternatives absolument possibles), puis j'ai montré que ces deux aspects avaient toutes les chances d'être *réellement*

incompatibilistes, si du moins l'on s'en tient aux arguments qui concernent directement notre phénoménologie agentive, sans se soucier pour l'instant de ses mérites épistémologiques ou de sa véridicalité.

Ce deuxième argument est plus simple, et plus « satisfaisant intellectuellement », que le premier qui parle du « concept de libre arbitre » sans jamais nous dire de quoi il s'agit exactement. Pourquoi donc ne pas s'en tenir à l'argument phénoménologique direct ? Parce qu'il est aussi plus « engageant », il suppose d'admettre le modèle phénoménal à propos de l'origine de notre concept du libre arbitre, et ce modèle phénoménal est sans doute contestable. Ce que j'ai cherché à montrer, c'est que *même si l'on ignore l'origine* du concept de libre arbitre (ou même si l'on adopte un autre modèle), on a tout de même une bonne raison abductive d'être incompatibiliste. L'analyse phénoménologique directe vient ensuite renforcer cette présomption initiale.

Si nous combinons les résultats de la première partie aux résultats du présent chapitre, voici la situation dialectique à laquelle nous arrivons :

- (1) Notre expérience du libre arbitre nous confère une justification *prima facie* à croire que nous sommes libres et que nous pouvons le *savoir* en nous fiant à cette expérience.
- (2) Notre expérience du libre arbitre est (très probablement) notre expérience phénoménale agentive, en particulier notre expérience des choix difficiles et de ses aspects de soi comme source et de possibilités alternatives.
- (3) Par ces aspects, notre expérience du libre arbitre a (très probablement) un contenu réellement incompatibiliste.

De ces résultats on peut déduire :

- (4) Nous avons une justification *prima facie* à croire que nous sommes libres dans un sens qui implique l'indéterminisme, et à croire que nous pouvons *savoir* cela sur la base de notre expérience.

Reste à savoir si ces justifications *prima facie* résistent aux objections directes et fondationnelles qu'on peut faire. En particulier, le résultat incompatibiliste de ce chapitre pourrait constituer en elle-même, aux yeux de certains auteurs, une réfutation de notre justification *prima facie* à croire au libre arbitre, parce que ce contenu serait trop « exigeant » pour être vrai (objection directe) ou pour être connaissable sur la base de notre expérience (objection fondationnelle). Si l'objection directe fonctionnait, alors nous aurions encore la possibilité soit de devenir incompatibiliste dur (notre concept est bel et bien incompatibiliste mais du coup, nous ne sommes pas libres), soit de réviser notre phénoménologie pour sauver le libre arbitre (en fin de compte, il doit y avoir une erreur dans nos analyses phénoménologiques) – ce serait une forme de compatibilisme d'ajustement. De même, si l'objection fondationnelle fonctionnait, nous aurions la possibilité soit de devenir sceptique à propos de la liberté (notre concept de libre arbitre est indéterministe, mais du coup nous ne pouvons pas savoir, en l'état, si nous sommes libres ou non ; la science future nous l'apprendra peut-être), soit de nouveau de réviser notre phénoménologie pour sauver la connaissance introspective du libre arbitre (notre expérience doit avoir un contenu plus faible qu'il ne semble, autrement elle ne nous donnerait pas une *connaissance* de son contenu). J'envisagerai l'objection directe dans le chapitre suivant, et l'objection fondationnelle dans le chapitre 6. Je m'attacherai à montrer que l'une et l'autre objections sont non concluantes, de sorte que notre justification *prima facie* à croire au libre arbitre et à croire que nous le savons par expérience restera ultimement comme justification *ultima facie*.

Chapitre 5 : Absence de réfutation directe du libertarisme

I finished the first part of Renouvier's second Essais and see no reason why his definition of free will—"the sustaining of a thought because I choose to when I might have other thoughts"—need be the definition of an illusion.
(William James, in Perry 1936, I:323)

Perhaps in the brain the unstable equilibrium is so delicate that [...] the choice between possible transitions is determined by a psychological cause called "volition". All this is *possible*.³⁶³
(Russell 2009, 42)

I.	Réfutations a priori.....	348
I.1.	L'argument de la chance (contre la liberté des alternatives).....	349
I.1.a.	L'argument traditionnel du dilemme	349
I.1.b.	Réponse par la causalité agitive.....	350
I.1.c.	Réponse par la causalité événementielle	355
I.1.d.	L'argument du contrôle diminué	358
I.1.e.	L'argument du rembobinage.....	365
I.1.f.	Conclusion.....	373
I.2.	L'argument basique (contre la liberté de source ultime).....	374
I.2.a.	L'argument	374
I.2.b.	Réponse dans le modèle agentif	376
I.2.c.	Réponse dans le modèle événementiel	377
I.2.d.	Conclusion	379
I.3.	Conclusion : la concevabilité du libre arbitre	380
II.	Réfutations empiriques.....	381
II.1.	L'objection du déterminisme	381
II.1.a.	Le déterminisme ne peut pas être établi a priori	381
II.1.b.	Preuve empirique du déterminisme universel ?	382
II.1.c.	Preuve empirique du déterminisme macroscopique ?	384
II.1.d.	Preuve empirique du déterminisme neuronal ?	385
II.1.e.	Conclusion : pas d'objection déterministe	388
II.2.	L'objection d'intelligibilité empirique	389
II.3.	La causalité agitive face aux objections empiriques	392
III.	Conclusion : une « information recevable ».....	398

³⁶³ Pour être honnête, la citation de Russell continue ainsi: « All this is *possible*, but no more than possible; there is not the faintest positive reason for supposing that anything of the sort actually takes place. » Mais la concession que « tout ceci est *possible* » est tout ce qui nous intéressera dans ce chapitre. Pour la seconde partie de la remarque de Russell, la partie épistémologiquement pessimiste, ce sont les chapitres 2, 3 et 6 qui y répondent.

Dans la première partie, nous avons établi que nous avons une justification *prima facie* à nous fier à notre expérience du libre arbitre, aussi longtemps que son contenu nous était inconnu. Mais nous avons vu également que ce contenu, une fois déterminé, pourrait éventuellement constituer une réfutation de la justification *prima facie*. Dans le chapitre 4, nous venons d'établir que le contenu de notre expérience était très probablement incompatibiliste, c'est-à-dire qu'il implique logiquement l'indéterminisme causal. Nous avons donc établi les prémisses (IM) et (IA) de ce qui pourrait ultimement constituer un problème d'intégration pour le libre arbitre (voir chapitre 3 p. 239) :

(IM) notre expérience E nous donne connaissance du fait que nous sommes libres

(IA) être libre suppose l'indéterminisme causal.

Cette nouvelle information, (IA), pourrait nous amener à remettre en cause la première, (IM), et ce de deux manières différentes, que j'ai appelées plus haut la « réfutation directe » et la « réfutation fondationnelle ». Puisque notre expérience du libre arbitre s'avère avoir un contenu « lourd » (dans le sens où son contenu a des *implications* lourdes, à savoir l'indéterminisme), on pourrait douter que notre expérience soit une source épistémique suffisamment solide ou autorisée pour donner une telle information. C'est là le problème d'intégration proprement dit, que j'envisagerai au chapitre 6. Mais de manière plus directe, l'analyse incompatibiliste pourrait constituer une réfutation immédiate de notre expérience du libre arbitre si ce contenu incompatibiliste est logiquement impossible. De fait, une objection traditionnelle au libertarianisme (ou plutôt une famille d'objections) consiste à dire que cette conception du libre arbitre est inconcevable, impossible, ou insatisfaisable dans le monde tel que nous le connaissons. C'est cette famille d'objections que j'envisagerai dans ce chapitre. Pour ce faire, je repartirai donc des éléments de contenu conceptuel mis en évidence au chapitre précédent (le concept du soi comme source ultime et le concept d'alternatives absolument possibles) ; j'étudierai les objections qui ont été formulées pour établir que ces éléments sont impossibles (soit logiquement impossibles, soit impossibles compte tenu de ce que nous savons aujourd'hui sur le monde) ; enfin je m'efforcerai de montrer que ces objections ne sont pas concluantes, au sens où l'on peut construire des modèles du libre arbitre qui satisfont ces conditions fortes et qui sont à la fois concevables, métaphysiquement possibles et compatibles avec ce que nous savons aujourd'hui sur le monde.

Ce chapitre s'attaquera donc au deuxième des deux défis traditionnels du libertarianisme, défis que Robert Kane (1996, 13) a décrits comme « la montée » et « la descente » de la « montagne libertarienne ». La montée consiste à élever les exigences ou les conditions de satisfaction du concept de libre arbitre, à montrer que l'indéterminisme est nécessaire pour être libre – c'est le défi incompatibiliste proprement dit, la « question de compatibilité », que j'ai essayé de résoudre au chapitre 4. Mais une fois qu'on s'est élevé à ces sommets métaphysiques exigeants, il s'agit de faire redescendre cette condition d'indéterminisme dans le monde réel et de montrer comment l'indéterminisme pourrait concrètement aider à être libre. Ce deuxième défi, cette descente de la montagne, Kane l'a appelée « Question d'Intelligibilité » :

« Pouvons-nous donner un sens à une liberté ou à un libre arbitre qui est incompatible avec le déterminisme ? Une telle liberté est-elle cohérente ou intelligible ? Ou est-elle, comme le prétendent certains critiques, essentiellement mystérieuse et irrémédiablement obscure ? » (Kane 1996, 13)

Kane est un des auteurs qui a le plus contribué à relever ce défi d'intelligibilité du libertarianisme. Pour présenter l'histoire du débat dans les dernières décennies, on pourrait dire (en simplifiant quelque peu) que le débat s'est déplacé progressivement du défi

d'incompatibilité au défi d'intelligibilité. Des années 60 aux années 80 environ, le cœur du débat était la question de compatibilité parce qu'une majorité de philosophes restaient convaincus qu'il n'y avait aucun argument incompatibiliste solide. Puis les arguments incompatibilistes (ceux de Lehrer ou de van Inwagen, notamment) sont parvenus à convaincre un certain nombre d'auteurs qu'il était peut-être, finalement, possible de faire apparaître une réelle tension entre déterminisme et libre arbitre. Certains auteurs ont été jusqu'à admettre la conclusion incompatibiliste dure en raisonnant ainsi : le déterminisme empêche le libre arbitre... mais l'indéterminisme ne peut pas nous aider non plus à l'obtenir, donc le libre arbitre est absolument impossible. Plus généralement, chez tous les auteurs (incompatibilistes durs, compatibilistes ou même libertariens) s'est installée la conviction que le grand défi, le défi vraiment difficile pour le libertarianisme, était le défi d'intelligibilité : comment l'indéterminisme pourrait-il concrètement favoriser la liberté ? Ce tournant, Robert Kane a été un des premiers à le percevoir (et à l'influencer). Mais nous verrons que le type de réponse qu'il a apporté n'est pas le seul à être apparu. Pour simplifier, on peut dire qu'il y a eu deux grandes familles de réponses pour relever le défi d'intelligibilité, c'est-à-dire deux grands « modèles » du libertarianisme : d'une part la réponse très novatrice et « scientifique » de Robert Kane, et d'autre part la réponse plus « traditionnelle » de la « causalité agentive ». Dans ce chapitre (et jusqu'à la fin de cette étude), je ne prendrai pas parti en faveur de l'un ou l'autre modèle, car ils me semblent tous les deux défendables en l'état actuel de nos connaissances. Mon objectif ne sera donc pas de dire exactement comment fonctionne le libre arbitre (je ne crois pas que cela soit possible aujourd'hui), mais seulement de montrer qu'il y a des modèles du libre arbitre concevables et scientifiquement possibles. Mon intention, en effet, est simplement de montrer que les objections d'impossibilité ne sont pas concluantes : il est faux de dire que notre expérience d'un libre arbitre libertarien est l'expérience de quelque chose d'*impossible* ou d'inconcevable ; c'est l'expérience de quelque chose qu'on peut modéliser, et qu'on peut même modéliser de différentes manières (au moins de deux manières différentes). Notre expérience n'est donc pas réfutée de manière directe.

Pour répondre aux objections d'impossibilité, il faut d'abord observer qu'il existe plusieurs objections assez distinctes dans cette grande famille. Certaines objections consistent à dire que la conception libertarienne est *absolument* (ou métaphysiquement) impossible, d'autres au contraire estiment que cette conception serait possible métaphysiquement mais est rendue impossible à cause des informations (contingentes) que nous avons sur la nature du monde dans lequel nous vivons (et en particulier sur la nature des êtres que nous sommes). Les premières sont donc des réfutations *a priori* tandis que les secondes peuvent être appelées *empiriques*. Nous verrons que chacune de ces catégories se subdivise encore en plusieurs objections possibles.

Je commencerai par aborder les objections *a priori* ; je m'efforcerai de montrer que le libre arbitre n'est pas *absolument* impossible, qu'il est logiquement et métaphysiquement modélisable. Puis j'envisagerai les objections empiriques, et je montrerai que les modèles du libre arbitre (certains modèles en tout cas) sont non seulement métaphysiquement possibles, mais aussi compatibles avec les données scientifiques dont nous disposons aujourd'hui sur la nature des agents humains.

I. *Réfutations a priori*

Dans le chapitre précédent, j'ai soutenu que notre concept expérientiel de libre arbitre contenait deux éléments importants, l'élément de source ultime et l'élément de possibilités alternatives. Or parmi les réfutations *a priori*, on peut distinguer entre certaines objections qui situent le problème d'impossibilité ou d'inconcevabilité dans la notion de source ultime, et

d'autres qui situent le problème dans la notion de possibilités alternatives (et plus précisément dans l'idée que les possibilités alternatives pourraient *augmenter* ou *favoriser* notre liberté). Le meilleur exemple du premier type est « l'argument basique » de Galen Strawson (1986, 291). Quant au deuxième type d'objection, il porte dans les débats contemporains le nom « d'argument de la chance » ou « argument du hasard ». Dans cette section, j'envisagerai donc ces deux types d'arguments. Je commencerai par l'argument de la chance, qui permet d'introduire de manière très claire les deux grandes familles de réponse, la réponse de la causalité agentive et la réponse « événementielle » de Robert Kane. Mon intention, dans cette section et la suivante, sera de montrer les ressources de l'un et l'autre modèle pour répondre aux objections d'impossibilité.

I.1. L'argument de la chance (contre la liberté des alternatives)

De manière générale, l'argument de la chance consiste à montrer que si deux actions alternatives sont réellement possibles, s'il n'y a absolument aucun déterminisme dans un sens ou dans l'autre, alors l'action n'est pas libre. Initialement, cette conclusion n'est pas très intuitive : ainsi que nous l'avons vu au chapitre suivant, la menace la plus évidente sur la liberté, c'est que quelque chose *détermine* une des deux alternatives et que nous n'ayons pas réellement le choix ; pourquoi l'existence d'un vrai choix ouvert serait-il un *obstacle* à la liberté ? La réponse tient à l'affinité étroite entre la notion d'indéterminisme et la notion de hasard : s'il est absolument indéterminé que je fasse A ou B, alors il semble que c'est un pur hasard que je fasse A plutôt que B (ou B plutôt que A). Et si c'est par pur hasard que je fais A, alors on ne peut pas dire que je fasse A de manière libre. Plus précisément, si le hasard s'oppose à ce que mon action soit libre, c'est parce que le hasard constitue une *perte de contrôle* à l'égard de mon action, et qu'une action est plus libre lorsque je possède un meilleur contrôle à son égard. On ne voit donc pas comment l'insertion d'un élément indéterministe pourrait augmenter ma liberté d'action puisqu'il ne peut que diminuer mon contrôle. La notion libertarienne est donc incohérente, elle revient à rechercher dans l'indéterminisme un accroissement de contrôle agentif que ce dernier ne saurait procurer.

Telle est l'intuition de base de l'argument de la chance. Elle a pu prendre ensuite diverses formes. Je classerai ces formulations en trois grandes catégories, l'argument traditionnel du dilemme, l'argument du contrôle diminué et l'argument du rembobinage.

I.1.a. L'argument traditionnel du dilemme

Les premières formulations de l'argument de la chance ont été proposées par des auteurs compatibilistes qui cherchaient à établir que non seulement le libre arbitre est *compatible* avec le déterminisme, mais que même il *implique* le déterminisme ou qu'il est « inconcevable sans détermination » (Hobart 1934). Plus précisément, le libre arbitre ne requiert peut-être pas un déterminisme absolu dans toutes les parcelles du monde physique, mais « dans la mesure même où nous avons le libre arbitre » (*ibid.*, 2) nos actions libres quant à elles doivent être causalement déterminées.

La version historiquement la plus remarquable est sans doute celle de Hume :

D'après mes définitions, la nécessité est une composante essentielle de la causalité ; et par conséquent la liberté, en supprimant la nécessité, supprime également les causes, et c'est donc la même chose que le hasard. (Hume 1978, pt. III, sect. 1)

Au cœur de cette objection se trouve donc un dilemme fondamental :

(1) soit une action A est déterminée causalement, soit elle est un pur hasard

Et Hume ajoute en outre un argument pour arriver à cette conclusion :

(1a) Si une action A n'est pas déterminée causalement, elle n'est pas *causée* (du tout)

(1b) Si une action A n'est pas *causée*, elle est un pur hasard.

donc

(1) Si une action A n'est pas déterminée causalement, elle ne peut être que pur hasard.

A partir du dilemme (1), on peut ensuite se diriger vers une conclusion compatibiliste en ajoutant ceci :

(2) si l'action A est un pur hasard, alors elle n'est pas libre.

(C) donc l'action A ne peut être libre qu'en étant déterminée causalement.

Plus récemment, le même dilemme a été utilisé par des auteurs incompatibilistes pour composer l'argument d'impossibilité du libre arbitre :

(3) si l'action A est déterminée causalement, elle n'est pas libre (non plus)

(C) donc l'action A ne peut absolument pas être libre.

Quelle que soit la conclusion, compatibiliste ou impossibiliste, que l'on tire, le résultat est le même en tout cas pour la conception *libertarienne* du libre arbitre : une telle conception est jugée absolument impossible ou incohérente.

Que penser de cet argument du dilemme ? Est-il vrai qu'une action est soit déterministe soit hasardeuse, *tertium non datur* ? De nombreux auteurs ont fait observer que cette alternative avait de bonnes chances de n'être pas exhaustive.³⁶⁴ De fait, cette formulation de l'argument de la chance n'est plus réellement défendue aujourd'hui. Mais elle reste importante car elle permet de poser le problème dans des termes assez clairs : si une action libre (au sens libertarien) n'est ni causalement déterminée, ni hasardeuse, alors *comment* est-elle non hasardeusement causée ? A ce problème bien posé, il existe deux familles de réponses. Les deux familles de réponses consisteront à dire que l'action libre est « causée d'une manière qui n'est pas déterministe », mais c'est la nature de cette « causalité non déterministe » qui les distinguera.

1.1.b. Réponse par la causalité agentive

Une première famille de réponses part de l'intuition suivante : l'action libre, bien qu'elle ne soit pas déterminée, n'est pas non plus « hasardeuse » parce qu'un événement « hasardeux » est un événement qui arrive pour ainsi dire « par lui-même » sans être ni déterminé ni contrôlé, un événement qui se produit « sans raison aucune ». Mais l'action libre n'est pas de cette sorte. L'action libre n'est pas « hasardeuse » tout simplement parce qu'elle est contrôlée par l'agent. Décrire l'action libre comme hasardeuse, c'est mal comprendre le

³⁶⁴ C.A. Campbell est un de ceux qui formulent le plus clairement l'objection sous la forme d'un faux dilemme : « How often, for example, do we find the Determinist critic saying, in effect, 'Either the act follows necessarily upon precedent states, or it is a mere matter of chance and accordingly of no moral significance'. The disjunction is invalid, for it does not exhaust the possible alternatives. » (1957, pt. I, lect. IX, sec. 10). L'alternative que propose Campbell entre dans la famille de la causalité agentive (ou *creative activity* dans sa terminologie).

concept de hasard ou mal comprendre le concept d'action d'un agent. Le dilemme était donc un faux dilemme, car il existe en réalité trois possibilités pour un événement donné : être déterminé par des causes antérieures, être produit par l'agent de manière libre, ou être totalement sans cause et sans raison (c'est-à-dire hasardeux).

Telle est l'intuition de départ, il s'agit ensuite de lui donner une formulation métaphysique plus précise, en particulier pour savoir que répondre au sous-argument de Hume, à savoir :

- (1a) Si une action A n'est pas déterminée causalement, elle n'est pas *causée* (du tout)
- (1b) Si une action A n'est pas *causée*, elle est un pur hasard.
- (1) Si une action A n'est pas déterminée causalement, elle ne peut être que pur hasard.

Faut-il refuser (1a) ou (1b) ? Les actions supposées « produites par l'agent de manière indéterministe » sont-elles donc causées et non déterminées, ou bien non causées et pourtant non hasardeuses ? La réponse traditionnelle est la suivante.

Les actions produites par l'agent de manière indéterministe sont bel et bien *causées*, mais elles ne sont pas causées *dans le sens humien*. Cela veut dire évidemment qu'elles ne sont pas causées *de manière déterministe*, mais plus radicalement, cela veut dire aussi qu'elles ne sont pas causées *de manière événementielle*. Le concept humien de causalité est le concept d'une *connexion nécessaire entre deux (types d')événements A et B*. Par exemple, si le (type d') événement A « choc de la boule rouge par la boule blanche avec telle force » entraîne de manière constante et nécessaire le (type d') événement B « départ de la boule rouge », on peut dire que (une instance du type) A a été la cause de (une instance du type) B. On voit donc qu'il y a deux aspects dans cette définition, l'aspect de nécessité, et l'aspect de la causalité comme *relation entre des événements* : ce qui peut causer un événement, c'est un autre événement. Ce deuxième aspect est traditionnellement appelé l'idée de causalité *événementielle*.

La première réponse au dilemme traditionnel consiste à refuser l'un et l'autre élément de la définition humienne : il est possible pour un événement d'être causé sans être causé de manière indéterministe *notamment parce que* il est possible pour un événement d'être causé de manière non « événementielle », c'est-à-dire d'être causé par autre chose qu'un événement. Par quoi l'action libre est-elle causée si ce n'est par un événement ? Par l'agent lui-même, qui n'est pas un événement mais une substance. C'est ce qu'on appelle la solution par la « causalité de l'agent » ou « causalité agentive ». Cette notion de causalité agentive peut être (et a été) retrouvée ou identifiée dans un passé philosophique très ancien, notamment dans cette célèbre phrase d'Aristote :

Le bâton meut la pierre, et est mue par la main mue par l'homme. (*Physique*, 256a8-9)

Dans cette citation, Aristote semble indiquer une certaine chaîne causale entre événements : le mouvement de la pierre est un événement causé par un autre événement, le mouvement du bâton, lui-même causé par un autre événement, le mouvement de la main. Mais au bout de cette chaîne se trouve une nouvelle relation de « motricité » ou de causalité : le mouvement de la main est causé non pas par un autre événement mais par l'homme, par l'agent. Malgré cela, le moment où émerge vraiment la notion de causalité agentive, en tant qu'alternative métaphysique à la causalité événementielle, est dans la controverse entre Hume et Reid. Après Reid, la notion a été reprise par différents auteurs, en particulier (Chisholm 1964), mais n'a

retrouvé toute son importance que dans le débat récent sur « le problème d'intelligibilité », avec des défenseurs comme (O'Connor 2000) ou (Clarke 2003).³⁶⁵

Comment concevoir cette nouvelle notion de cause agentive ? Si elle est censée apporter une réponse au problème d'intelligibilité, il est crucial qu'elle soit elle-même une notion intelligible. Mais avons-nous vraiment une telle notion ? A ce défi, souvent soulevé contre la causalité agentive, il existe à ma connaissance deux types de réponse : l'une, plus traditionnelle (la solution « reidienne ») défend l'existence et la connaissance d'une notion proprement agentive de la causalité ; l'autre, plus récente dans le débat contemporain (bien que « néo-aristotélicienne »), part d'une métaphysique non-humienne de la causalité en général, à savoir une métaphysique de la causalité par les substances, et consiste à montrer que la causalité agentive entre sans difficulté et sans discontinuité dans ce « nouveau » cadre général.

La réponse traditionnelle est à la fois très forte sur le papier, et pourtant assez délicate du point de vue de la simple persuasion dialectique. L'idée, qui trouve son origine chez Thomas Reid, est la suivante : non seulement nous avons une notion de causalité proprement agentive, mais c'est même notre toute première notion de causalité ; la notion de causalité événementielle est *dérivée* de notre notion de causalité agentive. Comment arrive-t-on à ce résultat ? A partir de la discussion humienne principalement : notre expérience de la causalité événementielle (d'un mouvement de boule causant un autre mouvement de boule) est purement externe comme le fait observer Hume. Ce à quoi nous avons un accès, c'est seulement à la contiguïté spatiale et temporelle des deux mouvements, et pas à *l'influence* d'un mouvement sur l'autre. L'idée d'influence causale ne peut donc pas venir de cette expérience-là. Hume en déduit que nous n'avons pas du tout de notion d'influence causale, mais cette déduction, pour Reid, est contraire aux faits les plus évidents de l'introspection : nous *avons* bel et bien une telle notion. D'où la tirons-nous donc ? De notre propre expérience lorsque nous exerçons notre influence causale sur un mouvement de notre corps. Autrement dit, notre notion même de causalité (d'influence causale par opposition à la simple contiguïté) nous vient de l'expérience de la causalité agentive, et elle est ensuite étendue aux faits de causalité événementielle dont nous n'avons d'expérience qu'extérieure. La réponse traditionnelle repose donc sur une épistémologie de la notion de causalité qui fait de la causalité agentive un primitif par rapport à la causalité événementielle. Cet aspect épistémologique est repris de nos jours notamment par (Lowe 2002, 206 sqq). Reid lui-même en tirait la conclusion qu'il n'existe qu'une seule notion de causalité, à savoir la causalité agentive. La conséquence métaphysique d'une telle réduction de toute la causalité à la causalité agentive, c'est que Reid devait considérer tous les événements naturels comme agentivement causés par Dieu directement. Les auteurs comme Lowe qui reprennent aujourd'hui l'épistémologie primitiviste de Reid ne vont pas aussi loin : ils considèrent au contraire qu'à partir de notre concept de causalité agentive, nous pouvons construire un concept, dérivé mais distinct, de causalité événementielle, de telle sorte que notre image de la réalité contient finalement deux concepts de causalité : la causalité agentive et la causalité événementielle. Cette métaphysique de la causalité « à deux vitesses » est très souvent présentée comme une faiblesse de cette théorie : il serait ontologiquement plus simple, ou moins « coûteux », de travailler avec un seul type de causalité dans la nature ; la notion de causalité agentive, comprise ici comme un type de causalité *sui generis*, revient à ajouter à notre image de la réalité une sorte d'anomalie dont il serait préférable de se passer. Ce manque de simplicité ontologique a amené certains défenseurs de la causalité agentive

³⁶⁵ Pour une excellente historique et critique de la notion de causalité agentive, voir (O'Connor 2000, chap. 3).

(notamment Timothy O'Connor) à passer de cette réponse traditionnelle à la réponse néo-aristotélicienne que nous allons voir à présent.³⁶⁶

La deuxième réponse au défi d'intelligibilité de la causalité agentive consiste à dire que la causalité agentive n'est pas une notion *sui generis* mais qu'elle n'est qu'un cas particulier de notre notion commune de causalité – si celle-ci est bien comprise, c'est-à-dire si elle est comprise comme une notion de causalité *par les substances* (*substance causation*). Cette deuxième réponse repose sur les évolutions récentes en métaphysique de la causalité : l'état de la question aujourd'hui n'est guère comparable avec l'époque de Reid et Hume, ni même avec le début du XX^e siècle. Jusque dans le courant du XX^e siècle, la conception dominante de la causalité naturelle était la conception humienne (événementialiste et régulariste), si bien que toute défense de la causalité agentive était forcément la défense d'un type de causalité qui *rompait* avec le schéma général de la causalité naturelle. Ce n'est pas le lieu d'exposer en détail la richesse et la diversité des progrès et des inventions en philosophie de la causalité au XX^e siècle. Nous en retiendrons simplement quelques étapes « anti-humiennes » qui ont été essentielles pour arriver à une conception de la causalité naturelle en termes de causalité par les substances.

Le premier aspect de la théorie humienne qui ait été mis en question est sans doute l'aspect *régulariste* : pour Hume, un énoncé causal « A cause B » est fondamentalement l'énoncé d'une régularité, c'est-à-dire l'affirmation qu'il existe, ailleurs et à d'autres moments, des événements de type A qui ont été suivis d'événements de type B. Par conséquent, un énoncé causal est un énoncé qui ne concerne pas simplement les deux événements singuliers que l'on décrit, mais qui concerne également un grand nombre d'autres événements éloignés dans l'espace et dans le temps. Cet aspect de la théorie humienne a fait l'objet de critiques dès le début du XX^e siècle, notamment chez Ducasse (1926) : Ducasse s'est efforcé de donner une analyse des énoncés causaux qui soit strictement « singulariste », c'est-à-dire dont les conditions de vérité concernent strictement ce qui arrive aux événements singuliers auxquels l'énoncé fait référence. Bien qu'elle soit « singulariste », il est intéressant de noter que l'analyse de Ducasse préserve l'intuition « contiguïste » de Hume, c'est-à-dire l'intuition selon laquelle un énoncé causal décrit simplement des conditions de contiguïté spatio-temporelle entre des événements, et non pas une quelconque opération « d'influence causal » ou de « pouvoir causal ».³⁶⁷ Ce deuxième pilier de la *doxa* humienne a été remis en question à son tour par la renaissance de la conception aristotélicienne des pouvoirs causaux. Cette renaissance a eu deux sources indépendantes. En Grande-Bretagne, la notion de pouvoirs causaux est réapparue dans les années 70s, dans les travaux de philosophie des sciences de Harré et Madden (1975), puis a été développée par d'autres philosophes des sciences, notamment Nancy Cartwright (1979). Ces études portaient d'une analyse du vocabulaire des théories scientifiques et de l'ubiquité des termes dispositionnels dans ce vocabulaire. Un courant indépendant, et métaphysique plutôt que scientifique, s'est développé en Australie, notamment sous l'influence de C.B. Martin (1993), et a trouvé son expression la plus systématique dans le livre posthume de George Molnar (2006). Ces deux traditions ont commencé, depuis une dizaine d'années, à interagir l'une avec l'autre, notamment dans les travaux de Stephen Mumford (Mumford 2008; Mumford and Anjum 2011). Bien qu'elles aient toutes abandonné la conception simplement contiguïste de Hume, ces théories ont, pour certaines, gardé le vocabulaire de la causalité par les événements – attribuant les pouvoirs

³⁶⁶ Dans (Jacobs and O'Connor 2013), cette première version de la théorie est appelée « the standard agent-causal alternative ». C'est celle que O'Connor défendait notamment dans (O'Connor 2000). A partir de (Jacobs and O'Connor 2013), O'Connor défend la deuxième version que nous allons voir à présent.

³⁶⁷ L'analyse de Ducasse, dans sa version simplifiée, est en effet la suivante : « we may say that *the cause of the particular change K was such particular change C as alone occurred in the immediate environment of K immediately before.* » (Ducasse 1926, sec. 3).

causaux non pas aux agents ou aux substances mais aux événements. Bien que le vocabulaire néo-aristotélicien soit en lui-même adapté à l'adoption d'une métaphysique des substances comme causes, il faut donc distinguer comme troisième étape le moment où l'on passe d'une simple métaphysique des pouvoirs à une métaphysique dans laquelle les pouvoirs appartiennent explicitement aux substances elles-mêmes. Cette étape a été réalisée tout particulièrement par (Lowe 2008).³⁶⁸ Dans une théorie néo-aristotélicienne des pouvoirs causaux, toute instance de causalité est l'exercice d'un pouvoir causal, et ce pouvoir causal dépend de certaines propriétés.³⁶⁹ Par exemple, la propriété qu'a la lame d'être affûtée lui donne le pouvoir causal de trancher le fruit. Tout en disant cela, on peut maintenir une description des instances de causalité en termes de causalité événementielle (par exemple, ce sera le *mouvement* de la lame affûtée qui, en vertu de la propriété, causera la découpe du fruit). Mais il est clair qu'en donnant la place centrale aux propriétés dans l'explication causale, la théorie néo-aristotélicienne suggère naturellement une description dans laquelle c'est le *porteur* de la propriété (donc la substance, la lame elle-même et non pas son mouvement) qui doit être appelée la cause.

Si les instances de causalité naturelle que nous rencontrons quotidiennement sont en fait des cas de causalité par les substances, alors la causalité par une substance humaine (Socrate, vous ou moi) est directement intelligible. En un sens trivial (et pas encore agentif), si je tombe par terre, je peux causer l'écrasement d'une fleur, exactement dans le même sens qu'une pierre causerait l'écrasement de la fleur en tombant au même endroit. Mais je peux aussi causer l'écrasement de la fleur d'une autre manière, à savoir volontairement et dans un certain but. Dans ce cas, le pouvoir causal que j'exercerai sera appelé un pouvoir causal *agentif*. Mais un tel pouvoir n'introduit pas un nouveau sens de la causalité : il s'agit toujours de causalité par une substance en vertu de certaines de ses propriétés. Seulement il s'agit d'une certaine *manière* pour une substance de causer un effet. Si donc on adopte une métaphysique générale de la causalité par substance, la causalité agentive cesse d'être une originalité et un problème.

Les défenseurs de la causalité agentive ont traditionnellement employé un deuxième moyen pour montrer que nous pouvons comprendre cette notion. C'est l'idée que la notion de causalité agentive (quel que soit son rapport avec la causalité naturelle) est donnée dans notre expérience de l'action, c'est-à-dire dans notre phénoménologie agentive au sens où nous l'avons étudiée au chapitre 4. Pour cette raison, les défenseurs de la causalité agentive appuient souvent non seulement leur libertarianisme mais aussi leur *version* particulière du libertarianisme sur notre expérience commune de l'action. Ainsi, Timothy O'Connor que nous avons déjà cité :

[La causalité agentive] est attrayante parce qu'elle capture la manière dont nous expérimentons ordinairement notre propre activité. Il ne me semble pas (au moins ordinairement) que je suis causé à agir par les raisons qui favorisent de le faire ; il semble, plutôt, que *je produis ma propre décision en vue de ces raisons*.
(O'Connor 1995, 196)

³⁶⁸ On la retrouve également dans (Clarke 2003), (Steward 2012), et comme on l'a dit un peu plus haut, dans (Jacobs and O'Connor 2013).

³⁶⁹ Un débat entre néo-aristotéliciens concerne le rapport entre les propriétés en question et les pouvoirs causaux : pour certains (les dispositionnalistes), les propriétés sont intrinsèquement des pouvoirs causaux ; pour d'autres (les relationnalistes), les propriétés n'acquièrent des pouvoirs causaux qu'en vertu de certaines relations extrinsèques dans lesquelles elles se trouvent les unes avec les autres. Sur ce débat, voir notamment (Clarke 2003, sec. 10.1). Comme Clarke, je pense que ce débat n'a pas de conséquences pour la théorie de la causalité agentive, qui peut être développée aussi bien selon l'une ou l'autre réponse. C'est pourquoi je ne l'aborderai pas ici.

Si notre phénoménologie agentive est bel et bien une phénoménologie de la causalité agentive, alors tout ce que j'ai dit jusqu'à présent pour défendre le statut épistémologique du libertarianisme vaut également pour la causalité agentive, et ces éléments *suffisent* à apporter une réponse à l'argument de la chance, au moins dans cette première formulation du dilemme traditionnel. Si nous avons une expérience de l'action libre comme causée par nous, alors il suffit d'utiliser notre répertoire conceptuel le plus commun pour savoir qu'un événement peut n'être ni causalement déterminé ni hasardeux.

Cette défense phénoménologique présente, à mes yeux, une difficulté. C'est que l'interlocuteur peut toujours refuser le fondement phénoménologique sur lequel elle repose. Un interlocuteur peut toujours prétexter qu'il ne fait *pas* l'expérience de causer ses actions d'une manière irréductible à toute causalité événementielle. Si l'interlocuteur refuse ce point d'introspection, pourrions-nous le convaincre par des *arguments phénoménologiques*, comme nous l'avons fait au chapitre précédent pour prouver le contenu incompatibiliste de notre expérience commune ? C'est précisément ce dont je ne suis pas assuré, et ce qui m'amène à ne pas m'engager entre le modèle de la causalité agentive et l'autre modèle que nous allons aborder bientôt. En effet, l'argument le plus fort sur lequel peuvent s'appuyer les opposants de la causalité agentive, c'est l'argument de la réduction. Lorsque nous *disons* que le ballon a causé le bris de la vitre, nous parlons comme si une substance (le ballon) était une cause ; mais en réalité nous *voulons dire* que c'est le *mouvement* du ballon (ou son choc contre la vitre) qui a causé le bris de la vitre. Pourquoi en irait-il autrement pour l'expérience de l'action : lorsque nous *disons* que nous avons causé le mouvement du ballon, ne voulons-nous pas dire que c'est par un certain événement mental (notre désir de shooter) que nous avons causé cet événement physique ? Dans ce cas, nous utiliserions un langage ou une conceptualité où *apparemment* l'agent est cause, mais pour décrire une situation dans laquelle *en réalité* nous décrivons des relations entre événements. Nous aurions donc bien une certaine conceptualité naturelle de la causalité agentive... mais entièrement réductible à la causalité événementielle. Or la réponse de la causalité agentive (au problème du dilemme) repose *fondamentalement* sur l'idée d'une causalité agentive *irréductible* à la causalité événementielle.³⁷⁰ Peut-on prouver que notre concept naturel de causalité agentive est *irréductiblement* agentif ? Certains auteurs le soutiennent.³⁷¹ D'autres soutiennent plus prudemment que le concept irréductible constitue une analyse qui « correspond mieux » à notre phénoménologie. Pour ma part, ce qui me semble évident c'est seulement la chose suivante : que la description de notre phénoménologie en termes agentifs est une description plus fidèle de l'expérience du soi comme source ultime. Mais il ne me paraît pas évident, en l'état actuel des recherches, que cette description doive être forcément *irréductiblement* agentive. Le modèle de la causalité (irréductiblement) agentive a donc *peut-être* un avantage phénoménologique sur son concurrent (que nous allons voir à présent), mais cela ne me paraît pas aujourd'hui évident.

1.1.c. Réponse par la causalité événementielle

La deuxième famille de réponse consiste à dire que l'action libre peut être causée de manière indéterministe *dans le sens le plus général du terme* « cause ». L'idée est donc que le sens général de la causalité ne suppose pas du tout l'idée de détermination. En somme, cette première réponse revient à refuser seulement le premier aspect de l'analyse Humienne de la causalité, en préservant l'aspect de la causalité événementielle : une bonne analyse du concept

³⁷⁰ Autrement, la causalité agentive n'est pas une *alternative* à la causalité événementielle ou à l'absence de causalité, elle est seulement un sous-cas de la causalité événementielle.

³⁷¹ Voir notamment (Lowe 2002, 199).

de cause événementielle ne contient *pas* le concept de nécessité. Un événement A peut causer un événement B même si (le type) A n'est pas corrélé constamment et nécessairement avec (le type) B.

Cette réponse a le très grand avantage d'être appuyée par des considérations absolument indépendantes du problème du libre arbitre. Dans le débat philosophique contemporain, et en particulier en philosophie des sciences, il est aujourd'hui quasi universellement admis qu'il peut exister des formes non déterministes de causalité. Ce point a été établi philosophiquement par G.E.M. Anscombe dans « Causality and Determination » (Anscombe 1971). Nous pouvons très bien concevoir et comprendre que le choc du ballon a causé le bris de la vitre même si l'on nous apprend que le même choc *aurait pu* ou *avait une certaine probabilité de* ne pas causer le bris de la vitre. Ce point philosophique a également été considérablement renforcé par la physique contemporaine, et notamment par la physique quantique dont l'interprétation standard suppose justement des formes indéterministes de causalité.

L'argument de Hume n'est donc pas bon : une action peut très bien être causée sans être causée de manière déterministe. En particulier, il se pourrait que nos actions libres soient causées par nos événements de désirs (et de croyances) d'une manière indéterministe au même sens où la physique quantique (dans l'interprétation standard) soutient que des événements quantiques peuvent en causer d'autres de manière indéterministe. Nous verrons plus loin *comment* cela pourrait se réaliser dans la nature telle que nous la connaissons – et en particulier nous verrons quel rôle les indéterminismes quantiques *eux-mêmes* peuvent jouer pour faire apparaître de l'indéterminisme naturel dans l'action – mais pour l'instant, le problème que nous traitons est purement un problème de concevabilité *a priori* sur la notion de cause. Et *a priori*, rien n'empêche de concevoir une action qui serait causée de manière indéterministe par nos raisons (plus nos croyances). Ce modèle, qui s'inspire largement de la physique quantique pour défendre une causalité événementielle indéterministe, est le modèle défendu notamment par Robert Kane (1996).

Si l'on admet un tel modèle, a-t-on répondu totalement à l'objection du dilemme traditionnel ? Ce que montre cette réponse, c'est qu'on ne peut pas considérer une action indéterministe comme hasardeuse *au titre qu'elle serait non causée* ; une action indéterministe peut être causée de manière indéterministe. Donc le sous-argument Humien en faveur du dilemme ne marche pas. Mais ne peut-on pas néanmoins garder une forte intuition du dilemme ; ne peut-on pas garder l'intuition qu'une action causée de manière événementielle est soit causée de manière déterministe, soit causée d'une manière totalement hasardeuse ? Auquel cas, l'action libre ne serait pas hasardeuse *en tant que non causée*, mais simplement *en tant que causée de manière indéterministe*. Après tout, si l'on pense aux événements quantiques causés de manière indéterministe, n'a-t-on pas justement l'intuition que le résultat de tels événements est purement hasardeux ? L'intuition du dilemme peut donc survivre à l'idée de causalité indéterministe.

Comment Robert Kane répond-il à cette intuition qu'un événement indéterministe, même *causé*, serait purement hasardeux ? La réponse de Kane consiste essentiellement à montrer que l'action indéterministe n'est pas « simplement hasardeuse » comme un événement quantique, parce qu'elle est causée *par des raisons* et qu'elle aurait également été causée *par des raisons* si l'autre choix possible avait été fait. Prenons en effet l'exemple d'un événement quantique, la désintégration possible (mais non déterminée) au temps *t* d'une particule élémentaire *e*. Dans un tel événement, quel que soit le résultat (la désintégration ou la non désintégration), le fait que ce résultat se soit produit aura été purement fortuit ou hasardeux, bien qu'on puisse établir les antécédents causaux de cette situation (et par exemple un certain événement qui pouvait causer la désintégration ou ne pas la causer). Imaginons à

présent un agent S qui délibère pour savoir s'il va lever le bras ou non, et imaginons que son choix soit indéterministe ; à la fin de la délibération, S choisit de lever le bras sans aucune raison. Dans ce cas aussi, on pourra dire que son choix était purement hasardeux. Mais imaginons que S ait des *raisons* de choisir de lever le bras, et que ce soient *ces raisons* qui causent (de manière indéterministe) son lever de bras. Dans ce cas, on ne sera plus autant tenté de dire que son choix était un pur hasard : son choix au contraire était un choix rationnel. Il n'a pas choisi de lever le bras *par hasard*, mais (par exemple) *parce qu'il voulait prendre la parole*.

Cela n'est peut-être pas encore totalement suffisant cependant ; on pourrait en effet objecter qu'il y a un hasard de deuxième ordre. S'il avait des raisons de lever le bras, c'est un hasard qu'il ait agi pour ces raisons plutôt que pour aucune raison (et donc de manière hasardeuse). On se place alors dans un cas où S n'avait *que* des raisons de lever le bras, et *pas* de raisons de ne pas le lever. Dans ce cas, on peut en effet estimer que l'élément d'indétermination est un pur effet de hasard. C'est pour répondre à cette situation que Kane a produit le concept de « rationalité duale » (Kane 1996, 107) : pour qu'une action ou plutôt un choix indéterministe soit non hasardeux, il faut qu'il soit causé par des raisons de faire ce choix *dans l'une et l'autre alternative*. Par exemple, l'agent S avait des raisons personnelles de lever la main pour prendre la parole (il voulait faire l'intéressant), et des raisons morales de ne pas lever la main (cela gênerait la progression du cours). S'il choisit de lever la main, son choix aura été causé (de manière indéterministe) par son désir de faire l'intéressant ; s'il choisit de ne pas lever la main, son choix aura été causé (de manière indéterministe) par sa motivation morale de ne pas gêner le cours. Dans l'une et l'autre alternative, son choix est fait *pour des raisons* et non pas simplement *par hasard*, et qui plus est l'élément d'indéterminisme provient ici *essentiellement* de la présence de motivations contradictoires. Elle ne provient pas, par exemple, d'un risque probabiliste d'échouer à agir de manière rationnelle.

Pour qu'il y ait vraiment rationalité duale, une dernière précision est indispensable, c'est qu'il doit y avoir incommensurabilité entre les raisons qui entrent en conflit pour le choix. En effet, si les deux raisons de l'agent étaient commensurables – par exemple, si sa raison de ne pas lever le bras était seulement une raison prudentielle, pour se faire le plaisir d'être mieux apprécié du professeur en ne gênant pas le cours, un plaisir commensurable au plaisir de faire l'intéressant – alors les deux raisons pourraient être pesées ensemble, et l'agent S aurait une unique « raison contrastive » : par exemple, malgré sa raison de lever la main, il serait plus rationnel en ne la levant pas plutôt qu'en la levant, pour s'attirer un plus grand avantage. Dans ce cas, l'une des deux actions ne *pourra pas* être choisie de manière pleinement rationnelle. Mais si le conflit a lieu entre une raison prudentielle et une raison morale, alors ces deux raisons seront incommensurables ; le fait de les comparer ne permettra de faire émerger aucun contraste en faveur de l'une ou l'autre. Donc S aura non seulement, au premier niveau, une raison de lever la main et une raison de ne pas la lever, mais il aura aussi, au niveau comparatif, une raison (prudentielle) de la lever *plutôt que* de ne pas la lever pour des raisons morales, et une raison (morale) de ne pas la lever *plutôt que* de la lever pour des raisons prudentielles. Ainsi, les deux raisons étant incommensurables, l'agent S agira de manière pleinement rationnelle quel que soit le choix qu'il fera finalement.

Cette théorie de la rationalité duale permet de montrer qu'une action libre au sens de Kane n'est pas comparable à un événement quantique causé d'une manière indéterministe qu'on pourrait dire *hasardeuse*. En effet, une action libre est causée de manière indéterministe mais causée *par des raisons* et ce sont ces raisons elles-mêmes (par leur conflit) qui introduisent un élément d'indétermination dans cette *causalité pour des raisons*.

Tel est le deuxième grand type de réponse à l'argument du dilemme : une action peut n'être ni déterministe ni hasardeuse en étant causée de manière indéterministe par des raisons.

Quels sont les avantages et les inconvénients de ce modèle par rapport au modèle de la causalité agentive ? Nous verrons dans les sections suivantes certains avantages et inconvénients *pour répondre* à telle ou telle objection. Pour ce que nous avons vu pour l'instant, nous pouvons dire simplement ceci : le modèle de la causalité événementielle n'a sans doute pas l'avantage phénoménologique que la causalité agentive réclame pour elle-même. Mais comme on l'a vu, cet avantage n'est pas absolument évident. Par ailleurs, un avantage majeur du modèle de Kane, c'est son *économie* : il ne suppose pas d'ajouter dans notre métaphysique un certain type de causalité. Il fait fonds sur les éléments métaphysiques dont on dispose *indépendamment* du problème du libre arbitre. C'est là sûrement un « intérêt » dialectique pour répondre aux incompatibilistes ou aux compatibilistes, mais je ne crois pas que cela suffise, en l'état, pour préférer ce modèle à celui de la causalité agentive.

1.1.d. L'argument du contrôle diminué

Maintenant que nous connaissons les deux grands modèles du libertarianisme, voyons comment ils peuvent répondre aux autres objections d'impossibilité qu'on a soulevées contre cette théorie. Et tout d'abord, comment permettent-ils de répondre aux autres versions de l'argument de la chance ?

J'ai dit un peu plus haut que l'argument traditionnel du dilemme était largement abandonné aujourd'hui, parce qu'un grand nombre d'auteurs reconnaissent qu'une action pourrait très bien à la fois contenir un élément d'indétermination et pourtant n'être pas purement hasardeuse. Il n'y a donc pas une alternative absolue entre « être déterminé » et « être hasardeux ». Mais l'intuition de base de l'argument de la chance peut être formulée d'une manière qui n'est pas si facile à éliminer. Certes une action peut « contenir de l'indéterminisme » et être néanmoins rationnelle et contrôlée ; mais n'est-il pas évident que *dans la mesure* où elle est rationnelle et contrôlée, cette part de rationalité et de contrôle est due à des éléments déterministes ? Ou pour le dire dans l'autre sens, qui est sans doute plus clair : comment les éléments d'indéterminisme présents dans l'action pourraient-ils être la source d'une *augmentation* de rationalité ou d'une augmentation de contrôle ? Pour autant qu'on ajoute de l'indéterminisme, n'est-il pas évident que cela tendra à *diminuer* le contrôle ? C'est là l'argument du « contrôle diminué » selon l'expression de (Clarke 2003, chap. 5), qui est bien plus prudent que l'objection du dilemme, et bien plus difficile à rejeter. Un auteur qui a particulièrement bien vu l'importance de cet argument « quantitatif » à la place de l'argument « absolu » de Hume ou Hobart, est Mark Balaguer (2010, 66).

Comme le montre bien Balaguer, cet argument repose fondamentalement sur l'idée que les notions de *hasard* et de *contrôle* sont antinomiques et que par conséquent l'un ne peut augmenter ou favoriser ou permettre l'autre. L'idée de Balaguer est qu'un auteur libertarien doit postuler des indéterminismes *pour* permettre la liberté (le type libertarien de liberté). Par conséquent, ces indéterminismes doivent être capables de permettre, d'augmenter, ou de favoriser un certain type de liberté. Mais comme une action ne saurait être libre sans être contrôlée, sans être « non-aléatoire », cela veut dire que les indéterminismes en question doivent permettre, augmenter ou favoriser un certain type approprié de contrôle ou de « non-aléatoire ». Or, comme le dit Balaguer :

Il se pourrait qu'il existe, dans nos processus de prise de décisions, des indéterminismes qui ne détruisent pas totalement le non-aléatoire approprié, et donc l'argument de Hobbes-Hume-Hobart est trop fort. Néanmoins, il est difficile de voir comment il pourrait y avoir dans ces processus des indéterminismes qui *gènèrent* ou *favorisent* ou *augmentent* le non-aléatoire approprié. Le non-aléatoire approprié a à voir avec le fait

que l'agent est au contrôle. Comment cela pourrait-il être augmenté par l'addition d'un événement indéterministe dans un processus de prise de décision ? Il semble que lorsque nous ajoutons un événement indéterministe dans un tel processus, nous ajoutons un élément d'*aléatoire*. Donc comment cela pourrait-il augmenter le *non-aléatoire* ? Il semble que ce soit impossible, et donc il semble que le libertarianisme ne pourrait pas être vrai. (Balaguer 2010, 66)

L'argument de la chance dans sa version du « contrôle diminué » consiste donc à dire qu'il ne semble pas possible d'insérer des indéterminismes dans le processus décisionnel d'une manière qui augmente le contrôle. On pourrait le lire de manière forte comme l'intuition qu'il est absolument impossible qu'un ajout d'indéterminisme augmente le contrôle, mais dans une lecture plus modeste, il s'agit plutôt d'un « défi » consistant à dire ceci : « toutes les manières d'ajouter un indéterminisme que je peux imaginer diminuent le contrôle au lieu de l'augmenter ; pouvez-vous trouver une manière qui intuitivement augmente le contrôle ? »

Pour répondre à cet argument, je procèderai en deux temps. Je commencerai par écarter les « mauvaises » manières d'insérer un indéterminisme dans le processus décisionnel, les manières qui, effectivement, diminuent le contrôle au lieu de l'augmenter. La seule manière dont l'indéterminisme puisse éventuellement augmenter le contrôle, c'est en l'insérant au moment même du choix, pas avant, pas après. Deuxièmement, il faudra voir en quoi un indéterminisme bien placé peut être dit augmenter le contrôle de l'agent. J'envisagerai la réponse originale de Balaguer à cette question, qui ne me semble pas entièrement satisfaisante, puis j'envisagerai les réponses possibles dans le modèle de la causalité agentive et dans le modèle de Kane.

Il est évident que *certaines* manières d'insérer de l'indéterminisme dans un processus décisionnel ont pour conséquence de diminuer le contrôle en ajoutant de l'aléatoire. Une première manière consisterait à insérer l'indéterminisme *avant* le moment de la décision ou du choix, une autre manière consisterait à l'insérer *après* le choix, c'est-à-dire entre la décision et sa mise en pratique.

La première solution a été positivement proposée comme un type de libertarianisme par divers auteurs.³⁷² Randolph Clarke (2000) l'a baptisée « libertarianisme délibératif », bien que son statut de « libertarianisme » soit compliqué, comme on va le voir. L'idée est la suivante : au moment où l'agent délibère pour faire un choix, au moment où il considère en son esprit les différentes raisons de faire le choix A ou de faire le choix B, il se pourrait qu'il se trouve des indéterminismes dans le processus qui l'amène à prendre telle donnée ou telle autre en considération. Par exemple, il se pourrait que lorsque je délibère sur l'achat de telle ou telle voiture, il y ait un moment où mon cerveau *pourrait* me faire prendre en considération le fait que mes enfants préféreraient tel modèle, et il *se pourrait* également que mon cerveau ne m'amène pas à prendre cela en considération. L'indéterminisme en question concerne donc « ce qui me vient (ou ne me vient pas) à l'esprit ». ³⁷³ Pourquoi cette forme d'indéterminisme a-t-elle pu être considérée comme une forme de libertarianisme ? Parce que l'indéterminisme ainsi situé peut avoir une répercussion sur la décision finalement prise. Il se peut, par exemple, que *si* les préférences de mes enfants me viennent à l'esprit, alors je choisirai la voiture A, et *si* les préférences de mes enfants ne me viennent pas à l'esprit, alors je choisirai la voiture B. En ce sens, il est donc vrai que ma décision elle-même, mon choix final, est indéterminée, ainsi que le veulent les tenants d'un libre arbitre indéterministe. Par ailleurs, il semble bien

³⁷² Notamment par Daniel Dennett qui l'envisage à titre hypothétique (Dennett 1978), mais aussi par des auteurs plus ouverts au libertarianisme comme Alfred Mele (1999).

³⁷³ L'indéterminisme délibératif peut évidemment aussi concerner à quel moment telle considération me vient à l'esprit s'il est déterminé que je vais penser à X, mais indéterminé à quel moment je vais y penser.

que ma décision soit indéterminée d'une manière qui ne remet pas en cause ma liberté : on peut très bien concilier cet aléa dans mon processus délibératif et continuer de dire que j'ai librement choisi la voiture A. Après tout, j'ai finalement pris ma décision de manière rationnelle à partir des considérations qui me sont venues à l'esprit ; on ne voit donc pas pourquoi cet aléa aurait *supprimé* ma liberté. On a donc une forme d'indéterminisme du choix qui ne supprime pas le contrôle nécessaire à la liberté. Mais a-t-on pour autant une forme d'aléa qui *augmente* ou *permet* ou *favorise* mon contrôle et ma liberté ? La réponse est clairement non : cet indéterminisme ne retire pas le contrôle de l'action en fonction de la décision prise, ou le contrôle de la décision en fonction des considérations envisagées, mais il n'ajoute aucune sorte de contrôle ; là où il apparaît, en amont des considérations envisagées, il ne fait que *diminuer* le contrôle que j'ai, puisqu'il m'enlève tout véritable contrôle des considérations envisagées. Ce n'est pas moi qui choisit les décisions qui me viennent ou non à l'esprit, c'est simplement un processus indéterministe (une roulette pour ainsi dire insérée dans mon cerveau) qui détermine de manière hasardeuse les considérations qui me viennent à l'esprit. Cette théorie ne fait donc pas apparaître un nouveau type de contrôle ou un type particulier de libre arbitre, et en ce sens, on ne peut pas réellement l'appeler un type de *libertarianisme* si le libertarianisme est l'idée d'un certain type de liberté qu'on ne peut avoir qu'avec l'indéterminisme.³⁷⁴

Passons à la deuxième « mauvaise » manière d'insérer de l'indéterminisme dans le processus décisionnel. Cette deuxième manière consiste à insérer l'indéterminisme *après* la décision, c'est-à-dire dans le moment qui mène de la décision prise à sa mise en pratique par l'action. Supposons par exemple qu'après avoir délibéré et choisi d'acheter la voiture A, il se trouve certains événements indéterministes qui peuvent m'amener à *ne pas* mettre en œuvre cette décision. On peut imaginer par exemple que j'aie une certaine probabilité d'oublier ma décision et d'en reprendre une autre en sens contraire. Supposons par ailleurs que finalement, malgré cet aspect indéterministe, je mette tout de même ma décision en pratique. Dira-t-on que l'élément d'indéterminisme a totalement supprimé ma liberté ? Sans doute pas. On a donc une forme d'indéterminisme qui concerne le fait de faire une action ou de ne pas la faire, et qui ne supprime pas (pas absolument en tout cas) le contrôle nécessaire à la liberté. Dira-t-on pour autant qu'on a là une forme de libertarianisme ? Non, car de nouveau il est clair que cet indéterminisme *n'augmente* ni ne *favorise* ni ne *permet* aucun nouveau type de contrôle. Au moment où il intervient, le seul contrôle pertinent est celui de mon action par ma décision, et *ce* contrôle-là, même si il ne le supprime pas totalement, il est clair qu'il ne peut que le *diminuer*. En effet, si j'étais assuré de réaliser toutes mes décisions, il est clair que j'aurais un meilleur contrôle de mes actions par mes décisions que s'il y a ce type d'indéterminisme.

A quel moment l'indéterminisme doit-il être situé pour pouvoir « aider » la liberté ? Exactement au moment où notre phénoménologie de l'action le place, d'après les résultats du chapitre 4, c'est-à-dire au moment du choix lui-même ou de la décision elle-même. Tout indéterminisme avant ou après le choix ne fera que diminuer notre contrôle, mais cela n'a pas d'importance car ces indéterminismes ne sont *pas* ce que notre phénoménologie nous dicte. Quant à l'indéterminisme *au moment* du choix, il est loin d'être aussi évident qu'il soit « incapable d'augmenter notre contrôle ». Nous allons voir dans un moment qu'on peut même comprendre en quoi cet indéterminisme *permet* un certain type de contrôle. Avant cela, nous pouvons éliminer une autre forme d'objection qui est très liée aux considérations présentes.

Ces considérations sur la « mauvaise insertion de l'indéterminisme » sont pertinentes pour répondre à ce qu'on présente parfois comme une version particulière de l'argument de la chance (voir Franklin 2011, sec. 6), à savoir « l'argument de l'assurance ». L'argument

³⁷⁴ C'est là la définition de « libertarianisme » retenue par Balaguer.

soutenu en particulier par (Haji 2001) est le suivant : si un agent accomplit à t une action indéterministe A, alors il n'aurait pas pu *s'assurer* ou *garantir* qu'il ferait A plutôt que B à t . Si donc l'indéterminisme a pour conséquence que l'agent ne pouvait pas *s'assurer* de quelle action il accomplit, alors il semble clair qu'il diminue le contrôle, et donc qu'il n'introduit pas un nouveau type de liberté. Mais la réponse à cet argument est très simple dès qu'on la rapproche des considérations précédentes sur le bon moment de l'indéterminisme. Il suffit en effet de se demander à *quel moment* l'agent S est supposé être privé de la possibilité de « s'assurer » ou de « garantir » qu'il fera A plutôt que B. Si on se place *après* le moment de la décision, alors dans le modèle libertarien que nous venons de décrire, l'indéterminisme pertinent est déjà *passé*, donc l'agent normalement est assuré à ce stade d'appliquer la décision qu'il a prise. Si l'on se place *avant* le moment de la décision (donc pendant la délibération et, supposons, à un moment où l'agent n'a pas encore la possibilité de faire cesser la délibération en prenant une décision), alors il est clair qu'à ce stade l'agent ne peut pas s'assurer qu'il fera A plutôt que B (en décidant A plutôt que B) parce que l'indéterminisme de sa décision est encore à venir et qu'il ne peut pas, pour l'instant, le résoudre. Mais cette absence « d'assurance » est-elle problématique ? Qu'est-ce qu'elle signifie exactement ? Elle signifie simplement que nos décisions libres *futures* sont vraiment libres et vraiment futures ; qu'elles ne m'appartiennent pas à moi *maintenant* et que je ne peux donc pas, pour l'instant, les « préempter ». Mais loin d'être un problème, il semble que cela soit une part essentielle de la notion de liberté : si j'avais, avant le temps t , un moyen de manipuler mon moi futur de telle sorte que je m'assure de décider que A, alors il est clair que mon moi futur ne serait pas *libre* dans sa décision de faire A. Pour que ma décision de faire A soit pleinement libre à t , il faut qu'elle soit une décision prise à t et non pas une décision manipulée (pas même par l'agent lui-même) à $t-1$. Autrement dit, un défenseur du libertarianisme n'a aucune raison d'être soucieux de cette absence de « contrôle antécédent » ou « d'assurance antécédente ». Ce qui peut être problématique, ce n'est donc pas l'avant ni l'après de la décision (avant, il n'a pas de contrôle antécédent, mais ne doit pas en avoir ; après, il a l'assurance de faire ce qu'il a décidé) ; ce qui peut être problématique, c'est le moment même de la décision : peut-on dire que, dans un modèle libertarien, l'agent ne peut à t s'assurer qu'il décide à t de faire A plutôt que B ? La réponse est clairement non. A t , il y a quelque chose de très simple que peut faire A pour s'assurer qu'il décide à t de faire A à t : il peut décider de faire A. Qu'est-ce en effet que « s'assurer de faire A à t » ? C'est faire quelque chose X telle que si S fait X, alors nécessairement S fera A à t . Nous avons vu qu'avant t , il n'y avait rien que S puisse faire tel que si S le faisait alors nécessairement S fera A à t . Mais si l'on se situe à t , au moment où S peut décider de faire A ou de faire B, il est clair qu'il peut faire quelque chose (A) telle que s'il le fait (à t), alors nécessairement (de nécessité logique) S fait A à t .

Le problème de l'assurance n'est donc qu'une autre manière de mal comprendre le bon emplacement de l'indéterminisme dans un modèle libertarien, à savoir au moment même du choix. Reste à voir en quel sens un indéterminisme placé au moment même du choix pourrait concevablement *augmenter* ou *permettre* un type de contrôle, alors que l'indéterminisme semble augmenter l'aléatoire et donc s'opposer au contrôle.

Même si on élimine les mauvais modèles de l'indéterminisme pré-décisionnel et post-décisionnel, il demeure une certaine tension entre les notions de contrôle et d'indéterminisme dans la mesure où l'indéterminisme est une forme d'aléatoire et que l'aléatoire ne semble pouvoir que diminuer notre contrôle. On peut résumer cette intuition par la métaphore de la « roulette » indéterministe : si le libertarianisme consiste à intégrer dans notre cerveau une certaine roulette indéterministe, il y a sans doute des mauvais endroits où l'insérer, mais comment pourrait-il y avoir un *bon* endroit pour l'insérer, un endroit qui permet d'augmenter

notre contrôle ? Comment résoudre cette tension ? Je distinguerai ici les ressources du modèle de la causalité agentive et du modèle de Kane.

Commençons par la théorie de la causalité agentive. En quoi l'insertion de l'indéterminisme au moment du choix pourrait-elle permettre ou augmenter le contrôle de l'agent ? Dans la théorie de la causalité agentive, la réponse est simple : sans cet indéterminisme du choix, il existe un type de contrôle particulier qui serait impossible, à savoir précisément le contrôle des alternatives par causalité agentive. La causalité agentive constitue un *pouvoir* dont dispose l'agent de contrôler des alternatives qui s'ouvrent à lui, et comme on l'a vu, ce pouvoir est irréductible à toute forme de causalité événementielle telle que le pouvoir causal de nos raisons sur nos choix, de nos choix sur nos actes, ou encore de nos raisons de deuxième ordre sur nos raisons de premier ordre. Il s'agit d'un pouvoir non analysable de l'agent directement sur une alternative qui s'ouvre à lui, le pouvoir de déterminer laquelle des deux branches de l'alternative s'actualisera, c'est-à-dire le pouvoir de *choisir* au sens métaphysiquement fort du terme.

Comment cette réponse permet-elle de remédier à notre embarras d'origine, c'est-à-dire au fait que l'ajout d'une roulette aléatoire serait censée augmenter le caractère contrôlé et donc non-aléatoire ? Le cœur de la solution est le suivant : ce n'est pas l'ajout d'indéterminisme *en lui-même*, ou *tout seul*, qui augmente ou permet le contrôle de l'action ; seulement l'ajout d'un indéterminisme est une condition *nécessaire* (non suffisante) pour que soit applicable le pouvoir agentif, un type irréductible de pouvoir qui ne peut s'appliquer que sur des alternatives ouvertes. C'est donc l'ajout du pouvoir agentif qui « augmente » le contrôle de l'agent, ou plutôt, c'est ce pouvoir qui *constitue un nouveau type de contrôle* par rapport à toutes les formes de contrôle qu'on peut trouver dans une action purement déterministe. Si l'on se contentait d'ajouter une roulette indéterministe à une action initialement purement déterministe, il est vrai (pour le théoricien de la causalité agentive) qu'on ne ferait que diminuer le contrôle de l'agent. Mais la causalité agentive en tant que telle ajoute en outre un contrôle *sui generis* et métaphysiquement irréductible, un type de contrôle qu'on ne peut ajouter *que* si l'on est en présence d'un indéterminisme.³⁷⁵

Passons à présent aux théories « à la Kane », qui reposent sur une causalité événementielle indéterministe. Ces théories, on l'a vu, ont « l'avantage » théorique de ne pas postuler de forme nouvelle ou *sui generis* de causalité ou de contrôle... mais cet avantage constituera manifestement un inconvénient pour répondre à l'objection du contrôle diminué. En effet, en l'absence d'un type métaphysiquement original de contrôle, les théoriciens tels que Kane seront obligés d'admettre que c'est le simple ajout d'indéterminisme *en tant que tel* qui permet ou augmente le contrôle, c'est-à-dire que c'est l'insertion de la roulette indéterministe dans le processus décisionnel qui est en elle-même une condition *suffisante* pour augmenter ou permettre le contrôle de l'agent. L'objection du contrôle diminué se repose donc ici de manière plus radicale.

Pour bien voir le problème, il faut insister sur le fait que Kane (et ceux qui adoptent le même genre de modèle) a recours, fondamentalement, à des types de contrôle que le compatibiliste peut également satisfaire. Dans sa théorie de la rationalité duale, ou du contrôle dual, Kane soutient que l'agent a deux raisons contraires et que ces deux raisons contraires l'amènent à faire deux efforts contraires et parallèles. Quel que soit le choix finalement effectué, il aura été effectué à cause d'une raison, et à la suite d'un effort pour effectuer ce

³⁷⁵ Cette réponse est essentiellement celle que donne Timothy O'Connor à une version légèrement différente de l'argument : « The agent causationist contends that these objections fail to take seriously the concept of agent causation. It is conceived as a primitive form of control over just such undetermined, single-case outcomes. » (O'Connor 2011, 324).

choix. C'est *cela* qui permet de dire que son choix sera *contrôlé* quel que soit le choix effectué. Mais le contrôle compris comme « décision prise pour une raison suite à un effort pour faire valoir cette raison » peut tout aussi bien avoir lieu dans un contexte entièrement déterministe lorsque l'agent n'a pas de raisons contradictoires mais seulement une raison prévalente dans un seul sens. Autrement dit, le conflit des raisons et l'indéterminisme « permettent » le caractère dual du contrôle, mais ils ne permettent pas un *type* de contrôle fondamentalement nouveau par rapport à une décision « compatibiliste » (il s'agit toujours de contrôle de la décision par les raisons en vertu d'un effort). N'est-il pas clair, par conséquent, que cet indéterminisme ne fait qu'ajouter de l'aléa et donc *diminue* le contrôle (bien qu'il ne le supprime pas).

La réponse satisfaisante à ce problème a été apportée par Mark Balaguer (2010, sec. 3.3.2). Son idée consiste à se focaliser sur les « choix difficiles » (*torn decisions*) au sens phénoménologique, ceux pour lesquels nous n'avons pas de raison prévalente en faveur de l'une ou l'autre option. Il est clair que tout être humain normal a l'occasion de faire de temps en temps cette expérience phénoménologique. La question est alors la suivante : ne perdriions-nous pas une forme de contrôle si *ces* choix étaient entièrement déterministes ? Pour Balaguer, la réponse est clairement positive. En effet, si un choix difficile était entièrement déterministe, cela voudrait dire qu'il y a certains facteurs causaux qui entrent dans la sélection de la décision finale *et qui ne sont pas représentés mentalement comme raisons prévalentes*. Parmi les facteurs causaux qui influencent une décision, certains seront normalement représentés au niveau mental comme *raisons* de faire tel ou tel choix. Dans les choix confiants, un facteur causal *déterminant* pourra notamment être représenté par une raison *déterminante* (par exemple, la raison déterminante de sauver ma vie en donnant mon portefeuille). Mais dans les cas de choix difficile, *aucune* des raisons n'est phénoménologiquement déterminante ou prévalente ; par conséquent, si un certain facteur causal est déterminante, il ne sera pas représenté mentalement par une raison déterminante. C'est pour cette raison que le choix ne sera pas pleinement contrôlé comme on va le voir. Supposons tout d'abord que le facteur déterminant ne soit pas représenté *du tout* au niveau mental. Il s'agira par exemple d'un facteur causal purement biologique ou bien d'une motivation totalement inconsciente. Dans ce cas, on se trouvera dans une situation de « court-circuit », une situation dans laquelle le facteur déterminant du choix est un facteur non mental. Or nous avons déjà vu que ces situations de court-circuit ne sont pas des situations de contrôle ou de liberté *même du point de vue d'un compatibiliste*. Supposons à présent que le facteur causal déterminant soit représenté comme raison dans la phénoménologie de l'agent, mais de nouveau comme une raison non déterminante ou non prévalente. Par exemple, l'agent ressentira un désir de choisir la vanille et un désir de choisir la fraise, les deux seront pour lui phénoménologiquement aussi forts, mais du point de vue neuronal, le désir de vanille aura des caractéristiques causales qui le détermineront à emporter la décision. Dans ce cas, on ne se trouvera pas dans une situation grossière d'illusion par court-circuit, mais il est clair qu'il y aura tout de même une *illusion* de la même nature : l'agent aura *l'impression* qu'aucune motivation ne prévaut, alors qu'*en fait* l'une d'elles prévaut. Ou pour le dire autrement, même si tout *facteur causal* sera représenté par un *facteur rationnel*, il n'en restera pas moins que le *poids causal* des différents facteurs ne correspondra pas à leur *poids rationnel*. Et par conséquent, l'explication du choix au niveau de la base causale neuronale sera plus complète et prévaudra sur l'explication du choix au niveau rationnel. A ce titre, on pourra donc dire que le « surpoids » causal de l'une des deux raisons constituera une diminution du « contrôle par les raisons » ; l'action ne sera pas « entièrement » ou « ultimement » déterminée par les raisons. Là encore, c'est en se fixant sur une notion compatibiliste de contrôle qu'on pourra conclure que ce choix difficile ne sera pas pleinement contrôlé par l'agent, et ce *à cause* du déterminisme.

Comparons maintenant cette situation avec la situation où l'on « insère » un indéterminisme au niveau du choix (difficile). Par exemple, pour le choix entre vanille et fraise, on fera en sorte que, d'un point de vue neuronal, le poids causal des deux facteurs soit exactement le même (et donc qu'il y ait réellement indéterminisme causal entre les deux issues). Cette insertion d'indéterminisme augmentera-t-elle le contrôle de l'agent ? La réponse est oui, pour la raison suivante : le fait d'insérer cet indéterminisme ne consistera pas simplement à insérer une roulette n'importe où dans le processus décisionnel ; cette insertion d'indéterminisme consistera à rééquilibrer le poids causal des facteurs du choix *pour le faire correspondre* au facteur rationnel de ces mêmes facteurs au niveau phénoménologique. Autrement dit, si on compare le choix difficile déterministe et le choix difficile indéterministe, c'est le second qui est plus « contrôlé » parce que c'est le second qui satisfait la pleine correspondance entre facteurs mentaux et facteurs causaux, supprimant ainsi toute forme de « court-circuit » (qu'elle soit absolue ou simplement quantitative). C'est donc bien l'ajout d'indéterminisme *en tant que tel* qui permet un véritable contrôle *dans le cas des choix phénoménologiquement difficiles*, parce que ces choix sont phénoménologiquement indéterminés et qu'il y aurait donc une forme d'*illusion* agentive à ce qu'ils ne soient pas *réellement* indéterministes.

Cette solution de Balaguer est ingénieuse. Il faut bien comprendre cependant qu'elle ne remplit pas le même cahier des charges que la solution par la causalité agentive. Si l'une et l'autre solution « augmentent » le contrôle de l'agent, ce n'est pas par rapport au même point de référence. En effet, la causalité agentive augmente le contrôle de l'agent *par rapport* à une action contrôlée de manière simplement compatibiliste : ainsi, un choix difficile indéterministe et contrôlé par pouvoir agentif est plus au contrôle de l'agent qu'un choix confiant contrôlé de manière simplement compatibiliste. Il n'en va pas de même pour la réponse de Balaguer : pour lui, l'indéterminisme augmente le contrôle simplement *par rapport* à une situation de choix difficile déterministe. Un choix difficile indéterministe est plus au contrôle de l'agent qu'un choix difficile déterministe (et partant illusoire)... mais il n'est pas plus au contrôle de l'agent qu'un choix confiant déterministe (dans lequel des raisons prévalentes correspondent aux facteurs causaux déterminants). Au contraire, il reste probable qu'un choix difficile indéterministe soit *moins* contrôlé qu'un choix confiant déterministe. Pour le dire plus simplement, si nous n'avions jamais de choix difficile, nous pourrions être libres dans un monde totalement déterministe, et nous ne « perdriions » aucune forme de liberté ou de contrôle. On peut donc imaginer un monde compatibiliste dans lequel nous sommes autant (et peut-être même plus) libres que dans le monde libertarien imaginé par Kane et Balaguer : il s'agirait simplement d'un monde sans choix difficile. En ce sens, on peut dire que cette position est moins radicalement « libertarienne » que celle de la causalité agentive. Dans cette autre théorie, il existe une propriété de contrôle ou de liberté que nous ne pouvons avoir *que* dans un monde indéterministe. Cependant, si l'on utilise les définitions proposées dans cette étude, la position de Kane et Balaguer reste clairement libertarienne, puisqu'elle revient à dire que *notre phénoménologie* du choix nous oblige à postuler de l'indéterminisme dans la nature si l'on veut posséder le type de contrôle agentif que nous croyons avoir. Certains auteurs considèrent que cette différence entre les deux théories donne un avantage à la causalité agentive pour répondre à l'objection du contrôle diminué ; comme si la causalité agentive avait les moyens de rendre compte que les choix indéterministes sont contrôlés, et que la causalité événementielle n'en avait pas les moyens. Mais comme on vient de le voir, ce n'est pas le cas : la causalité événementielle a tout à fait les moyens de considérer nos choix indéterministes comme contrôlés, même si c'est avec un sens de contrôle qui, dans d'autres mondes possibles, est compatible avec le déterminisme.

Pour conclure, les deux théories peuvent répondre à l'objection du contrôle diminué. Dans la causalité agentive, l'idée est que l'insertion d'un indéterminisme n'augmente pas le contrôle *en elle-même*, mais parce qu'elle est une condition nécessaire à un type de contrôle métaphysiquement irréductible, le pouvoir agentif. Dans les théories événementielles, l'insertion d'un indéterminisme augmente le contrôle *dans le cas des choix difficiles* si cet indéterminisme résulte d'une concordance entre le poids causal et le poids rationnel des motivations de l'action, laquelle concordance est essentielle pour qu'il y ait un véritable « contrôle par les raisons », ou une véritable « causalité mentale » au sens défini au chapitre précédent. De ces deux solutions, on peut dire que la première est plus satisfaisante parce qu'elle fait apparaître un type de contrôle qui est fortement incompatibiliste (un type de contrôle qui ne peut être satisfait dans *aucun* monde déterministe), alors que la solution événementialiste fait seulement apparaître un type de contrôle fondamentalement *compatibiliste* qui « devient » incompatibiliste pour les choix phénoménologiquement difficiles. Mais cet avantage ne me semble pas être un avantage *nécessaire* à une théorie libertarienne : il reste vrai dans la théorie de Kane ou Balaguer que notre phénoménologie du choix (du choix difficile) est une phénoménologie qui oblige à postuler de l'indéterminisme dans la nature pour satisfaire une condition de contrôle ou de liberté.

1.1.e. L'argument du rembobinage

Il y a une dernière version de l'argument de la chance qui mérite d'être considérée, une version qui est censée poser un problème de hasard *même* lorsqu'on suppose une forme métaphysiquement irréductible de contrôle des indéterminismes. Cet argument, imaginé par Peter van Inwagen (qui n'en accepte pourtant pas la conclusion) repose sur l'idée de *répéter* ou de *rejouer* plusieurs fois la même situation de choix indéterministe, en *rembobinant* chaque fois l'histoire pour replacer l'agent dans la situation initiale. On l'appelle donc l'argument du *replay* (rejouer) ou du *rollback* (rembobiner).

Le scénario dont part van Inwagen est le suivant :

Supposons qu'il y ait des actes libres indéterminés. Supposons, par exemple, que dans une situation difficile Alice a été confrontée au choix entre mentir et dire la vérité et qu'elle a librement choisi de dire la vérité [...] Supposons qu'immédiatement après qu'Alice a dit la vérité, Dieu a ramené l'univers à l'état qui était le sien précisément une minute avant qu'Alice ne dise la vérité (appelons le premier moment où l'univers était dans cet état « t1 » et le second moment où l'univers était dans cet état « t2 »), et laissons les choses se poursuivre à nouveau. Que serait-il survenu la deuxième fois ? [...] Aurait-elle menti ou aurait-elle dit la vérité ? Puisque la décision « originale » d'Alice était indéterminée, sa « seconde » décision aurait aussi été indéterminée. [...] Imaginons maintenant que Dieu ait ramené *un millier de fois* l'univers à l'état qui était le sien à t1. [...] Que devrions-nous nous attendre à observer ?
(Van Inwagen 2000, 14, traduction de Michon 2011, 107–108)

Tel est le scénario. La réponse à la question posée constitue la première prémisse de l'argument :

Nous ne pouvons pas dire ce qui serait arrivé, mais nous pouvons dire ce qui serait *probablement* arrivé : parfois Alice aurait menti, et parfois elle aurait dit la vérité. Au fur et à mesure que le nombre des répétitions [*replays*] augmente, nous, les observateurs, allons, presque certainement, observer que la proposition de sorties « vérité » et la

proportion de sorties « mensonge » se stabilise et converge vers une certaine valeur. (*ibid.*)

La première prémisse est donc la suivante :

(1) Si Dieu replaçait Alice dans la situation de choix indéterminé un très grand nombre de fois, le ratio de « vérité » et de « mensonge » convergerait vers une valeur définie.

On peut envisager ici différents ratios, 30% de mensonges, 70% de mensonges, ou plus simplement 50% de mensonges. L'important, c'est qu'à l'intérieur de cette série, chaque acte particulier commence à sembler totalement hasardeux. Plus précisément, van Inwagen raisonne ainsi :

A mesure que nous voyons le nombre de répétitions [*replays*] augmenter, nous sommes de plus en plus convaincus que ce qui arrivera dans la prochaine répétition [par exemple la sept cent vingt-septième] est une pure question de hasard. [...]

Evidemment, ce qui vaut de la sept cent vingt-septième répétition vaut pour chacune, y compris celle qui n'était pas au sens strict une répétition, la première séquence d'événements. Mais ce résultat concernant la « répétition initiale », devrait valoir que Dieu se soucie ou non de produire des répétitions. [...] Par conséquent, une action indéterminée est simplement une pure question de hasard. [...] Mais alors comment pouvons-nous dire que c'était un acte libre d'Alice ? (*ibid.*)

Autrement dit, les prémisses supplémentaires de l'argument sont les suivantes :

(2) Si le ratio vérité/mensonge converge vers une valeur définie, alors l'issue de la 1000^{ème} répétition (par exemple) est une pure question de hasard

(3) Si l'issue de la 1000^{ème} répétition est une pure question de hasard, l'issue du premier choix de la série est aussi une pure question de hasard

(4) Que le premier choix soit ou non suivi d'une série de répétitions ne change rien à son caractère hasardeux ou non.

(C1) Donc le choix d'Alice, même sans les répétitions divines, est une pure question de hasard.

(5) Si l'issue d'un choix est une pure question de hasard, ce n'est pas un choix libre.

(C2) Donc le choix d'Alice (ou tout autre choix indéterministe) n'est pas un choix libre.

Il y a, à mon sens, deux manières de lire l'argument de van Inwagen, et corrélativement deux manières d'y répondre.

La réponse la plus courante consiste à accorder la conclusion intermédiaire (C1), à savoir que le choix d'Alice ou tout choix libre est une pure question de hasard, mais à se concentrer sur la notion de hasard qui est ainsi établie pour montrer que ce type de hasard n'est pas incompatible avec la liberté ou le contrôle de l'agent. Quel sens van Inwagen donne-t-il au concept de hasard qu'il emploie ? Notre auteur a le mérite de donner une définition explicite : un événement est hasardeux s'il a une *probabilité objective inférieure à 1*. Ainsi, ce que prouverait la première partie de l'argument (avec le scénario imaginaire des rembobinages), c'est que les actions libres indéterministes ont certainement une probabilité objective. D'après (Buchak 2013), on peut donc résumer les étapes (2), (3) et (4) par la prémisse suivante :

(P2) Si le ratio de mensonges et de vérités converge vers une valeur définie dans le scénario de répétitions, alors le mensonge d'Alice (ou le fait qu'elle dise la vérité) dans le scénario normal ont tous les deux une probabilité objective à t_1 , à savoir le ratio de mensonges (ou de vérités) par rapport au nombre total de cas.

Et Buchak souligne que cette prémisse est une inférence à la meilleure explication bien difficile à rejeter : pourquoi la proportion de mensonges (ou de vérités) convergerait-elle vers une valeur définie, si ce n'est parce que cette valeur définie représente la *probabilité objective* de cet événement ? De fait, il semble bien qu'aucun auteur n'ait cherché à contester cette prémisse. Dans cette lecture de l'argument, le seul rôle du scénario des répétitions est donc de prouver qu'il existe (même dans le cas sans répétition) une probabilité objective. La stratégie majoritaire pour répondre à cette lecture consiste à accepter l'existence d'une probabilité objective et à contester la prémisse (5) : « van Inwagen a-t-il raison de dire que si il y a une probabilité objective inférieure à 1 qu'Alice choisisse de dire la vérité, alors elle n'a pas la capacité de choisir de dire la vérité ou de faire autre chose, comme de choisir de mentir ? » (Franklin 2011, 216). Autrement dit, une action ne pourrait-elle pas à la fois être « hasardeuse » dans le sens (faible) où elle a une probabilité objective et pourtant être contrôlée et libre ?³⁷⁶

Pour répondre à cette question, il faut tout d'abord éliminer un mauvais argument avancé par van Inwagen pour nous convaincre de l'absence de contrôle pour les actions indéterminées, à savoir l'argument de la promesse. Van Inwagen (2000, 17) imagine l'ami d'un candidat à des élections, qui connaît des informations compromettantes sur ce candidat, et qui se trouvera dans une situation où il pourra choisir librement de les révéler ou de les taire. Si le candidat est conscient de l'indéterminisme futur de son choix, d'après van Inwagen il n'est pas en position de promettre aujourd'hui à son ami qu'il ne révélera pas le secret. Et comme il n'est pas en position de promettre tel ou tel choix, cela montre qu'il n'est pas au contrôle de ce choix. Compte tenu des discussions qui précèdent, la nature de l'erreur devrait ici sauter aux yeux : l'erreur concerne le *moment* du choix et du contrôle. A quel moment l'ami n'est-il pas au contrôle de son choix d'après cet argument ? Non pas au moment du choix lui-même, mais *avant* ce choix. Autrement dit, tout ce que montre l'argument, c'est que l'agent n'a pas de « contrôle antécédent » sur son choix futur. Mais comme on l'a vu un peu plus haut, l'absence de contrôle antécédent n'exclut pas qu'il y ait un contrôle simultané (un contrôle *au moment du choix*) et loin de rendre impossible la notion de choix libre, c'est au contraire une condition indispensable pour que le choix, au moment où il est fait, soit vraiment libre. On peut donc accepter la conclusion de l'argument de la promesse (l'absence de contrôle antécédent) sans que cela pose le moindre problème pour la thèse libertarienne.³⁷⁷

³⁷⁶ Buchak elle-même tente une autre stratégie : elle considère qu'on peut rejeter la prémisse (1). Il n'est pas évident du tout que la proportion de mensonges et de vérités *doive* converger au bout d'un certain nombre de cas. Pour affirmer cela, il faudrait un argument. Or, d'après Buchak, le seul argument avancé par van Inwagen est le principe de la loi des grands nombres... un principe qui *présuppose* l'existence d'une probabilité objective. Cette loi dit en effet que *lorsqu'un événement a une probabilité objective de p* , alors au bout d'un grand nombre de cas, sa répartition convergera vers p . Donc van Inwagen raisonnerait ici de manière implicitement circulaire. Cette stratégie ne me semble pas convaincante ; mon problème principal avec cette stratégie est qu'une suite peut effectivement ne *pas* converger, mais qu'il faut pour cela lui donner une « allure » *très spécifique* (par exemple, 1 vérité, 2 mensonges, 4 vérités, 8 mensonges, ... $2n$ vérités, $2n+1$ mensonges ; cette suite divergera constamment), et cette allure très spécifique ne me semble pas être du tout ce qu'on « devrait s'attendre » à trouver dans le cas des actions libres répétées.

³⁷⁷ Pour cette réponse, voir notamment (Franklin 2011, 217). Si l'on accepte la conclusion de cet argument, on pourra peut-être en revanche douter de sa prémisse : il n'est pas absolument évident que je ne puisse pas promettre de faire quelque chose même lorsque j'ai conscience que mon choix futur sera libre et que son issue est donc indéterminée au moment de la promesse. Une raison que je pourrais avoir de promettre lorsque mon

La vraie question est donc de savoir si le fait qu'un acte ou un choix ait une probabilité objective inférieure à 1 supprime le contrôle *au moment* du choix. Puisque l'argument de la promesse ne prouve pas ce point, on peut dire tout d'abord que van Inwagen n'offre aucun argument pour accepter la prémisse (5). Mais par ailleurs, différents auteurs, aussi bien dans la tradition de la causalité agentive que dans celle de la causalité événementielle, se sont appliqués à montrer positivement que cette prémisse était probablement fausse.

La réponse des défenseurs de la causalité agentive est très simple : elle consiste à insister sur le fait que la notion de causalité agentive ou de pouvoir agentif est *en elle-même* une notion de contrôle, métaphysiquement irréductible :

Le fait qu'Alice cause son intention de dire la vérité est en lui-même un *exercice* de contrôle. Et puisque, par hypothèse, c'est littéralement l'agent elle-même qui produit l'effet, il est difficile de voir comment cette forme de contrôle pourrait possiblement être améliorée. Et donc où se trouve le hasard ? (O'Connor 2011, sec. IV)

La réponse par la causalité agentive consiste donc à dire que l'objecteur a simplement ignoré ou mal compris le pouvoir explicatif de la notion de causalité agentive.

Les défenseurs de la causalité événementielle, comme (Franklin 2011), sont parfois tombés d'accord avec van Inwagen pour dire que la notion de causalité agentive ne pouvait pas aider contre l'argument du rembobinage. Je ne les suivrai pas dans cette direction car il me semble que la réponse traditionnelle que je viens d'évoquer est dialectiquement légitime. En revanche, je les suivrai lorsqu'ils s'attachent à montrer qu'un traitement par la causalité événementielle peut *aussi* montrer qu'un choix qui a une probabilité objective inférieure à 1 peut néanmoins être contrôlé. La réponse de Franklin est essentiellement la suivante : avoir le contrôle d'un choix, ou encore être *capable* de faire un choix, ne suppose pas qu'on ait une capacité *infaillible* à le faire. Par exemple, on peut dire que Tony Parker est capable de réussir un lancer-franc, même s'il y a des situations où il rate son lancer-franc. Dans les situations où il le réussit, il serait absurde de dire qu'il l'a réussi « par pur hasard » simplement parce qu'il existait une probabilité objective qu'il le rate. Cette probabilité objective n'enlève pas son contrôle. Autrement dit, pour qu'un agent S ait le contrôle d'un choix ou d'une action A, pour qu'il soit « capable de faire A », il n'est pas nécessaire qu'il réussisse sa « tentative » de faire cette action dans tous les mondes possibles où il essaie, il suffit qu'il y ait un grand nombre de mondes possibles dans lesquels (i) il réussit à faire A et (ii) son effort joue un rôle explicatif essentiel pour rendre compte du fait qu'il a fait A.³⁷⁸ Or il est tout à fait possible que ces conditions soient remplies dans le cas d'un choix difficile et indéterministe, tel que le choix d'Alice de mentir ou de dire la vérité. Dans la théorie défendue par (Kane 1996), Alice fait deux efforts à la fois, un effort pour dire la vérité et un effort pour mentir, et c'est la concurrence de ces deux efforts qui produit une situation indéterministe. Si Alice choisit finalement de dire la vérité, ce sera un effet de l'effort qu'elle faisait pour dire la vérité, et il y a un grand nombre de mondes possibles où cela se passe (disons environ 50%) ; de même, si Alice choisit finalement de mentir, ce sera un effet de l'effort qu'elle faisait pour mentir, et il y a un grand nombre de mondes possibles où cela se passe (environ 50%). Dans les deux éventualités, on pourra dire que Alice a réussi (avec compétence) à mener à bien un effort qu'elle faisait pour faire X, et donc qu'elle avait la capacité ou la compétence de faire X. La réponse des défenseurs de la causalité événementielle est donc la suivante : pour dire que les

choix reste indéterminé pourrait être la suivante : il se pourrait que le *fait* de promettre de dire la vérité, par exemple, augmente ma probabilité objective de dire la vérité, sans annuler totalement la probabilité objective de mentir. Dans ce cas, il n'est pas évident qu'une promesse serait irrationnelle ni insincère.

³⁷⁸ Les conditions données par (Franklin 2011, 218) sont encore plus souples que celles données ici. La réponse que j'apporte est en fait inspirée à la fois de (Franklin 2011) et de (Kane 1996).

actions ne sont libres et contrôlées *que* lorsque leur probabilité objective est de 1, il faudrait défendre une conception de la capacité ou de la compétence qui est incroyablement exigeante et qu'on a toutes les raisons de rejeter, indépendamment du problème du libre arbitre.³⁷⁹

Ainsi, il me semble que la théorie agentive et la théorie événementielle ont toutes deux les ressources pour répondre à l'argument de manière satisfaisante : si l'on considère l'acte indéterminé d'Alice et qu'on remarque le fait qu'il possède une probabilité objective inférieure à 1, cela ne devrait pas suffire à nous convaincre que cet acte est non contrôlé et non libre, puisqu'il a par ailleurs « tout ce qu'il faut » pour générer au contraire l'intuition qu'il est contrôlé et libre, soit un pouvoir agentif métaphysiquement irréductible, soit une compétence à faire réussir un effort de manière suffisamment fréquente.

Reste que cette première lecture de l'argument du rembobinage n'est pas la seule. On a vu que, selon cette lecture, le *seul* intérêt de l'hypothèse des rembobinages était de prouver l'existence d'une probabilité objective même dans le cas d'un choix d'Alice isolé (ou sans rembobinage). Et l'argument consistait alors à essayer de produire l'intuition que, pour ce choix isolé avec probabilité objective, Alice n'avait donc pas de contrôle et de liberté. Mais cette intuition est plus que difficile à produire : si l'on s'en tient au cas d'un choix isolé, le simple fait qu'il y ait probabilité objective ne suffit pas à renverser notre intuition de contrôle. Mais il n'en demeure pas moins que *lorsqu'on situe ce choix dans la série des rembobinages* une intuition inquiétante persiste, une intuition de « hasard » problématique qui va au-delà de la simple notion de probabilité objective. Et contre cette intuition problématique, il ne suffit pas de rappeler que, pour le cas isolé, nous avons une intuition de contrôle, car l'argument consiste justement à mettre cette intuition en *concurrence* avec d'autres intuitions, peut-être plus fortes, que nous avons lorsque le choix d'Alice est situé dans une série de répétitions. On pourrait alors reconstruire l'argument de van Inwagen comme un argument « par cas » (formellement comparable à l'argument des quatre cas de Pereboom).

Dans le cas 1, le choix d'Alice de dire la vérité est isolé, et nous avons pour ce cas l'intuition normale qu'il s'agit d'un choix contrôlé (quel que soit le modèle de contrôle que l'on défende, par causalité agentive ou événementielle).

Dans le cas 2, le choix d'Alice est considéré au début d'une série de répétitions où Alice tantôt dit la vérité, et tantôt ment.

Dans le cas 3 enfin, le choix d'Alice de dire la vérité est considéré comme le 1000^{ème} choix (par exemple), c'est-à-dire qu'il est considéré comme étant précédé par (voire comme étant « au milieu de ») une série de répétitions.

Pour le cas 3, il me semble difficile de nier que nous avons l'impression problématique d'un « hasard » assez radical, d'un hasard qui pourrait fort bien menacer la notion de contrôle et de liberté. Et la force de l'argument tient au fait qu'il ne semble pas y avoir de différence *pertinente* entre les trois cas. C'est le sens même des prémisses (3) et (4) de l'argument d'origine – prémisses que la reformulation traditionnelle revient à ignorer :

(3) Si l'issue de la 1000^{ème} répétition est une pure question de hasard, l'issue du premier choix de la série est aussi une pure question de hasard

³⁷⁹ On pourrait être tenté de dire la chose suivante : pour avoir la compétence de faire X, il ne faut peut-être pas réussir « à tout coup », mais au moins réussir plus de la moitié des fois. Or cela n'est pas le cas d'Alice ; elle ne peut pas réussir « à la fois » à mentir et à dire la vérité plus de 50% des fois. Donc elle ne peut pas avoir la compétence « duale ». Pour se convaincre que la proportion de succès n'a pas besoin d'être supérieure à 50% mais seulement « suffisamment élevée », on peut penser au cas d'un archer extrêmement fort qui réussit le mille à 150 mètres 40% des fois. *Lorsqu'il le réussit*, on ne sera normalement pas tenté de dire qu'il l'a réussi « sans compétence » et par pur hasard. Bien sûr, on pourra dire qu'il y a eu une *part* de chance dans son succès, mais une part de chance qui ne lui enlève pas le mérite de la compétence.

(4) Que le premier choix soit ou non suivi d'une série de répétitions ne change rien à son caractère hasardeux ou non.

Ces prémisses reviennent en effet à dire ceci :

(3*) Si le cas 3 est un cas de hasard problématique pour la liberté, alors le cas 2 aussi.

(4*) Si le cas 2 est un cas de hasard problématique pour la liberté, alors le cas 1 aussi

Face à cet argument, il ne sert à rien de répéter que nous avons une forte intuition de contrôle *dans le cas 1* (grâce à la causalité agentive ou grâce aux efforts). L'argument accepte tout à fait l'existence de cette intuition initiale, mais il cherche à la renverser par d'autres intuitions plus fortes. Comment répondre alors à l'argument sous cette forme ? Le choix d'Alice est-il contrôlé dans le cas 3 où il se trouve au milieu d'une série de répétitions alternativement mensongères et véridiques ? Ou bien ce cas de répétitions est-il significativement différent de la situation actuelle où Alice fait un seul choix sans répétition ? Mais alors quelle est cette différence ?

Je n'apporterai pas à cet argument une réponse parfaitement explicite, parce que je ne sais pas, personnellement, laquelle de ses prémisses est fausse. C'est-à-dire que je ne saurais pas dire si le choix d'Alice pris au milieu d'une série de répétitions est contrôlé, ou bien s'il est purement hasardeux et non libre mais que cette absence de liberté ne se propage pas au cas normal d'un choix isolé. En revanche, je crois pouvoir montrer que *l'une ou l'autre* des prémisses doit être fausse, et ce par un argument d'analogie avec la notion de connaissance.

Comme on l'a vu dès l'introduction, il y a une incompatibilité entre connaissance et « certaines formes » de hasard. Et différents scénarios de croyances hasardeuses ont été proposés pour déterminer exactement *quelles* formes de hasard sont incompatibles avec la connaissance. L'un des scénarios les plus célèbres est celui des fausses granges imaginé initialement par (Goldman 1976). Imaginons qu'un homme conduise dans la campagne et aperçoit une série de granges, ou en tout cas de façades qui semblent être des granges. En réalité, bien qu'il ne le sache pas, il se trouve au milieu d'un gigantesque studio de cinéma en plein air où se trouve toute une rangée de fausses façades de granges au milieu de vraies granges, et la millième façade se trouve être la façade d'une vraie grange. L'homme conduit au milieu de ces façades, vraies et fausses, pendant quelques minutes puis arrive à la 1000^{ème} (la vraie), et *devant celle-ci*, se met à penser : « je pourrais m'arrêter quelque part ; tiens, voici justement une grange ». Autrement dit, il forme à propos de la 1000^{ème} façade la croyance perceptive, vraie et justifiée, qu'il s'agit d'une grange. Mais évidemment, il aurait pu s'arrêter quelques mètres plus tôt et former la même croyance, de la même façon, à propos d'une fausse façade. En un sens, c'est donc par pure chance que sa croyance se trouve être vraie. Cette croyance est-elle néanmoins une connaissance, du simple fait qu'il s'agit d'une croyance perceptive formée d'une manière normale par quelqu'un qui n'a pas de problèmes de vue ? On peut tout à fait en douter. Il y a même une certaine intuition que le hasard en question empêche la croyance du conducteur d'être une connaissance dans ce cas. C'est ce que j'appellerai le cas 3.

Mais imaginons maintenant le cas 2, identique au cas 3 sauf que le conducteur s'arrête à la toute première façade, qui se trouve être également la façade d'une vraie grange. Il n'a même pas le temps de voir les autres façades parce qu'il ne poursuit pas sa route. Il s'arrête simplement devant ce qui lui semble, à juste titre, être une grange, puis repart dans l'autre sens après quelques minutes.

Enfin imaginons le cas 1, dans lequel la même grange que celle du cas 2 ne se trouve *pas* suivie d'une rangée de vraies et fausses façades mais se trouve néanmoins exactement au

même endroit et notre conducteur s'y arrête après avoir formé exactement de la même façon la croyance qu'il s'agit d'une grange, avant de repartir dans l'autre sens.

Dans le cas 1, tout est « normal » : un homme voit une grange et forme la croyance perceptive qu'il s'agit d'une grange de la même manière que nous formons n'importe quelle croyance perceptive. Certes, cette croyance est *faillible* dans le sens où il y a des mondes possibles où elle est fausse (et où il s'agit d'une fausse façade). Mais si cette faillibilité suffisait à lui refuser le statut de connaissance, alors nous n'aurions aucune connaissance perceptive du tout. Autrement dit, si nous avons des connaissances perceptives, il faut accepter que le cas 1 est un cas de connaissance.

Et pourtant, il demeure l'intuition que le cas 1 ne devrait pas différer du cas 2 de manière pertinente. C'est-à-dire qu'on peut avoir l'intuition suivante :

(4G) Que la première croyance perceptive soit ou non suivie d'une série de façades vraies ou fausses ne change rien à son caractère hasardeux ou non.

Et de la même façon, on peut avoir l'intuition qu'il n'y a pas de différence pertinente entre les cas 2 et 3, c'est-à-dire :

(3G) Si la croyance produite devant la 1000^{ème} façade est une pure question de hasard, alors la croyance produite devant la 1^{ère} façade est aussi une pure question de hasard

Si on accepte que les cas 1, 2 et 3 ne sont pas pertinemment différents, alors on est obligé de renoncer à l'une ou l'autre de nos deux intuitions de départ : soit nous n'avons aucune connaissance perceptive, même dans les cas normaux comme le cas 1, soit la croyance du conducteur à la 1000^{ème} façade est une connaissance (c'est-à-dire qu'elle souffre d'un type de hasard qui n'empêche pas la connaissance).

L'analogie entre cet argument et l'argument du rembobinage dans sa deuxième formulation devrait être claire. Pour répondre au second, on peut donc s'appuyer sur la réponse au premier. Mais comment répondre au premier ?

Dans le cas de la connaissance, faut-il devenir sceptique en abandonnant notre intuition sur le cas 1 ? Ou bien faut-il abandonner l'intuition d'un problème pour le cas 3 ? Ou bien faut-il penser qu'il y a une différence significative soit entre 1 et 2, soit entre 2 et 3 ? Je ne suis pas sûr de savoir quelle est ici la meilleure réponse, mais je crois savoir quelle est la pire. La pire réponse consisterait à embrasser la conclusion sceptique. Notre intuition la plus forte est certainement que la connaissance perceptive, dans des conditions normales au moins, est possible (bien qu'elle ne soit pas infaillible). Restent donc les deux autres options. La première option serait de dire que dans le cas 3, le conducteur *sait* en fait qu'il se trouve devant une grange. Bien sûr, on ne peut pas nier qu'il y ait une certaine forme de hasard dans la manière dont il a acquis cette connaissance, mais la réponse consisterait à dire que *ce* type de hasard est de ceux qui n'empêchent pas la connaissance, et pour nous en convaincre, le défenseur de cette solution insisterait sur tous les mérites « intrinsèques » de cette croyance : il s'agit d'une croyance formée par un système visuel qui fonctionne bien, qui a reçu les informations pertinentes pour former cette croyance ; par ailleurs, le hasard en question ne concerne que des éléments *extrinsèques* au processus de formation de croyance, comme on peut le voir si on observe les différences entre les cas 1, 2 et 3. Les seules différences entre ces cas ne concernent pas le processus de formation de croyance lui-même, ni l'objet perçu, mais ce qui se trouve « autour », ou le contexte. Par conséquent, les mérites intrinsèques de cette croyance permettent sans doute de la considérer comme une croyance formée par un bon processus et comme une connaissance. La deuxième réponse consisterait à dire que le hasard épistémique du cas 3 est trop « grave » pour permettre la connaissance, et à identifier une

différence pertinente entre les cas 1 et 2 ou entre les cas 2 et 3. Il me semble que la différence la plus plausible serait entre les cas 1 et 2, c'est-à-dire entre le cas où les possibilités d'erreur sont simplement contrefactuelles (dans d'autres mondes possibles, cette façade est une fausse façade, ou bien elle se trouve au milieu de fausses façades) et les cas (2 et 3) où le contexte actuel rend pour ainsi dire réelles les « possibilités d'erreur », sous la forme de fausses façades actuellement présentes. Une manière de considérer cette différence serait la suivante : lorsque les fausses façades sont actuellement présentes, cela rend « proches » les mondes possibles où le conducteur fait une erreur (il aurait pu *facilement* s'arrêter à la 2^{ème} façade plutôt qu'à la 1^{ère}). Quelle que soit la raison exacte pour laquelle cette différence est pertinente, la deuxième réponse consisterait donc à dire que lorsque la vraie façade est isolée, le conducteur peut avoir une connaissance, mais lorsqu'elle se trouve dans le contexte hostile de la série hasardeuse, le conducteur ne peut plus avoir de connaissance. Entre ces deux solutions, je ne chercherai pas à déterminer quelle est la plus forte pour ce scénario précis des façades de granges.

Ce qui importe ici, c'est que l'on peut appliquer le même type de réponses à l'argument du rembobinage. Il existe théoriquement trois types de réaction à cet argument (dans sa seconde lecture) : soit on considère que le cas 1 (cas d'un choix normal isolé) est en fait aussi hasardeux que le cas 3 et empêche le contrôle et la liberté ; soit on considère que le cas 3 (le 1000^{ème} choix d'Alice) a suffisamment de contrôle pour être libre, malgré l'impression de hasard ; soit on considère qu'il y a une différence pertinente entre 1 et 2 ou entre 2 et 3. Comme dans le cas de la connaissance, je ne suis pas sûr de savoir quelle est la meilleure réaction, mais je crois savoir quelle est la pire. La pire réaction consisterait à devenir sceptique à propos de la liberté, c'est-à-dire à affirmer que le concept de libre arbitre tel que nous en faisons l'expérience (cf. chapitre 4) est insatisfaisable, même dans les cas normaux et favorables tels que le cas 1. Restent les deux autres réactions possibles. La première réaction consisterait à défendre le caractère libre et contrôlé du choix d'Alice dans le cas 3, en insistant sur ses mérites intrinsèques. Ici les mérites intrinsèques divergeront selon qu'on adopte le modèle agentif ou le modèle événementiel : dans un cas, le 1000^{ème} choix d'Alice pourra lui aussi être considéré comme libre et contrôlé *parce qu'il est produit par causalité agentive* et qu'« il est difficile de voir comment cette forme de contrôle pourrait possiblement être améliorée » (O'Connor 2011, sec. IV) ; dans l'autre cas, le 1000^{ème} choix d'Alice sera considéré comme libre et contrôlé *parce qu'il est l'effet compétent d'un effort* pour réaliser ce choix. Mais dans l'un et l'autre cas, on fera observer que la seule différence entre ce 1000^{ème} choix et le choix normal est purement extrinsèque ou contextuelle ; elle tient à d'autres choix qui ont eu lieu ou n'ont pas eu lieu *avant* ou *après* le choix considéré ; elle ne tient pas au choix lui-même et à la manière dont il a été produit. Quant à ses mérites intrinsèques, les seuls pertinents d'après ce raisonnement, le 1000^{ème} choix d'Alice est donc parfaitement contrôlé. La deuxième réponse consisterait à dire que les différences extrinsèques entre le cas 1 et le cas 3 peuvent être pertinentes. En particulier, la différence pertinente se situera probablement entre le cas 1, dans lequel le choix d'Alice est isolé, et les cas 2 et 3 pris ensemble, dans lesquels le choix d'Alice se trouve accompagné d'une série d'autres choix hasardeusement similaires ou différents au choix considéré. On pourrait donc défendre que le hasard est problématique lorsque l'éventualité d'agir autrement est pour ainsi dire actualisée par une série de choix similaires, mais qu'il n'est pas problématique lorsque cette éventualité reste purement contrefactuelle. Et la raison, là encore, pourrait être que la présence de la série actuelle de choix alternatifs rend modalement *très proche* cette éventualité d'agir autrement, tandis que lorsque la possibilité d'agir autrement est simplement contrefactuelle, on a toujours

la possibilité de supposer qu'elle se trouve réalisée dans des mondes possibles *lointains*.³⁸⁰ Ou pour le dire plus simplement, cette réponse consisterait à dire que nous avons raison de ne pas avoir la même intuition pour les choix isolés et pour les choix en série hasardeuse parce que c'est *justement* la série hasardeuse qui fait apparaître le hasard problématique.

Encore une fois, je ne sais pas laquelle de ces deux réponses est la meilleure, mais je suis confiant que l'une ou l'autre doive être la bonne. L'important pour nous est seulement que la première réaction, celle qui revient à réviser notre intuition de contrôle dans le cas de choix isolé, puisse être légitimement rejetée. En effet, tous les choix que nous accomplissons sont, dans le monde actuel, des choix isolés, et donc il nous importe peu de savoir si les choix accompagnés d'une série de répétition sont hasardeux mais dissemblables ou semblables mais également contrôlés et libres. La seule conclusion à retenir, c'est que nos deux modèles du libertarianisme, causalité agentive et événementielle, peuvent tous les deux répondre (de la même façon) à ce problème du rembobinage, dans sa deuxième interprétation aussi bien que dans la première qui mettait l'accent sur les probabilités objectives.

1.1.f. Conclusion

Nous venons donc de voir que les différents arguments soulevés contre l'idée de liberté des alternatives n'étaient pas ultimement convaincants, et qu'on peut leur apporter essentiellement deux grands types de réponse, par la causalité agentive ou par la causalité événementielle.

Si l'agent est la cause agentive de son action, alors cette action n'est ni déterministe ni hasardeuse (contre l'argument du dilemme), et l'indéterminisme joue un rôle pour rendre possible un type *sui generis* de contrôle (contre l'argument du contrôle diminué). Par ailleurs (contre l'argument du rembobinage), l'existence d'une probabilité objective ou la possibilité théorique de rembobinages ne peuvent pas renverser notre intuition pour le cas normal d'un choix isolé, à savoir que ce choix est contrôlé *du fait* qu'il est causé agentivement.

Si d'autre part nos choix sont causés par des efforts compétents contradictoires qui font apparaître un indéterminisme dans l'issue du choix, alors de nouveau, ces choix ne sont ni déterministes ni hasardeux, et l'indéterminisme joue un rôle pour le contrôle car s'ils étaient déterministes, ces choix phénoménologiquement difficiles seraient en réalité causés d'une manière « court-circuitante » et donc pas mentalement contrôlée. Enfin, le fait que ces choix soient produits par un effort compétent à faire tel choix peut suffire à établir solidement l'intuition qu'ils sont contrôlés, cette intuition résistant à l'existence d'une probabilité objective du choix et à la possibilité théorique d'une série de rembobinages.

En conclusion, nous pouvons donc dire que nous avons une expérience phénoménale d'avoir (parfois) des alternatives absolument possibles (cf. chapitre 4), que cette expérience phénoménale nous donne une justification *prima facie* à croire que nous avons (parfois) des alternatives absolument possibles (cf. première partie), et enfin que cette justification *prima facie* n'est pas réfutée de manière *directe* et *a priori*, c'est-à-dire que rien n'empêche logiquement que nous ayons *effectivement* de telles alternatives. Peut-être y a-t-il des réfutations *a posteriori* (II de ce chapitre) ou des réfutations fondationnelles (chapitre 6). Mais pour l'instant, il semble que notre expérience des alternatives soit une bonne route pour établir l'existence d'un libre arbitre qui implique l'indéterminisme causal.

Voyons à présent si l'autre aspect important de notre phénoménologie agentive, l'aspect de la source ultime, constitue une aussi bonne route. Chacune de ces deux routes pourrait être

³⁸⁰ Alice serait donc libre (au sens des alternatives) parce qu'elle pourrait faire autrement (il existe un monde possible dans lequel elle fait autrement) *et parce que* cette éventualité n'est pas modalement trop proche (si elle était trop proche, son choix ne serait plus libre car il serait hasardeux).

suffisante à elle seule, mais je m'efforcerai de montrer que chacune fonctionne, de telle sorte que si un lecteur n'est pas convaincu par mes arguments sur l'une de ces deux routes, il peut toujours me suivre dans mon libertarianisme à cause de la seconde.

I.2. L'argument basique (contre la liberté de source ultime)

L'idée de source ultime, dont j'ai dit (chapitre 4) qu'elle faisait partie du contenu de notre expérience phénoménale, a également fait l'objet de diverses critiques, mais en particulier d'un argument impossibiliste essentiellement développé par Galen Strawson (1986, 291; 1994), bien qu'on puisse en trouver des versions au moins à partir de Leibniz. Il s'agit d'un argument de régression à l'infini dont l'idée fondamentale est que l'idée de source ultime nous engagerait dans une régression infinie de choix que nous ne pouvons pas satisfaire, étant des agents finis et incapables d'être des *causa sui*, ou causes de soi-même.

Strawson lui-même est un auteur incompatibiliste dur ; c'est-à-dire qu'en plus de croire que l'idée de source ultime est insatisfaisable, il croit que cette idée fait partie de notre expérience de nous-mêmes et de nos actions et même qu'elle devrait être satisfaite pour que nous soyons moralement responsables. Par conséquent, il pense que notre phénoménologie de l'action est fondamentalement illusoire et que nous ne pouvons pas être moralement responsables. Mais son argument peut très bien être utilisé (et a été utilisé) sans ces thèses supplémentaires, pour montrer simplement que l'idée *libertarienne* de liberté est insatisfaisable et que nous devons donc penser que notre phénoménologie ne contient pas cet aspect. Quel que soit l'usage qu'on retient, l'argument reste le même.

I.2.a. L'argument

L'argument a été présenté dans différentes versions. La plus claire est sans doute celle proposée par Robert Kane (2005, 71–72), qui se rapproche beaucoup de (Strawson 1986, 23–24), car elle fait apparaître explicitement, et de manière simple, la structure de régression à l'infini. L'argument est donc le suivant :

- (1) Un agent S fait ce qu'il fait à t à cause de ce qu'il est à t (de sa nature ou de son caractère à t)³⁸¹
- (2) Pour être la source ultime de ce qu'il fait à t , S doit être la source ultime de ce qu'il est à t .³⁸²
- (3) Mais pour être la source ultime de ce qu'il est à t , S doit avoir fait quelque chose dans le passé (à $t-1$) dont S était également la source ultime et qui l'a rendu tel qu'il est à t .
- (4) Mais si S était la source ultime de ce qu'il a fait dans le passé, à $t-1$, pour se rendre tel qu'il est (à t), alors S doit aussi avoir été la source ultime de ce qu'il était à $t-1$.

³⁸¹ L'expression anglaise est la suivante : « you do what you do [...] because of the way you are at t » (Strawson 1986, 291).

³⁸² Strawson lui-même ne parle pas de « source ultime » mais de « true responsibility », ou de « ultimately responsible ». Ce qui, dans ces expressions, est essentiel à l'argument, c'est le concept « d'ultime » qui suppose la régression jusqu'à un terme fondamental de la source et/ou de la responsabilité. Le fait que Strawson considère une notion morale (la *responsabilité* morale) est secondaire dans l'argument, qui peut fonctionner aussi bien avec une notion non morale d'attribution de l'action (l'idée de source).

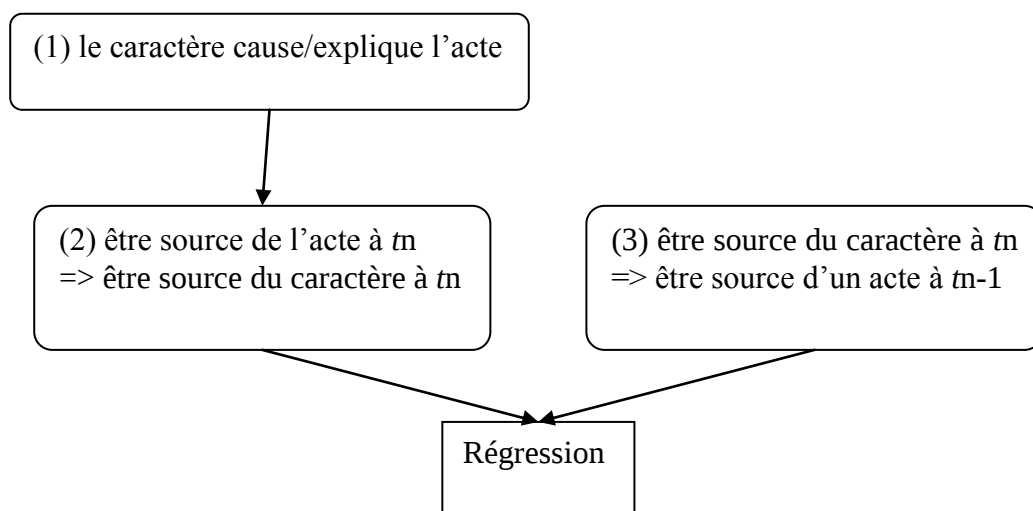
(5) Mais pour être la source ultime de ce qu'il était à $t-1$, S doit avoir fait quelque chose auparavant (à $t-2$) dont il était également la source ultime et qui l'a rendu tel qu'il était à $t-1$.

(...) Et ainsi de suite *ad infinitum*.

Cette série de réquisits fait apparaître la régression à l'infinie. L'argument proprement dit consiste simplement à ajouter la prémisse que cette régression est insatisfaisable pour nous : il est impossible que nous soyons la source d'une série infinie d'actes passés par lesquels nous avons formés nos caractères successifs. Nous sommes en effet des êtres finis, accomplissant un nombre fini d'actes depuis une date passée limitée puisque nous n'avons pas toujours existé. Peut-être qu'un être infini tel quel Dieu pourrait satisfaire une telle régression et devenir ainsi, en un sens absolu, la source ultime de son propre caractère et donc de chacun de ses actes. Il serait alors, selon la terminologie traditionnelle, *causa sui*, cause de soi. Mais pour nous, nous ne pouvons pas être des *causa sui* ; nous ne pouvons donc pas être la source ultime de qui nous sommes et de nos actes.

On peut noter que l'argument basique n'est pas strictement *a priori* ; il utilise une certaine information « contingente » à notre propos, à savoir que nous ne sommes pas des êtres infinis capables comme Dieu d'être *causa sui*. Mais cette « information » est si évidente et si essentielle au genre d'êtres que nous sommes qu'il serait artificiel de la prendre comme une « donnée empirique ».

Pour simplifier la discussion à venir, il faut remarquer que la régression qu'on vient de présenter, même si on la suit jusqu'à l'étape 5, repose en fait entièrement sur deux principes et deux seulement, représentés par les prémisses (2) et (3). En effet, la prémisse (4) n'est qu'une deuxième instance du même principe que (2) – pour être la source d'un acte, il faut être la source de son caractère au moment de l'acte. Et la prémisse (5) n'est qu'une deuxième instance du principe exprimé par (3) – pour être la source de son caractère à un moment, il faut avoir produit ce caractère par un acte antérieur. Ce sont fondamentalement ces deux principes (2) et (3) qui entraînent la régression de l'acte 1, au caractère 1, à l'acte 2, au caractère 2, etc. Quant à la prémisse (1), elle est une *raison* d'adopter le principe (2) : s'il faut être la source de son caractère pour être source d'un acte, c'est parce que le caractère est la cause, ou peut-être plus généralement le facteur explicatif fondamental, de l'acte. On pourrait donc représenter l'argument de régression de la manière suivante :



Telle est la structure de l'argument. Si l'on essaye de suivre la régression de Strawson de manière « réaliste », on remontera de nos actions d'adulte à nos actions d'enfant, puis à nos toutes premières actions du début de l'enfance qui ont été produites avec un caractère probablement entièrement déterminé par notre environnement familial ou social. Autrement dit, on réussira à trouver certaines actions qui sont déterminées par un caractère lui-même déterminé « par nous » au sens où il est déterminé par nos choix antérieurs, et en ce sens, un compatibiliste sera peut-être heureux de souligner que nous pouvons tout de même avoir cette forme de contrôle proximal de nos actions ; mais *ultimement*, ces chaînes remonteront toujours à des actes et à un caractère dont nous ne sommes pas la source. Donc l'idée de source *ultime*, dont nous avons vu qu'elle faisait partie de notre expérience phénoménale, n'est pas possible.

Il est important de noter enfin que cet argument est fait pour fonctionner aussi bien *que le déterminisme soit vrai ou faux*. C'est l'idée même de source ultime qui est censée être insatisfaisable, et l'indéterminisme, où qu'on le place, ne semble pas pouvoir aider puisque l'argument ne présuppose pas le déterminisme.

Nous allons voir à présent comment on peut répondre à cet argument, dans l'un et l'autre modèles libertariens.

1.2.b. Réponse dans le modèle agentif

Ainsi que l'a fait observer Tim O'Connor (2011), le présupposé le plus problématique de l'argument de Strawson est de s'appuyer sur une notion de « responsabilité / liberté pour notre caractère », alors que « la liberté est principalement une caractéristique de nos actions ». C'est cette idée de responsabilité / liberté pour notre caractère qui permet d'initier la régression d'un acte à un caractère, à un autre acte, etc. Mais sommes-nous vraiment à la source de notre caractère ? Et surtout, *devrions-nous* être à la source de notre caractère pour pouvoir être considérés comme la source ultime de nos actions ? C'est-à-dire, doit-on accepter le principe (2) ?

Comme on l'a vu, Strawson apporte une justification en faveur du principe (2) : s'il pense que nous *devons* être la source de notre caractère pour être la source de notre action, c'est parce que le caractère « explique » ou « cause » l'action (prémisse 1). Que répondra le théoricien de la causalité agentive à cette prémisse ? Une première chose qu'il pourra répondre, c'est que nous n'avons pas à considérer notre caractère (nos inclinations, nos préférences, nos jugements, etc.) comme des *causes* de l'action. Tous ces paramètres sont des circonstances dans lesquelles *l'agent* cause l'action. Mais cette réponse n'est pas l'essentiel car le théoricien de la causalité agentive accordera que notre caractère contribue à « expliquer » l'action : notre caractère, en effet, sera responsable de la *probabilité objective* que l'agent aura de prendre telle ou telle décision. Par exemple, si l'agent a un fort désir de lever la main pour prendre la parole, et une forte conviction morale qu'il ne faudrait pas le faire, ces deux éléments de « caractère » pourront contribuer à lui donner, par exemple, une probabilité objective de lever la main de 55% (et 45% de ne pas la lever). Donc la prémisse 1 semble pouvoir être acceptée dans le modèle de la causalité agentive : le caractère explique l'action.

Mais l'important est ici de noter que le caractère n'est *qu'un facteur* qui contribue à expliquer l'action, mais qu'il ne constitue pas une explication *complète* du fait que l'agent a finalement choisi de lever la main. Le caractère est une explication *non suffisante* de l'action. « Le choix est *expliqué par* 'l'état dans lequel j'étais, du point de vue mental', mais il n'est pas complètement le *résultat* de cet état » (O'Connor 2011, sec. IV). En effet, l'*autre* élément explicatif fondamental est la contribution causale apportée par l'agent lui-même sous la forme

de la causalité agentive. A ce titre, il est important de noter que cette contribution causale peut très bien aller à l'encontre de la probabilité la plus élevée : si l'agent choisit de ne pas lever la main, alors qu'il avait une plus faible probabilité objective de faire ce choix, on ne sera pas tenté de considérer l'explication par le caractère comme explication suffisante. Au contraire, l'élément explicatif le plus important est bel et bien la contribution causale apportée par l'agent. Or l'agent est clairement la source ultime de cette contribution causale, il peut donc être la source ultime de l'élément essentiel de l'action.

Cette réponse peut sembler insatisfaisante dans la mesure où elle semble rejeter totalement l'idée de responsabilité pour notre caractère. Or si le caractère est au moins une *contribution* explicative de notre acte, ne faudrait-il pas en être responsable pour être *totale*ment responsable de notre acte ? Si je ne suis responsable que d'un aspect explicatif de mon acte (la causalité agentive), en suis-je vraiment totalement responsable ?

Pour répondre à cette inquiétude, O'Connor fait observer que le théoricien de la causalité agentive peut très bien admettre l'idée que la responsabilité (ou la notion de source de l'acte) peut avoir des *degrés*. Par exemple, si j'ai un désir extrêmement fort, quasi irréprouvable, de voler dans certaines circonstances, si ma cleptomanie me donne une probabilité objective de voler de 95%, et si je vole effectivement, alors on pourra dire que mon vol est *grandement* expliqué par mon caractère, d'une manière telle que cela réduit beaucoup ma responsabilité malgré l'existence de ma contribution causale agentive. Pour que je sois *pleinement* responsable de mon acte, il faudrait que j'aie développé cette cleptomanie par des actes antérieurs pleinement libres et non malades. Mais si l'on suppose que cette maladie m'est survenue de manière totalement involontaire, alors je ne suis pas pleinement responsable de mon vol. Il faut cependant distinguer entre responsabilité complète et responsabilité ultime. La notion de responsabilité complète est une notion de *degré* de responsabilité ; la notion de responsabilité ultime est une notion d'*origine* de la responsabilité. Dans le cas d'un vol par cleptomanie, si j'avais ne serait-ce qu'une faible probabilité objective de ne pas voler, alors la contribution causale agentive était logiquement nécessaire. Et cette contribution causale n'est expliquée et explicable par rien d'autre ; elle est une source ultime de responsabilité (sinon de responsabilité morale, du moins de « responsabilité causale » au sens de l'attribuabilité). Je suis la source *ultime* de *cette part explicative*, bien que je ne sois pas la source *complète* de tous les facteurs explicatifs de la décision que j'ai prise.

Si on accepte cette idée de responsabilité ou de source *par degrés*, alors l'argument de Strawson montre bel et bien quelque chose : il montre qu'il est impossible pour nous d'être la source ultime *complète* de nos actions, au sens où il y aura toujours une *part* de responsabilité causale de nos actions qui nous sera ôtée par les circonstances brutes de nos actions. Mais cela ne veut pas dire que nous n'ayons *aucune* responsabilité ultime, ou que nous ne soyons *en rien* des sources ultimes. Nous sommes sources ultimes d'une contribution fondamentale (qui dans les cas vraiment ouverts en 50%-50% est même *la* contribution explicative décisive).

I.2.c. Réponse dans le modèle événementiel

La réponse à l'argument de Strawson dans le modèle événementiel ne sera pas très différente. La différence principale tient au fait que Robert Kane reste plus proche, dans les détails métaphysiques, de la condition recherchée par Strawson.

En effet, Robert Kane admet que notre caractère (ou nos états doxastiques et préférentiels antérieurs à l'action) soit une *cause* de l'action, et pas simplement une contribution explicative. Bien plus, il accepte que ces éléments causaux soient les *seuls* éléments causaux en jeu puisqu'il n'y a pas de contribution causale agentive. Par ailleurs,

Robert Kane a accordé beaucoup d'importance à l'idée de responsabilité pour notre caractère. Il défend en effet que bon nombre de nos actions, celles que nous accomplissons sans délibération et sans hésitation, du simple effet de notre caractère et des circonstances, sont libres dans un sens *dérivé* du fait que nous pouvons être responsables de notre caractère. Et pour cette raison, Kane s'est intéressé, comme Strawson, à la manière dont notre responsabilité de caractère pouvait être fondée dans des choix antérieurs qui ont formé notre caractère. L'idée d'une régression à des choix antérieurs, Kane l'accepte et la considère positivement. La seule chose qu'il conteste, c'est que cette régression soit nécessairement *infinie*.³⁸³

Pour Kane, la régression qu'initie la responsabilité pour le caractère ne nous renvoie pas à l'infini mais nous renvoie à ce qu'il appelle des « actions formatrices du soi » ou des « volitions formatrices du soi » (*self-forming actions* ou *self-forming willings*, SFAs ou SFWs)³⁸⁴, c'est-à-dire des choix par lesquels nous déterminons et formons, de manière ultime, notre caractère. Une fois que nous avons formé notre caractère de manière ultime par une SFA libre, toutes les actions que nous produirons du fait de ce nouveau trait de caractère seront responsables *de manière dérivée* ou indirecte.

Reste à savoir comment des « actions formatrices du soi » sont possibles, et comment nous pouvons en être la source *ultime*. Ces actions auraient-elles le privilège d'être formées dans une absence totale de « caractère » ? Cela n'est pas possible. Mais ce qui est possible, c'est que notre caractère au moment d'une SFA ne *détermine pas* l'action que nous ferons. Comme nous l'avons déjà vu, dans les situations de choix difficile notre caractère nous place devant un choix non tranché ; notre caractère ne suffit pas à déterminer un choix plutôt qu'un autre. Comme dans le cas de la causalité agentive, on pourra dire que notre caractère *explique* la situation de choix dans laquelle nous nous trouvons, et même en un certain sens il expliquera *le choix* que nous ferons quel que soit ce choix (puisque notre caractère est tel que nous faisons un effort pour faire chacun des deux choix contradictoires). Mais bien qu'il *contribue* à expliquer notre choix, il ne l'explique pas de manière *complète* puisqu'il peut *aussi bien* expliquer une issue que l'autre. Ici, dire que l'explication par le caractère est incomplète ne revient pas à dire qu'il y aurait un *autre* facteur explicatif (comme c'était le cas pour la causalité agentive) ; il n'y en a pas. Le choix n'est tout simplement pas explicable de manière complète par des faits ou par des circonstances du choix ; il s'agit (au moins partiellement) d'un « fait brut ». Un fait brut peut-il nous aider à satisfaire la condition de source ultime ? Oui, parce qu'il ne s'agit pas de n'importe quel fait brut ; le fait en question est le *choix* de l'agent. Autrement dit, lorsqu'on cherche à remonter la chaîne d'explications qui explique tel trait de caractère, nous pouvons typiquement remonter jusqu'à un choix de l'agent comme explication ultime, ultime en l'occurrence *parce qu'il* s'agit d'un fait brut. C'est en ce sens que l'agent (ou ses SFAs) peut être considéré comme source ultime de toutes ses actions.

Le modèle de Kane pour rendre compte de la condition de source ultime revient donc à faire une différence entre deux types d'actions. Pour les choix « faciles » ou les actions « confiantes », l'action repose entièrement sur notre caractère, mais nous sommes ultimement à la source de ce caractère par le biais de certaines actions passées, qui étaient formatrices de notre caractère (les SFAs). Pour les choix difficiles en revanche (les SFAs elles-mêmes), nous n'avons pas besoin d'être à la source du caractère avec lequel ils ont été formés, parce que

³⁸³ Ce sont les thèses T6 à T9 de (Kane 1996, 74–75).

³⁸⁴ Plus précisément, Kane montre d'abord que la régression suppose des *actions* ultimes formatrices du soi, et ils *argumente* ensuite que ces actions doivent être des actions simplement mentales, des choix ou des formations de volitions (c'est la thèse T19, *ibid.*, p.124). Comme nous avons déjà eu l'occasion de voir que les actions libres du modèle événementiel étaient en réalité des choix ou des formations de volitions, je ne reviens pas ici sur cette étape du raisonnement.

notre choix n'est pas déterminé par ce caractère ; nous sommes la source de ces actions de manière immédiate puisqu'il s'agit d'un fait brut qui est *notre* choix. Ce sont ces SFAs que le raisonnement de Strawson n'avait pas prévues. Ces SFAs sont bel et bien causées et expliquées par notre caractère (prémisse 1) mais elles sont causées de manière indéterministe et expliquées de manière incomplète. Par conséquent la prémisse (2) est fausse en ce qui les concerne : on peut être la source d'une SFA sans être la source du caractère qui l'a causée (parce qu'une SFA est notre choix et qu'elle n'est pas déterminée par notre caractère).

Comme pour la causalité agentive, une réponse complète devra admettre que nous ne sommes pas la source *complète* des SFAs et que notre responsabilité à leur égard est susceptible de degrés. Par exemple, supposons que devant un mendiant mon caractère me détermine à hésiter entre passer sans rien donner et donner quelque chose et qu'il s'agisse là d'un choix difficile et d'une action qui formera pour la suite mon caractère, avare ou généreux. Supposons également que mon caractère me détermine à ne *pas* envisager la possibilité de passer sans rien donner *et en lançant un regard de mépris*, ce qui aurait formé pour la suite un caractère méprisant. Supposons enfin que je ne donne rien (mais sans regard de mépris). Dans ce cas, je suis responsable de n'avoir rien donné, mais je ne suis pas responsable de n'avoir pas jeté un regard de mépris puisque mon caractère initial déterminait *cela*. Corrélativement, je serai plus tard responsable (indirectement) de ce que mes actions seront avares plutôt que généreuses, mais je ne serai pas responsable (même indirectement) de ce qu'elles ne seront pas méprisantes. Notre caractère initial – celui avec lequel nous formons nos (premières) SFAs, et que nous n'avons pas contribué à déterminer – peut donc nous ôter certaines responsabilités dans une certaine mesure. Mais la seule chose qui importe pour notre propos, c'est qu'il ne nous enlève pas *toute* responsabilité ; en particulier, il ne nous enlève pas la responsabilité des SFAs elles-mêmes.

1.2.d. Conclusion

L'argument basique de Strawson est, à ma connaissance, la seule grande objection formulée contre le concept de source ultime. C'est une objection importante et qu'on peut retrouver à différents endroits de la tradition, mais il me semble qu'il est possible d'y apporter une réponse satisfaisante, aussi bien dans le modèle de la causalité agentive que dans le modèle de la causalité événementielle.

Pour résumer la discussion, on peut dire que le grand défaut de cet argument est de présupposer implicitement que nous ne pouvons être à la source d'une action que si cette action est *déterminée* par notre caractère. C'est cette présupposition implicite qui amène ensuite le réquisit d'être à la source de ce caractère lui-même. Mais on peut très bien être considéré comme la source ultime d'une action même si cette action n'est pas déterminée par notre caractère.

Dans le modèle de la causalité agentive, nous sommes sources ultimes en tant que causes agentives de nos actions indéterminées par le caractère. Dans le modèle de la causalité événementielle, nous sommes aussi la source ultime en tant qu'auteurs du choix pour nos SFAs qui ne sont pas déterminées non plus par le caractère. Chacun de ces deux types d'actions peut ensuite être considéré comme formateur d'un caractère avec lequel nous accomplirons certaines actions. Mais qu'il y ait ou non cette série d'actions futures, l'action causée agentivement ou la SFA de Kane sont l'une et l'autre des actions dont nous sommes la source ultime.

La condition de source ultime, avec ses conséquences incompatibilistes, n'est donc pas une condition logiquement ou métaphysiquement impossible. Elle est même satisfaisable pour des agents limités dans le temps tels que les êtres humains.

I.3. Conclusion : la concevabilité du libre arbitre

Nous avons donc vu que les grandes objections formulées contre la possibilité ou la concevabilité du libre arbitre n'étaient pas totalement convaincantes. La condition des alternatives possibles, tout comme la condition de la source ultime, peuvent être modélisées d'une manière développée et compréhensible. On pourrait se demander quelle est la plausibilité des modèles que nous avons présentés, mais pour l'instant, ce n'est pas notre propos. Nous considérons ici des objections qui cherchent à montrer l'*impossibilité* ou l'inconcevabilité de ces conditions libertariennes. Et ces objections ont été trouvées non concluantes. Par conséquent, le libre arbitre tel que nous l'expérimentons, le libre arbitre libertarien, n'est pas quelque chose de logiquement impossible et la justification *prima facie* que nous tirons de notre expérience phénoménale n'est pas renversée par une réfutation directe *a priori*.

Il est important de noter que les deux discussions menées dans cette section, celle sur la condition de source ultime et celle sur la condition des alternatives, sont indépendantes et se renforcent mutuellement. En effet, supposons que ma réponse à l'argument basique ne marche pas. Supposons, par conséquent, que la liberté de la source ultime soit métaphysiquement inconcevable. Dans ce cas, ma réaction serait de considérer que *probablement* l'analyse phénoménologique qui identifiait ce critère dans notre expérience agentive était erronée. C'est-à-dire que je réviserais mon jugement phénoménologique pour préserver la véracité de notre expérience agentive : puisque la condition de source ultime est inconcevable, il est probable qu'elle ne fasse pas, finalement, partie de notre expérience agentive. Mais dans cette expérience agentive nous avons identifié un *autre* aspect incompatibiliste, l'aspect des alternatives possibles. Et *cet* aspect n'est pas renversé par l'argument basique, même si cet argument marche. Par conséquent, si l'argument basique marchait, ma réaction serait de dire que notre phénoménologie n'est probablement pas une phénoménologie de la source ultime, mais qu'elle est tout de même une phénoménologie libertarienne à cause de l'aspect des alternatives possibles, et je continuerais de soutenir que cette phénoménologie constitue une *justification* à croire que nous sommes libres de manière indéterministe. Autrement dit, je ne deviendrais compatibiliste (d'ajustement) que si les *deux* arguments impossibilistes marchaient, que si les *deux* voies d'accès à l'incompatibilisme étaient jugées incohérentes et inconcevables. Or si je ne suis pas absolument certain des réponses apportées dans cette section aux deux arguments, je suis assez confiant qu'*au moins un des deux* arguments impossibilistes peut être rejeté. La conclusion de cette section me semble donc tout à fait solide : le libre arbitre tel que nous l'expérimentons est à la fois incompatible avec le déterminisme et concevable, ou métaphysiquement possible.

S'il y a des raisons de rejeter le libertarianisme, ces raisons ne sont donc pas *a priori* et de premier ordre. Restent deux possibilités : une possibilité est que le libertarianisme soit réfuté de manière indirecte, ou par une objection de deuxième ordre. C'est ce que nous étudierons dans le dernier chapitre. Dans ce chapitre, nous nous en tenons aux réfutations directes, aux prétendues preuves que le libertarianisme est positivement *faux* (et pas seulement inconnaissable). S'il n'y a pas de preuve directe *a priori*, la dernière possibilité est qu'il y ait une preuve directe *empirique*, c'est-à-dire une ou des *informations* empiriques sur la nature des êtres humains ou du monde qui rende le libertarianisme faux ou inconcevable. C'est ce type d'informations scientifiques (informations concernant le déterminisme et l'indéterminisme) que nous allons voir à présent. Avant de les envisager, on peut déjà annoncer que les informations scientifiques en question n'étaient certainement pas disponibles

dans l'antiquité, notamment à l'époque du libertarien Epicure. Il semble donc légitime de dire que pour Epicure tout du moins, il n'y avait *aucune* réfutation directe de notre phénoménologie libertarienne, et que par conséquent sa croyance à l'indéterminisme de la liberté était justifiée *ultima facie* – tout du moins au premier ordre. A cette époque « pré-scientifique », notre expérience libertarienne n'était pas réfutée. Cette conclusion intermédiaire n'exclut pas la thèse de certains auteurs, qui pensent que les révélations de la science physique sont venues renverser la justification de nos croyances ancestrales sur nous-mêmes et nos actions. Seulement, comme on va le voir à présent, je ne crois pas que les révélations de la science physique constituent une réfutation plus solide que les arguments *a priori*.

II. *Réfutations empiriques*

Parmi les objections empiriques au libertarianisme, on peut en distinguer deux types. Certaines consistent à dire que le libertarianisme est directement réfuté par la science parce que celle-ci aurait prouvé une forme de déterminisme de nos actions et de nos choix. Selon cette objection (l'objection du déterminisme), le problème n'est pas tant que le libertarianisme serait « inintelligible », c'est qu'il est directement contraire aux *faits* établis par la science. D'autres considèrent que nos actions et/ou nos choix ne sont peut-être pas déterministes, compte tenu de ce que la science nous enseigne, mais que le *type* d'indéterminisme que la science moderne laisse possible ne pourrait en rien constituer une source intelligible de liberté. Ce deuxième type d'objection ressemble aux objections *a priori* parce qu'elle pose fondamentalement un problème d'*intelligibilité*. Je l'appellerai donc l'objection de l'intelligibilité empirique.

II.1. L'objection du déterminisme

Si nous avons des raisons de penser que le monde physique se comporte de manière purement déterministe, alors assurément nous aurions un puissant argument contre l'existence libre arbitre. Bien peu d'auteurs défendent aujourd'hui positivement l'existence du déterminisme causal, mais il est important de prendre le temps d'exposer les raisons de cette prudence et les limites dans lesquelles une forme d'indéterminisme est acceptée ou du moins envisagée. La science a-t-elle donc apporté des preuves empiriques du déterminisme ?

II.1.a. *Le déterminisme ne peut pas être établi a priori*

Avant de considérer les arguments *empiriques* en faveur du déterminisme, il convient de rappeler que le déterminisme a parfois été défendu sur la base de raisons *a priori*. En ce sens, l'objection du déterminisme pourrait en partie entrer dans la section précédente sur les objections *a priori* au libertarianisme. Mais ces arguments sont aujourd'hui suffisamment discrédités pour que le seul éventuel problème du déterminisme soit un problème empirique. Les arguments *a priori* reposaient sur le Principe de Raison Suffisante de Leibniz. Si on prend ce principe dans une version forte, il est incompatible avec l'indéterminisme. En effet, si tout événement contingent doit avoir une raison suffisante de se produire ainsi plutôt qu'autrement, alors cette raison doit bien se trouver quelque part dans ses antécédents causaux, sous la forme d'une cause déterminante. Ou pour le dire autrement, l'indéterminisme causal violerait le principe de raison suffisante : dire qu'un état donné d'un système physique pourrait résulter

indifféremment dans deux événements A ou B, c'est dire que l'événement qui se produira effectivement, par exemple A, se produira sans avoir eu de *raison suffisante* de se produire plutôt que l'autre événement, B.

Pourquoi ce genre d'argument *a priori* est-il abandonné aujourd'hui ? La première raison est l'abandon assez répandu du principe de raison suffisante. La deuxième raison tient aux découvertes du XX^{ème} siècle en physique, en particulier de la physique quantique dont l'interprétation majoritaire implique l'indéterminisme causal. Ainsi, la communauté philosophique s'est persuadée que la question de l'indéterminisme *pouvait* être établie *a posteriori* et que tout argument *a priori* allant à l'encontre de ces données empiriques était donc probablement sans aucun poids, ou en tout cas d'un poids assurément plus faible que ces données empiriques. Sur le premier point, il faut reconnaître qu'il existe encore aujourd'hui des défenseurs du principe de raison suffisante, mais il est important de souligner que ces défenseurs contemporains du principe admettent pleinement l'autorité de la physique quantique pour établir l'indéterminisme, et défendent par conséquent une version du principe plus faible, compatible avec l'indéterminisme (voir Pruss 2006, sec. 8.3). Dans la discussion contemporaine, le principe de raison suffisante est donc soit rejeté soit compatible avec l'indéterminisme ; il ne constitue donc pas une voie possible pour établir *a priori* le déterminisme.³⁸⁵

II.1.b. Preuve empirique du déterminisme universel ?

La vraie menace pour une croyance indéterministe se situe dans les résultats des sciences empiriques. Avons-nous donc une preuve empirique du déterminisme causal ? A certaines époques, notamment au XIX^{ème} siècle, un certain nombre de scientifiques et de philosophes des sciences auraient certainement répondu positivement à cette question. Mais de nouveau, le tournant essentiel a été celui de la physique quantique. Il convient donc de rappeler brièvement les raisons pour lesquelles la mécanique quantique a renversé la croyance au déterminisme.³⁸⁶

La question du déterminisme physique peut être ramenée à la question du déterminisme *des particules élémentaires*. En effet, si les particules élémentaires se comportent de manière déterministe, alors tout ce qui en est composé devra se comporter de manière également déterministe. Et si les particules élémentaires se comportent de manière *indéterministe*, alors quand bien même ces indéterminismes n'apparaîtraient pas au niveau macroscopique, toutefois le déterminisme *universel* serait faux puisque certains événements (microscopiques) seraient indéterministes.³⁸⁷ La mécanique quantique est précisément l'étude de ces particules élémentaires. Or voici ce que nous dit la mécanique quantique à propos du comportement d'une particule élémentaire, un électron, d'après l'expérience de Stern-Gerlach : dans cette expérience, un électron traverse un champ magnétique dans un certain état – un état très particulier dit de « superposition » – et dans ces circonstances, nous savons que l'électron

³⁸⁵ Pour une autre présentation de l'argument *a priori* en faveur du déterminisme, voir (Balaguer 2010, sec. 4.2). Balaguer ne mentionne pas explicitement le principe de raison suffisante et envisage l'argument *a priori* directement comme un argument d'inconcevabilité des événements arrivants sans cause déterminante. Mais comme le rappelle Balaguer, de tels événements sont tout simplement concevables.

³⁸⁶ L'exposé qui suit ne prétend ni à la moindre originalité, ni à une compétence spécialisée en philosophie de la mécanique quantique. Je me contente ici de présenter les résultats pertinents pour la question du libre arbitre, tels qu'ils sont classiquement présentés et admis. Pour une présentation de ces résultats de mécanique quantique appliqués au libre arbitre, voire notamment (Hodgson 2002; Balaguer 2010, sec. 4.4.1).

³⁸⁷ Evidemment, un déterminisme simplement macroscopique suffirait à constituer un problème pour le libre arbitre. Nous aborderons ce point un peu plus loin.

prendra *soit* un chemin vers le haut, *soit* un chemin vers le bas, sans qu'aucune observation antérieure de son état puisse permettre de dire lequel de ces chemins il empruntera.

L'état très particulier de l'électron est relatif à une certaine propriété de l'électron, son « spin », qui peut être *soit up soit down*, et cet état de superposition est très particulier parce qu'il revient à dire que l'électron, avant l'entrée dans le champ magnétique, n'est ni *up* ni *down* mais dans un état de « superposition » des deux. Il est important de noter deux choses à propos de l'état de superposition. Tout d'abord, il ne s'agit pas d'une description liée à notre ignorance. Ce n'est pas que l'électron a un spin *up* ou un spin *down* et que nous ne le savons pas. Il n'a, en lui-même, ni l'un ni l'autre des deux spins, mais pour ainsi dire « hésite » entre les deux.³⁸⁸ Deuxièmement, il faut comprendre que cette superposition est véritablement un état qui rebute notre compréhension commune ; on pourrait l'oublier si l'on parle seulement de cette propriété inconnue du profane qu'est le spin ; mais un électron peut aussi être dans un état de superposition à propos de propriétés plus communes, comme ses propriétés spatiales. Imaginez une boîte constituée exhaustivement des régions A et B, un électron peut être d'une certaine manière dans la boîte, et pourtant n'être *ni* dans A, *ni* dans B, mais dans un état de superposition de *dans A* et *dans B*.

Quoi qu'il en soit de cet état étrange de superposition, l'important est simplement de noter qu'un électron qui se trouve ainsi en état de superposition de son spin est tel que *si on cherche à observer* son spin, le résultat de l'expérimentation sera *soit* une observation *up soit* une observation *down* sans que rien ne nous permette de prédire quelle observation sera faite. C'est précisément ce que réalise l'expérience de Stern-Gerlach. Dans une telle observation, on peut dire que l'électron « se décide » en faveur du spin *up* ou du spin *down*. Plus proprement, c'est ce qu'on appelle un processus de « réduction » (très exactement de réduction du paquet d'ondes, ou réduction de la fonction d'ondes) : l'état de superposition se trouve « réduit » à l'un des deux spins par l'expérience d'observation.

Voici donc les données incontestables de la mécanique quantique : dans certaines dispositifs expérimentaux, les électrons se trouvent dans un état (la superposition des spins *up* et *down*) tel qu'il est *impossible* de prévoir quel comportement ils vont avoir par une observation préalable de leur état (de fait, le processus dont on ne peut prévoir le résultat par une observation préalable est *justement* le processus « d'observation » de leur spin). Tout ce que donne la mécanique quantique à propos de tels processus, ce sont des lois *probabilistes* : il y a une probabilité de 50% que l'électron soit observé en spin *up* et 50% qu'il soit observé en spin *down*.

Cela veut-il dire que la mécanique quantique est indéterministe ? Pas immédiatement. En effet, bien qu'on ne puisse prédire le résultat de l'observation par une observation préalable, il se pourrait que l'électron ait préalablement une certaine propriété inobservable qui le détermine à être observé en spin *up* ou en spin *down*. Certes, on vient de le voir, cette propriété *ne peut pas être* que l'électron aurait « déjà », à notre insu, un spin *up* ou *down* avant l'observation, puisque l'électron est irréductiblement en superposition de ces deux états. Mais il pourrait y avoir une *autre* propriété qui le détermine à *prendre* le spin *up* ou le spin *down* lors de l'expérience d'observation. S'il y a une telle variable cachée, alors le processus de réduction du paquet d'ondes pourrait être déterministe, mais pour l'instant, nous n'avons aucune indication sérieuse nous amenant à postuler l'existence de telles variables cachées. Or s'il n'y a pas de telle variable cachée, alors le processus de réduction du paquet d'ondes sera bel et bien indéterministe, puisque cela reviendrait à dire qu'aucune propriété antérieure du système ne détermine si l'électron prendra le spin *up* ou le spin *down*.³⁸⁹ Pour trancher ce

³⁸⁸ Cet aspect est établi par le très fort argument de Kochen et Specker.

³⁸⁹ Pour être complet, il faudrait mentionner ici une autre interprétation, l'interprétation dite « many worlds » d'Everett, selon laquelle aucune variable cachée ne détermine une observation ou l'autre, mais le processus est déterministe tout de même car l'électron prend d'une certaine manière *les deux valeurs*, spin *up* et spin *down*,

débat, il faudrait donner une théorie de ce qu'il se passe lors du processus de réduction du paquet d'ondes. Et c'est précisément là que les interprétations divergent. L'étude des différentes interprétations nous emmènerait trop loin dans le formalisme de la mécanique quantique ; ce que l'on peut dire sans entrer dans les détails, c'est qu'il existe une interprétation très importante de la mécanique quantique, l'interprétation dite de Copenhague, souvent présentée comme l'interprétation majoritaire ou standard, selon laquelle il n'y pas de variable cachée et le processus est irréductiblement indéterministe. Malgré l'existence d'autres interprétations, l'importance de l'interprétation de Copenhague suffit à établir le point suivant : la mécanique quantique, notre meilleure théorie actuelle des particules élémentaires, a des chances sérieuses d'être une théorie irréductiblement indéterministe.

Par conséquent, il n'y aurait rien de plus faux, aujourd'hui, que de prétendre que les sciences empiriques ont démontré un déterminisme universel. Au contraire, il semble que notre meilleure théorie du monde physique élémentaire tende vers une image radicalement indéterministe.

II.1.c. Preuve empirique du déterminisme macroscopique ?

Au moment où la mécanique quantique a été découverte (et interprétée), ses potentialités indéterministes représentaient une révolution dans un paysage largement dominé par une ferme conviction scientifique déterministe. Auparavant, le déterminisme n'était contesté que par quelques auteurs sur la base justement de la croyance au libre arbitre, mais les sciences expérimentales semblaient donner tort à cette croyance traditionnelle. On aurait donc dû s'attendre à ce que la révolution quantique redonne une force considérable aux thèses libertariennes. En effet, dans une conception Poppérienne de la science, la mécanique quantique apportait une possible réfutation du modèle déterministe et corrélativement une corroboration de la thèse adverse, libertarienne indéterministe. Pourtant, l'effet sur le débat du libre arbitre n'a pas été aussi considérable que cela. Certes, la mécanique quantique a diminué la pression qui pesait sur les défenseurs de l'indéterminisme libertarien, mais on n'a pas observé un effet de revirement massif vers le libertarianisme. Les résultats quantiques ont été largement considérés comme non pertinents pour la théorie du libre arbitre. Pour quelle raison ? La raison principale est qu'on ne voyait pas très bien comment ces indéterminismes pouvaient aider la thèse libertarienne. Et cette raison se subdivise en deux objections. Soit il s'agissait de dire que ces indéterminismes n'apparaissent aucunement comme indéterminisme *des actions* – ces indéterminismes « disparaîtraient » au niveau de l'action – soit il s'agissait de dire que *même* s'ils apparaissent au cœur de l'action, ces indéterminismes ne seraient pas le bon type d'indéterminisme. La deuxième objection est l'objection d'intelligibilité empirique que nous verrons plus loin. Voyons pour le moment l'idée selon laquelle l'indéterminisme « disparaîtrait » au niveau de l'action.

L'objection revient à dire que nous vivrions dans un monde *quasi* déterministe, ou déterministe *for all practical purposes* (FAPP) parce que les indéterminismes quantiques n'apparaîtraient pas au niveau de notre action. Une première version de cette objection consiste à dire que les indéterminismes quantiques disparaissent toujours *au niveau macroscopique* pour des raisons statistiques. C'est-à-dire que le comportement d'un système macroscopique serait entièrement déterministe même si ses composants élémentaires se comportent localement de manière indéterministe. Si l'indéterminisme quantique disparaissait ainsi au niveau macroscopique, alors il est clair que le libre arbitre serait impossible car nos

dans deux mondes possibles différents. Cette interprétation modale revient à dire que les mondes possibles se subdivisent à chaque observation ou à chaque réduction du paquet d'ondes, ces processus fonctionnant comme des « bifurcations modales ».

actions se situent évidemment au niveau macroscopique. Mais le monde macroscopique est-il réellement déterministe ? N'y a-t-il aucunes circonstances dans lesquelles les indéterminismes quantiques pourraient se répercuter au niveau macro en événements macroscopiques indéterministes ?

La réponse est clairement non : il y a des circonstances dans lesquelles l'indéterminisme quantique est « détecté » au niveau macroscopique sous la forme d'un indéterminisme macroscopique. A vrai dire, la description même de l'expérience d'observation quantique (Stern-Gerlach) suffit à établir ce point : ce qui est imprédictible, et indéterministe dans l'interprétation de Copenhague, ce n'est pas seulement le comportement microscopique de l'électron, son spin, c'est bel et bien un événement macroscopique, l'observation par le dispositif expérimental macroscopique du mouvement que l'électron prendra finalement. Tout phénomène de réduction du paquet d'ondes par *observation* expérimentale est un phénomène macroscopique indéterministe si l'on suit l'interprétation de Copenhague. En termes plus précis, ce qui est requis pour qu'un système physique répercute au niveau macro un indéterminisme quantique, c'est une sensibilité très fine aux conditions initiales du système. C'est-à-dire que deux états très peu différents du système (différents seulement pour une parcelle microscopique du système) devront avoir des conséquences très grandes (macroscopiques). Un système qui est ainsi sensible aux très faibles variations initiales est ce qu'on appelle un système *chaotique*. Ainsi, en combinant l'indéterminisme et la « théorie du chaos », on peut obtenir des systèmes dont le comportement macroscopique est indéterministe *du fait* d'une amplification des indéterminismes quantiques.

Il est donc faux de dire que nous vivons dans un monde macroscopique qui est entièrement déterministe parce que les indéterminismes disparaîtraient toujours au niveau macroscopique. Si les indéterminismes quantiques ne se répercutaient jamais au niveau macroscopique, nous n'aurions probablement jamais supposé leur existence car nous n'aurions aucun moyen de les *observer*.

II.1.d. Preuve empirique du déterminisme neuronal ?

S'il est faux que le monde macroscopique soit *entièrement* déterministe, il se pourrait tout de même qu'il soit déterministe FAPP, si les seuls indéterminismes macroscopiques existants sont restreints au cadre très exceptionnel des laboratoires de physique quantique. Il se pourrait que les indéterminismes quantiques ne se répercutent au niveau macroscopique *que* dans les dispositifs expérimentaux du type Stern-Gerlach, et *jamais* dans notre cerveau au moment de nos décisions. Et donc il se pourrait que notre cerveau, pour ne parler que de lui, fonctionne comme un système (macroscopique) entièrement déterministe, où tous les indéterminismes quantiques disparaissent. Si tel était le cas, il est clair que nous n'aurions pas de libre arbitre puisqu'il n'y aurait pas de réelles alternatives possibles *pour nos choix*.

Est-ce le cas ? Avons-nous des raisons de croire à un déterminisme macroscopique *des événements neuronaux* ? La réponse, là encore, est non. Il pourrait y avoir deux manières d'argumenter en faveur du déterminisme neuronal. La première manière consisterait à dire que les neurosciences d'une manière générale emploient des lois strictement déterministes ; ainsi, nous pourrions dire au moins que notre meilleure théorie du domaine neuronal est une théorie strictement déterministe. La deuxième manière consisterait à dire que les indéterminismes quantiques ne *peuvent pas* être répercutés au niveau neuronal, à cause du type de système physique qu'est le cerveau ; ainsi, même si nos lois neuroscientifiques n'étaient pas entièrement déterministes, cela relèverait simplement d'une ignorance de notre part. Aucun de ces deux arguments n'est convaincant comme on va le voir.

L'argument des lois neuroscientifiques déterministes a été soutenu notamment par Ted Honderich qui écrit la chose suivante :

Il ne fait pas de doute que les neurosciences utilisent partout la conception intuitive et standard des circonstances et des effets causaux, telle que les premières nécessitent les seconds. (Honderich 1988, 292)

Mais cette thèse de Honderich est tout simplement fausse, ainsi que l'a souligné (Balaguer 2010, 153), qui est allé se renseigner directement auprès des neuroscientifiques. En particulier, il existe certains événements, les événements de transmission synaptique et d'excitation neuronale, que les neurosciences ne peuvent traiter que de manière probabiliste ou statistique. Balaguer cite par exemple le manuel de Dayan et Abbott :

La libération de transmetteur synaptique est un processus stochastique. La libération du transmetteur au niveau dans un terminal présynaptique ne se produit pas de manière nécessaire à chaque fois qu'un potentiel d'action arrive et, inversement, une libération spontanée peut se produire même en l'absence de la dépolarisation due à un potentiel d'action.

[...]

Etant donné que la séquence de potentiels d'action générée par un stimulus donné varie d'expérience en expérience, les réponses neuronales sont typiquement traitées de manière statistique ou probabiliste. Par exemple, on peut les caractériser par des taux d'excitation, plutôt que par des séquences spécifiques de potentiels d'action.

(Dayan and Abbott 2001, 9)

Ces deux événements correspondent en fait aux deux étapes de la « contribution causale » que peut apporter un neurone dans un processus donné. En effet, le fonctionnement d'un neurone consiste tout d'abord à recevoir un stimulus donné, lequel stimulus sera codé (ou non) sous la forme d'un influx nerveux (potentiel d'action) ; le potentiel d'action se transmet ensuite le long de l'axone du neurone jusqu'aux terminaux synaptiques qui constituent la « porte de sortie » du neurone, le lieu où il délivre (ou non) une information pertinente (des neurotransmetteurs) pour la suite du processus neuronal. Le processus d'acheminement de l'information le long de l'axone est déterministe, mais il est important de noter que chacun des deux processus de transfert d'information (vers le neurone ou hors du neurone) est traité par les neurosciences de manière statistique ou probabiliste et non par des lois déterministes.

Il est donc tout simplement faux d'affirmer que les lois utilisées au niveau neuronal sont des lois déterministes. Bien sûr, cela ne suffit pas pour affirmer que les événements en question sont ultimement indéterministes. Il pourrait se faire qu'il y ait des processus strictement déterministes, à un niveau plus fin que le niveau neuronal, qui déterminent l'excitation des neurones et la transmission synaptique. Mais si ces processus existent, ce ne sont pas les neurosciences qui nous en donnent connaissance,³⁹⁰ et à vrai dire nous n'avons aucune connaissance de ce genre pour l'instant. Nous n'avons donc aucune raison *de niveau neuronal* pour supposer le déterminisme neuronal.

Le seul argument qui pourrait demeurer pour défendre le déterminisme neuronal est un argument d'impossibilité de principe : le problème serait que, même si nous ne connaissons

³⁹⁰ Pour reprendre une déclaration du neuroscientifique Sebastian Seung (MIT) en correspondance avec Mark Balaguer, « The question of whether [synaptic transmission and spike firing] are 'truly random' processes in the brain isn't really a neuroscience question. It's more of a physics question, having to do with statistical mechanics and quantum mechanics. » (Balaguer 2010, 154)

pas les lois déterministes qui régissent les neurones (et même si nous ne pouvons les décrire au niveau neuronal lui-même), le comportement des neurones *devrait* être strictement déterministe parce que les indéterminismes quantiques ne sauraient se répercuter au niveau des neurones. L'argument ici ne saurait être simplement une question de taille ou d'échelle, puisque le niveau des connexions neuronales n'est pas à l'échelle macroscopique proprement dite, mais soit à l'échelle microscopique (quelques dizaines de nanomètres pour les fentes synaptiques), soit au maximum à l'échelle mésoscopique (entre cent et mille nanomètres environ), laquelle échelle mésoscopique est définie précisément par le fait qu'elle contient des phénomènes plus gros que l'échelle microscopique, mais assez petits néanmoins pour que les phénomènes proprement quantiques y aient un effet détectable. Les neurones ne sont donc pas « trop gros » pour être sensibles aux effets de la physique quantique. Mais (Tegmark 2000) a défendu qu'ils présentaient un autre défaut qui pourrait les empêcher de répercuter adéquatement les indéterminismes quantiques : le monde neuronal, le cerveau humain, serait « trop humide et trop chaud » pour permettre aux indéterminismes quantiques d'y avoir une répercussion sensible. Ainsi que l'a montré (Balaguer 2010, sec. 4.4.4.1), cette objection ne pose en fait problème que pour un certain modèle de l'indéterminisme quantique, le modèle proposé par (Hameroff and Penrose 1996). Voyons donc brièvement la réponse de Balaguer en repartant du modèle de Hameroff et Penrose.

Dans le modèle de Hameroff et Penrose, les protéines de tubuline qui se trouvent dans les neurones sont présentées comme capables de se trouver dans un état de superposition quantique (plus précisément, une superposition de leurs deux états classiques *étendu* et *contracté*). Par ailleurs, Hameroff et Penrose suggèrent que ces superpositions sont intrinsèquement instables et sujettes à des réductions de paquet d'ondes, ces réductions de paquet d'ondes étant forcées par un mécanisme de gravité quantique décrit par Penrose. Si tel était le cas, on aurait donc un phénomène d'indétermination quantique directement à l'échelle neuronale, dont le déclenchement serait dû essentiellement à un mécanisme de niveau également neuronal. C'est à cette possibilité en général que Tegmark s'est opposé. Pour Tegmark, le niveau neuronal est trop chaud et trop humide pour que les processus neuronaux puissent affecter des états de superposition neuronale. Plus précisément, le problème est un problème d'échelle de temps : dans les conditions de chaleur et d'humidité du milieu neuronal, les états de superposition ne pourraient statistiquement durer qu'environ 10^{-13} secondes avant réduction du paquet d'ondes, tandis que l'échelle de temps des processus neuronaux est de l'ordre de 10^{-3} secondes. Par conséquent, les éventuelles superpositions neuronales dureraient bien trop peu de temps pour être susceptibles d'être affectées par un processus neuronal.

Il ne sera pas utile d'entrer davantage dans les détails de cette discussion, car ce résumé suffit à voir la portée très étroite de l'argument de Tegmark. Même si l'argument de Tegmark est valable contre le modèle de Hameroff et Penrose, il semble clair qu'il ne permet pas du tout d'établir l'impossibilité d'un indéterminisme neuronal en général. Tout ce qu'il permet d'établir, c'est l'impossibilité d'un indéterminisme neuronal qui (1) reposerait sur un état de superposition quantique situé à l'échelle neuronale elle-même, et (2) supposerait une réduction de paquet d'onde provoquée par un mécanisme neuronal. Mais rien n'empêche d'imaginer un modèle d'indéterminisme neuronal indépendamment de ces deux conditions. Pour qu'il y ait indéterminisme neuronal, il suffirait qu'un état de superposition quantique *microscopique* (infra-neuronal) connaisse une réduction de paquet d'ondes indéterministe et qu'il ait des conséquences au niveau mésoscopique (neuronal). Contre (1), et le modèle de superposition neuronale, on peut donc tout à fait imaginer un modèle de superposition microscopique. Par ailleurs, contre (2), il faut noter que la réduction de paquet d'ondes pourrait fort bien n'être pas déclenchée ou affectée par un mécanisme neuronal ; l'important n'est pas qu'elle *soit affectée* (déclenchée) par un mécanisme neuronal, l'important est qu'elle *affecte* un mécanisme neuronal (que ce mécanisme neuronal dépende dans son

fonctionnement de l'issue de la réduction de paquet d'ondes microscopique). A côté du modèle de la superposition microscopique, on pourrait donc également imaginer un modèle de superposition neuronale avec une réduction de paquet d'ondes « spontanée » et non provoquée du point de vue des mécanismes neuronaux.

L'argument de Tegmark ne permet pas de rejeter la possibilité qu'une réduction de paquet d'ondes indéterministe au niveau microscopique ait un effet discernable au niveau neuronal ; il ne permet donc pas, même s'il marche, de rejeter l'indéterminisme neuronal. Il permet seulement d'en rejeter un modèle particulier, un modèle de superposition neuronale dont il n'y a pas de raison de penser qu'il soit particulièrement privilégié par rapport à d'autres modèles, notamment aux modèles de superposition microscopique à effet neuronal.

II.1.e. Conclusion : pas d'objection déterministe

Pour conclure, il n'y a donc aujourd'hui aucun argument, scientifique ni philosophique, qui permette de réfuter notre expérience du libre arbitre à cause d'une certaine forme de déterminisme. Pour établir une forme de déterminisme, nous avons vu qu'il faudrait apporter un argument empirique car les raisons *a priori* en faveur du déterminisme ne sont pas convaincantes. Et les données empiriques ne confirment absolument pas une thèse déterministe. Elles ne confirment certainement pas une thèse déterministe absolue, car il existe des interprétations indéterministes de la physique fondamentale. Elles ne confirment pas non plus des formes de déterminisme plus restreintes (et qui pourraient être tout aussi problématiques pour le libre arbitre), telles que le déterminisme macroscopique ou le déterminisme neuronal. Dans l'état actuel de la science, la physique fondamentale semble tout à fait ouverte à ce qu'il y ait de réels indéterminismes dans la nature, et à ce que ces indéterminismes au niveau microscopique puissent avoir un effet discernable au niveau neuronal, puis au niveau macroscopique de nos actions.

Il est intéressant de noter que les données de la science empirique n'ont pas toujours été dans ce sens. Au XIX^{ème} siècle, une majorité de scientifiques estimaient que la science confirmait une conception entièrement déterministe de la nature, car les lois de la physique fondamentale d'alors étaient, sans ambiguïté ni marge d'interprétation, des lois déterministes. Dans une telle situation, on peut comprendre que la croyance au libre arbitre ait été conçue comme sérieusement mise en cause par la science. Nous verrons un peu plus loin qu'il existait, même dans ces circonstances, une manière de contester la conclusion strictement déterministe grâce à la notion de causalité agentive. Comme on le verra en effet en II.3., la réponse par la causalité agentive ne fait pas nécessairement appel aux indéterminismes des lois quantiques (ou de quelque loi physique que ce soit). Mais à première vue en tout cas, les lois fondamentales déterministes constituaient certainement une objection considérable contre l'expérience d'être libre de manière indéterministe. C'est donc seulement la révolution quantique qui est responsable de la situation dialectique telle qu'elle est présentée dans cette section. Un auteur qui soutenait l'existence d'indéterminismes de la liberté *avant* la révolution quantique était assurément dans une posture délicate ; mais inversement, si l'on conçoit cette révolution scientifique en termes Poppériens, on pourrait dire qu'elle a apporté une corroboration très forte de la « théorie de la liberté », puisqu'elle a permis « d'observer » les indéterminismes naturels dont seule cette théorie affirmait l'existence. Pourtant, on ne peut pas dire que la révolution quantique ait entraîné dans le monde scientifique ou simplement philosophique une brutale adhésion à l'existence du libertarianisme. Certes, il y a eu quelques philosophes et quelques rares scientifiques qui ont essayé rapidement d'appliquer l'indéterminisme quantique à la théorie de la liberté, mais cela est resté une faible minorité. Comment expliquer cette absence de « révolution libertarienne » ? Pourquoi les scientifiques

et philosophes ne se sont-ils pas mis à faire confiance à leur expérience indéterministe de l'action puisque cette expérience n'était plus réfutée par la science ?

La réponse à cette question est qu'une majorité de scientifiques ou de philosophes sont restés convaincus que l'indéterminisme quantique ne pouvait rien avoir à faire avec la question de la liberté humaine. Cette conviction a eu parfois un soubassement compatibiliste : puisque la liberté était *déjà* réinterprétée par certains en termes compatibilistes, l'indéterminisme quantique n'apportait rien d'intéressant. Mais même chez des penseurs incompatibilistes, on a rencontré majoritairement la conviction que *ce genre* d'indéterminismes n'était pas le genre d'indéterminismes qu'il fallait pour donner raison à notre expérience du libre arbitre. Cette difficulté-là n'est plus un problème de réfutation frontale du libre arbitre, mais plutôt une nouvelle objection d'intelligibilité : on ne peut pas facilement concevoir comment les indéterminismes quantiques pourraient être pertinents pour le libre arbitre.

II.2. L'objection d'intelligibilité empirique

De manière générale, l'objection d'intelligibilité empirique consiste à dire que les données des sciences contemporaines, si elles ne sont pas logiquement incompatibles avec le libre arbitre, le rendent néanmoins incompréhensible. C'est-à-dire qu'elles rendraient impossible d'en donner un modèle compréhensible. Dans ces termes généraux, l'objection peut se poser aussi bien contre les théories qui font reposer l'indéterminisme du libre arbitre sur l'indéterminisme de la physique fondamentale, que contre la théorie de la causalité agentive qui, comme on l'a annoncé, n'aura pas nécessairement recours aux lois physiques indéterministes. Dans cette sous-section, je ne m'intéresserai qu'aux premières, et je consacrerai le point II.3. à la causalité agentive et aux réponses qu'elle peut apporter à la fois à l'objection déterministe et à l'objection d'intelligibilité empirique.

Si donc on considère que l'indéterminisme du libre arbitre trouve sa source dans les indéterminismes de la physique fondamentale, comment comprendre le lien entre les deux ? Comment l'indéterminisme d'un électron dans le cerveau pourrait-il avoir quelque chose à voir avec le fait d'être libre ? Certes, il est *possible* scientifiquement qu'un indéterminisme quantique se répercute au niveau mésoscopique des neurones, et rien n'empêche que cet indéterminisme se situe dans le contexte neuronal qui correspond à une décision et à un choix. Mais même dans ces circonstances qui peuvent paraître exceptionnelles, comment l'indéterminisme en question pourrait-il avoir quelque chose à voir avec la liberté ? Pour répondre à cette question, il s'agira d'apporter un *modèle* de l'indéterminisme du libre arbitre. Il ne s'agira pas de montrer empiriquement que ce modèle est vrai, mais il suffira de montrer qu'il rend intelligible une notion de libre arbitre appuyée sur l'indéterminisme quantique. Je reprendrai les grandes lignes du modèle développé par Robert Kane (1996, 128–130).

Le modèle de Kane est construit comme une réponse à deux problèmes : premièrement, si l'on admet que les indéterminismes quantiques peuvent se répercuter au niveau de l'activation neuronale, comment faire pour qu'un indéterminisme dans l'activation d'un neurone puisse entraîner une différence remarquable à l'échelle macroscopique ? Deuxièmement, comment faire pour que l'indéterminisme neuronal joue un rôle dans la *décision* libre et pas à n'importe quel endroit du fonctionnement cérébral ? Ces deux questions se résument ainsi : comment faire pour que la différence entre activation / non

activation d'un neurone soit responsable de la différence entre choix 1 et choix 2 à l'occasion d'une décision ?

La réponse à la première question repose sur la théorie du chaos. De la même façon que l'indéterminisme quantique peut se répercuter au niveau d'un neurone du fait d'une sensibilité fine aux conditions initiales, c'est-à-dire d'une dépendance chaotique entre le niveau microscopique et le niveau mésoscopique, de même, l'indéterminisme neuronal peut se répercuter au niveau du fonctionnement global du cerveau si l'on suppose une dépendance chaotique entre ces deux niveaux, une sensibilité fine des processus cérébraux aux activations neuronales individuelles. Robert Kane fait observer qu'on peut trouver dans les sciences contemporaines l'un et l'autre type de dépendance chaotique. Pour la dépendance chaotique du niveau neuronal par rapport au niveau microscopique, il s'appuie notamment sur le chimiste Scott (1991, 265) qui souligne que « les vagues de potentiels d'action sur les membranes fines des cellules nerveuses [qui] représentent la principale activité de contrôle de l'activité humaine » sont des systèmes thermodynamiques de non-équilibre susceptibles d'amplifier des événements indéterminés au niveau quantique. Pour la dépendance chaotique entre le niveau neuronal et le niveau cérébral proprement dit, Kane s'appuie sur l'étude des réseaux neuronaux par le neuroscientifique Gordon Globus. Pour Globus (1995), les réseaux neuronaux fonctionnent comme des systèmes holistes au sens où « l'influence du réseau entier affecte chaque nœud [chaque neurone], et l'influence de chaque nœud [affecte] le réseau entier. [...] Apportez quelques changements dans les poids de connexion [des neurones individuels] et tout change fonctionnellement [dans le réseau neuronal] ». Grâce à ces deux processus chaotiques, il est possible de donner un modèle dans lequel des indéterminismes quantiques affectant certaines particules du cerveau se répercuteraient d'abord en indéterminisme neuronal et ensuite en indéterminisme cérébral.

Reste à s'assurer que ces indéterminismes cérébraux soient « bien situés » pour avoir quelque chose à voir avec le libre arbitre. Informellement, il faut que ces indéterminismes soient situés « à l'endroit du choix » (et en particulier du choix difficile). Mais quel est cet endroit du choix dans les processus cérébraux ? D'après Kane (2005, 137), le choix (difficile) est réalisé dans le cerveau par plusieurs réseaux neuronaux dont chacun, individuellement, correspond à une des motivations concurrentes de l'agent. Ces réseaux sont interconnectés et leur contribution causale résultera dans le choix de l'agent. La contribution causale de chaque réseau pris séparément correspond donc, au niveau mental, à l'effort que fait l'agent pour faire tel ou tel choix. On a déjà vu plus haut (p. 362) que le modèle de Kane reposait essentiellement sur l'existence de deux efforts contraires, qui permettent de dire, quel que soit le choix final de l'agent, que ce choix aura été produit de manière volontaire et contrôlée. Ce « contrôle dual » au niveau mental est donc représenté au niveau neuronal par l'interconnection de deux réseaux neuronaux. Plus précisément, Kane estime qu'il doit s'agir de « réseaux neuronaux récurrents », c'est-à-dire un type particulier de réseau neuronal fréquemment présent dans les processus cognitifs avancés, et qui a la particularité de présenter des boucles de rétroaction (*feedback loops*). Ces réseaux neuronaux semblent être de bons candidats pour assurer des conditions de fonctionnement holiste, et donc de chaos entre le niveau des neurones individuels et le niveau cérébral. De cette manière, on peut dire que « ces conflits [entre efforts concurrents] créent des tensions qui sont reflétées dans les régions appropriées du cerveau [et] qui augmentent la sensibilité aux indéterminismes microscopiques au niveau neuronal, et amplifient ces indéterminismes à travers le processus macro complexe. » (Kane 1996, 130). Pour résumer, le modèle de Kane est donc le suivant :

Modèle de Kane

Lorsqu'un agent S fait un choix difficile, ses deux efforts concurrents sont représentés physiquement par deux réseaux neuronaux récurrents, connectés entre eux. Cette

configuration neuronale particulière présente un fonctionnement chaotique susceptible d'amplifier des indéterminismes microscopiques. Par ailleurs, des indéterminismes microscopiques existent au niveau des neurones individuels, si bien que l'issue des contributions causales des deux réseaux neuronaux est réellement indéterministe.

Le modèle de Kane constitue une réponse au problème d'intelligibilité empirique pour deux raisons. Tout d'abord, il donne une traduction « scientifique » (compatible avec la science contemporaine) de la solution au problème général d'intelligibilité : au problème général, Kane répondait que l'indéterminisme du choix n'est pas un « pur hasard » parce qu'il correspond à deux efforts au niveau mental ; le modèle de Kane montre que cette solution anti-hasard est susceptible d'une traduction neuroscientifique précise. Deuxièmement, le modèle de Kane présente une configuration cérébrale dans laquelle le conflit des motivations est *responsable de* ou *explique* le caractère indéterministe du choix, puisque c'est le conflit de motivations qui « augmente la sensibilité aux indéterminismes microscopiques ». Donc ce n'est pas un pur hasard si les décisions que nous ressentons comme conflictuelles sont aussi indéterministes au niveau physique. Ce dernier point deviendra particulièrement important dans le prochain chapitre. Dans le contexte présent, il s'agit simplement de remarquer que l'indéterminisme est inséré par Kane à un endroit où il est étroitement dépendant du caractère indécis de nos choix difficiles. Ainsi, on ne peut pas dire que l'indéterminisme quantique du cerveau n'aurait « rien à voir » avec la liberté.

Du point de vue dialectique, on peut faire une dernière remarque sur le caractère « constructif » ou « progressif » du modèle de Kane ou d'autres modèles de même inspiration.³⁹¹ Un peu plus haut, nous nous sommes demandé pourquoi la découverte de l'indéterminisme quantique n'avait pas été massivement considérée comme apportant, en termes Poppériens, une corroboration de la « théorie du libre arbitre ». Dans la situation présente, il est clair que le modèle de Kane reste hautement spéculatif ou hypothétique, mais le caractère hypothétique d'une théorie n'est pas une faiblesse en sciences expérimentales dans la mesure où l'hypothèse est réfutable par d'éventuelles expériences futures. C'est en tout cas la leçon essentielle de Popper. Imre Lakatos (1978) a précisé cette leçon en établissant la distinction entre programmes de recherche « dégénératifs » et « progressifs » : lorsqu'une théorie donnée ne peut être défendue qu'en suscitant des hypothèses purement *ad hoc* et difficilement ou pas vérifiables, on peut dire qu'elle est dégénérative du point de vue empirique. Inversement, lorsqu'une théorie ouvre la possibilité d'enquêtes empiriques précises, on peut dire qu'elle constitue un programme de recherche « progressif ». D'après cette distinction, on peut dire, me semble-t-il, que le modèle de Kane constitue un programme de recherche progressif pour les neuroscientifiques : rien n'empêche de mener des investigations scientifiques sur la base neuronale de nos choix, de nos motivations, et de nos efforts de volonté, pour déterminer si l'on y trouve effectivement des réseaux neuronaux récurrents, ou d'autres types de réseaux neuronaux chaotiques. Bien que nous n'ayons pas aujourd'hui de données sur ces questions (à ma connaissance), il semble clair que les progrès des neurosciences devraient ouvrir dans les années ou les décennies à venir la possibilité de tester empiriquement la viabilité du modèle de Kane, en suscitant éventuellement des nouveaux modèles chaotiques de la décision en fonction des nouvelles données. Autrement dit, le modèle de Kane a, à mon sens, le caractère d'une authentique hypothèse scientifique, bien qu'il s'agisse d'une hypothèse dont la mise à l'épreuve et l'éventuelle amélioration restent dans l'immédiat assez difficiles (comme l'était l'hypothèse du boson de Higgs il y a à peine quelques années). Il ne s'agit donc pas d'une hypothèse « purement » spéculative ou gratuite,

³⁹¹ Balaguer (2010, 157) donne également un modèle dans lequel le libre arbitre dépend des indéterminismes quantiques du cerveau. Mais il est beaucoup plus schématique, et donc moins explicatif, que celui de Kane.

mais d'une contribution constructive à l'investigation neuroscientifique de la base neuronale du choix.

II.3. La causalité agentive face aux objections empiriques

Les deux réponses que nous venons de voir, la réponse à l'objection déterministe et celle à l'objection d'intelligibilité empirique, sont celles qu'apportent principalement les théoriciens de la causalité événementielle pour le libre arbitre. La situation dialectique est sensiblement différente lorsqu'on se place dans le modèle de la causalité agentive. Dans ce modèle, tout ce que nous venons de dire (en particulier la dépendance aux indéterminismes quantiques) *peut* être utilisé, mais ce n'est pas une nécessité, car la causalité agentive présente d'autres ressources, des ressources propres, pour répondre aux deux problèmes. La réponse des théoriciens de la causalité agentive repose essentiellement sur la notion « d'émergence » ou encore de causalité « de haut en bas » (*top down* ou *downward causation*). Il faut donc envisager ces notions et leur caractère explicatif avant de voir pourquoi la causalité agentive peut utiliser, mais peut aussi se passer de, la dépendance aux indéterminismes quantiques.

Pour les théoriciens de la causalité agentive, en particulier pour (O'Connor 2000, sec. 6.2), la causalité agentive est une propriété « émergente » des systèmes physiques hautement complexes que sont les corps humains. Qu'est-ce qu'une propriété émergente ? L'intuition fondamentale qui a présidé à l'invention de ce terme est l'idée que les phénomènes se produisant au niveau macroscopique du composé sont, dans un sens fort, irréductibles aux phénomènes microscopiques se produisant au niveau des composants. Des théories de l'émergence ont été proposées initialement pour rendre compte de la conscience humaine ou de la vie biologique. La notion a eu un grand succès au début du XX^{ème} siècle puis a connu un déclin important qu'O'Connor attribue à deux facteurs. D'une part, le fait que la notion de code biologique a pu être finalement réduite aux composants bio-chimiques de l'ADN, contre les théories émergentistes qui s'appliquaient à ce problème particulier. D'autre part, l'émergentisme a souffert d'un défaut intrinsèque, la difficulté à donner de la notion d'émergence une définition unique et claire. Au premier de ces deux problèmes, on peut répondre que l'échec de l'émergentisme dans un domaine particulier n'empêche pas qu'il faille y avoir recours dans d'autres domaines, et notamment dans la théorie de l'agentivité humaine. Au deuxième problème, il faut répondre en offrant une définition plus précise de l'émergence, ce que fait O'Connor : « Lorsque nous appelons un certain phénomène 'émergent', nous voulons exprimer l'idée qu'il introduit dans le monde une caractéristique de niveau macro qualitativement nouvelle. [...] On peut reformuler ce réquisit en disant qu'une propriété émergente est *non structurelle* ». (O'Connor 2000, 111) Une propriété émergente s'oppose donc essentiellement à une propriété structurelle. Mais qu'est-ce qu'une propriété structurelle ? Une propriété structurelle d'un système complexe est une propriété dont « l'influence causale se produit [uniquement] à travers l'activité des micropropriétés qui le constituent » (*ibid.*). Une propriété émergente est donc une propriété dont l'influence causale ne se réduit pas aux influences causales des micropropriétés du système. Ainsi, dans un système non émergent, tout ce qui se produit au niveau macro est ultimement expliqué causalement par les propriétés du niveau micro : c'est ce qu'on appelle la causalité « de bas en haut » (*bottom up*). Inversement, dans un système aux propriétés émergentes, ce qui se produit dans le système au niveau macro (et corrélativement au niveau micro) est parfois expliqué causalement uniquement par une propriété de niveau macro. Les propriétés émergentes introduisent donc une contribution causale qui va directement du macro au macro, mais aussi

du macro au micro, ce qu'on peut appeler « causalité de haut en bas » (*downward* ou *top down causation*). Pour résumer, O'Connor donne de l'émergence la définition suivante :

Définition de l'émergence

Une propriété émergente est une propriété naturelle non structurelle, qui est exemplifiée par des objets ou systèmes lorsqu'ils atteignent un certain niveau et un certain type appropriés de complexité organisationnelle, et qui exerce de la causalité de haut en bas.

A partir de cette définition, on peut voir le lien étroit qui unit la théorie de l'émergence et la théorie de la causalité agentive. En effet, la théorie de la causalité agentive consiste à dire que certains systèmes (les cerveaux ou les corps humains) sont suffisamment complexes dans leur constitution pour qu'apparaisse en eux une certaine propriété, la propriété de pouvoir *causer* tel ou tel choix ou telle ou telle action. Dans la mesure où cette propriété est une propriété du système dans son ensemble (de la personne), et dans la mesure où son influence causale consiste à produire des mouvements (au moins des « mouvements » cérébraux) qui sont réalisés au niveau micro, il est clair que la causalité agentive est une propriété qui exerce la causalité de haut en bas. Si l'on pouvait réduire l'explication d'une action à la contribution causale des processus micro se produisant dans le cerveau, alors dans la mesure où ces contributions causales relèvent de la causalité « événementielle », il est clair qu'on ne se trouverait plus dans le cadre de la théorie de la causalité agentive. La causalité agentive est donc nécessairement conçue comme une propriété émergente des êtres humains comme systèmes complexes.

Il existe diverses objections à la théorie de l'émergence. J'en considérerai ici deux principales, qui sont particulièrement importantes pour comprendre la réponse aux objections empiriques.

La première objection consiste à dire que les propriétés émergentes sont contraires à l'esprit ou à la démarche scientifique elle-même. La démarche scientifique consisterait à analyser les phénomènes complexes en réduisant l'explication de leur fonctionnement aux contributions causales des composants. Pour répondre à cette objection, il faut d'abord concéder que « la stratégie de micro-réduction a connu des succès énormes dans la science théorique moderne » (O'Connor 2000, 115). Mais ces succès ne permettent certainement pas d'affirmer que dans tous les domaines et absolument la microréduction est possible. Pour des propriétés comme la chaleur ou la couleur, on peut dire que la microréduction était « attendue » ou raisonnablement espérable avant d'avoir été effectuée. Mais s'il y a bien un domaine de la réalité où les succès d'une microréduction future ne sont pas indiscutablement espérables, c'est bien le domaine de l'action humaine. Savoir si la microréduction pourra ou non fonctionner n'est pas, en l'état actuel de nos connaissances, une question scientifique, mais une question d'attentes métaphysiques. Il se pourrait tout à fait que la microréduction échoue dans ce domaine. Par ailleurs, si la microréduction échouait pour décrire l'action humaine, cela ne voudrait pas dire que « l'approche scientifique » en est impossible. L'approche scientifique ne se réduit pas à l'ambition d'analyses réductivistes. En particulier, (O'Connor 2000, 112) souligne qu'un phénomène émergent peut tout à fait être « explicable naturalistiquement » à condition de bien spécifier les « conditions de base causale », c'est-à-dire les caractéristiques micro et structurelles qui doivent être réunies pour qu'il y ait émergence d'une nouvelle propriété (non structurelle). Les propriétés émergentes n'émergent pas au hasard mais dans des conditions naturelles précises et scientifiquement étudiables. Il n'y a donc pas d'incompatibilité entre une approche scientifique et naturaliste de l'action humaine et l'idée d'une propriété émergente de causalité agentive.

Les conditions de base causale nous amènent à la deuxième objection contre l'émergence. O'Connor la formule ainsi :

Si une propriété émergente est une conséquence nécessaire de certaines propriétés au niveau de la base, alors son instanciation est une des potentialités de cet ensemble de propriétés. Par conséquent, les potentialités *additionnelles* de cette propriété émergente ne sont-elles pas aussi un sous-ensemble des potentialités des propriétés de la base ? [...] On est alors amené à se demander pourquoi on irait imaginer une propriété émergente, vu qu'elle n'apporte aucun gain explicatif par rapport à une théorie qui supprime l'étape intermédiaire et rattache directement les potentialités 'additionnelles' aux propriétés de la base.

(O'Connor 2000, 112)

L'objection revient donc à dire que les propriétés émergentes sont un postulat superflu du point de vue explicatif, si du moins on suppose des conditions de base causale. Cette objection a deux prémisses :

- (1) les potentialités des propriétés émergentes pourraient être considérées directement comme des potentialités des propriétés de la base.
- (2) si (1) est vraie, alors il n'y a aucun gain explicatif à postuler une propriété émergente.

La prémisse (1) n'est pas vraiment contestable. C'est donc la prémisse (2) qu'on peut essayer de mettre en cause. O'Connor réfute l'objection en montrant sur un exemple particulier la nature du gain explicatif.

Imaginez que [la propriété émergente] *P* est instanciée dans les systèmes ayant un haut niveau de complexité *n*. Imaginez par ailleurs que les physiciens utilisent un ensemble de lois fondamentales, *L*, pour décrire avec précision les processus physiques dans tous les systèmes dont le niveau de complexité est inférieur à *n*, mais que les lois échouent à rendre compte des systèmes complexes de niveau *n*. Dans ces conditions, il y aurait de bonnes raisons de supposer qu'une (ou plusieurs) propriété émergente est à l'œuvre. (O'Connor 2000, 112–113)

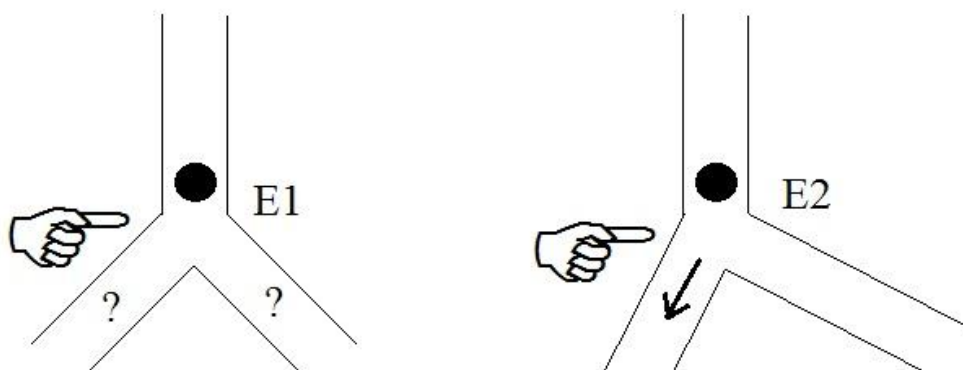
En postulant une propriété émergente, on apporterait une *explication* du fait que les systèmes se comportent différemment au-dessous et au-dessus du degré de complexité *n*. Par exemple, on expliquerait pourquoi un neurone ne se comporte pas de la même manière selon qu'il est inclus dans système de la complexité du cerveau humain, ou selon qu'il est séparé de ce système complexe. Cela ne remet pas en cause la prémisse (1), comme le montre O'Connor : les scientifiques *pourraient* choisir de ne pas postuler de propriétés émergentes et de réviser l'ensemble de lois *L* qui régissent les propriétés de la base : pour ce faire, ils devraient introduire dans leurs lois une disjonction de la forme « mais lorsque le système atteint le niveau *n*, le comportement sera tel et tel ». Mais une telle disjonction serait assurément une bizarrerie qui appelle une explication. Puisque les propriétés de la base ne se comportent pas de la même façon selon le « contexte » où elles se trouvent – selon qu'elles appartiennent ou non à un système de complexité *n* – il y a un gain explicatif à postuler une nouvelle propriété au niveau *n* de complexité, une propriété dont l'apparition est responsable du changement de comportement.

La réponse à cette deuxième objection est très importante pour comprendre le rapport entre déterminisme et théorie de la causalité agentive. Nous venons de voir en effet que,

lorsqu'on postule une propriété émergente, on peut distinguer entre deux ensembles de lois régissant les propriétés de la base : d'une part, les lois qui régissent ces propriétés lorsqu'elles ne sont pas dans le contexte d'un système assez complexe pour faire apparaître une propriété émergente ; d'autre part, les lois qui régissent ces propriétés lorsqu'elles se trouvent dans un tel contexte de propriété émergente. Et ces deux ensembles de lois n'ont pas de raison d'avoir les mêmes caractéristiques. En particulier, l'un pourrait être un ensemble de lois déterministes, et l'autre pourrait inclure des lois simplement stochastiques. Par conséquent, il se pourrait très bien que les lois qui régissent le fonctionnement des neurones *pris séparément* (hors contexte complexe du cerveau humain) soient des lois purement déterministes, et que néanmoins dans le contexte complexe du cerveau humain (où émerge une propriété de causalité agentive), tel neurone puisse être indéterminé dans son fonctionnement. L'important pour un auteur libertarien comme le théoricien de la causalité agentive, c'est que le fonctionnement du cerveau soit parfois indéterminé « en contexte » ; mais dans une théorie de l'émergence, cela est compatible avec le fait que les lois de fonctionnement des neurones *pris séparément* soient des lois déterministes. O'Connor (2000, 123) a décrit cette éventualité en disant que les processus de base seraient alors « quasi-déterministes » : « en dehors des contextes macroscopiques qui présentent le bon degré et le bon type de complexité pour l'apparition de caractéristiques émergentes, ces processus respectent des lois qui sont non statistiques. »

Cette possibilité a une conséquence considérable dans la discussion de l'objection déterministe. En effet, à supposer qu'il y ait un jour une preuve empirique du déterminisme micro ou bien du déterminisme neuronal, ces preuves consisteraient normalement à mettre en évidence des lois strictement déterministes du fonctionnement des particules ou des neurones *pris hors contexte*. Mais de telles lois ne remettraient pas en cause l'indéterminisme *en contexte complexe* d'émergence de propriété. Par exemple, si l'on suppose qu'au XIX^{ème} siècle les scientifiques étaient fondés à admettre le caractère déterministe des lois de la physique fondamentale, ce caractère déterministe des particules fondamentales *hors contexte d'émergence* n'aurait pas permis d'établir le caractère déterministe des systèmes complexes constitués de ces particules. Autrement dit, les « preuves du déterminisme » du XIX^{ème} siècle ne constituaient pas une objection au libertarianisme dans sa version de la causalité agentive. L'indéterminisme macro du système émergent ne *repose pas* sur des indéterminismes dans les lois qui régissent (hors contexte) les composants micro ; en particulier, l'indéterminisme de la causalité agentive est indépendant des indéterminismes quantiques.

Pour bien comprendre cette indépendance, on peut comparer l'influence causale de l'agent sur un neurone à l'influence causale d'un joueur sur une bille de flipper. Imaginez que la bille de flipper tombe dans un rail puis arrive à un embranchement E1 où elle peut aller soit à gauche soit à droite, et se « décidera » de manière indéterministe si le joueur n'intervient pas. Par ailleurs le joueur a la possibilité d'intervenir au moment où la bille parvient à E1, et de la faire tomber à gauche ou à droite (par exemple en la poussant du bon côté). Imaginez à présent un autre embranchement, E2, où elle peut arriver. Si elle arrive à E2, elle est déterminée à aller à gauche si le joueur n'intervient pas, mais là encore le joueur peut intervenir pour la faire aller à droite.



Le joueur peut donc intervenir dans deux cas de figure différents : dans un cas, il choisit une détermination pour un trajet qui, même sans la possibilité de son intervention, serait purement indéterministe ; dans l'autre, il choisit une détermination qui peut contrevenir à ce qui, autrement, serait déterminé différemment. Dira-t-on que le joueur a plus de contrôle dans le premier cas que dans le second, parce que dans le premier cas le mouvement de la bille était indéterministe quoi qu'il arrive ? Non, évidemment : ce qui compte, c'est qu'il puisse intervenir pour déterminer le mouvement de la bille dans le sens qu'il souhaite ; peu importe ce que la bille ferait si il n'intervenait pas. De la même façon, il est évident que les actions libres des agents entraîneront une forme d'influence causale « de haut en bas », du niveau macro vers le niveau neuronal. On peut donc imaginer que cette influence s'effectue sur un mouvement neuronal qui, autrement, serait indéterminé (pour des raisons quantiques), ou bien sur un mouvement neuronal qui, autrement, serait déterminé dans une direction différente, cela ne change rien à l'influence causale dont disposera (au niveau macro) l'agent. Pour cette raison, la théorie de la causalité agentive *pourrait* accepter tout ce que nous avons dit plus haut à propos des indéterminismes neuronaux – il *se pourrait* que notre influence causale agentive revienne à déterminer des mouvements neuronaux qui autrement ne seraient pas déterminés – mais elle n'a pas *besoin* de reposer ainsi sur des indéterminismes préexistants.

Plus haut, je me suis efforcé de montrer que le libertarianisme n'était pas réfuté par l'objection du déterminisme *parce que* nous n'avons pas (en l'état actuel des sciences) de preuve du déterminisme. Nous voyons à présent que la causalité agentive permet d'apporter une réponse plus originale : *même si* nous avons des preuves du déterminisme micro, cela n'empêcherait pas les systèmes complexes d'être indéterminés (et de « rendre » indéterminés leurs composants micro). Que penser de cette deuxième stratégie pour répondre à l'objection déterministe ? Elle présente des avantages et des inconvénients.

Un premier avantage est de rendre la thèse libertarienne indépendante des indéterminismes quantiques ; pour les personnes qui ont l'intuition que les indéterminismes quantiques ne peuvent pas être « pertinents » pour le libre arbitre, cette indépendance est certainement un avantage. Même si j'ai essayé de montrer qu'on pouvait présenter des modèles qui rendent compréhensible la dépendance aux indéterminismes quantiques, ces modèles sont assurément des réponses complexes à un « problème de pertinence », et à ce problème la solution par la causalité agentive peut apporter une solution beaucoup plus simple : le libre arbitre pourrait très bien n'avoir en effet rien à voir avec l'indéterminisme quantique.

Deuxième avantage, le libertarianisme de la causalité agentive ne dépend pas des résultats, déterministes ou non, de la physique fondamentale. Comme on vient de le voir, les « preuves déterministes » du XIX^{ème} siècle n'auraient pas été des réfutations de cette théorie, et par conséquent, la théorie peut être considérée comme « historiquement » plus résistante, par comparaison aux théories à la Kane qui ne sont rendues possibles que par l'état actuel,

historiquement situé, de la physique fondamentale. Ce deuxième avantage me permet également une remarque dialectique : si dans les années à venir la physique fondamentale devenait déterministe – soit que le modèle quantique soit remplacé par un modèle plus fondamental et déterministe, soit que le modèle quantique soit finalement interprété de manière claire comme déterministe – alors les modèles à la Kane seraient réfutés. Mais dans la mesure où notre justification expérientielle est une justification à croire que nous sommes indéterminés, et non pas une justification pour tel ou tel modèle particulier, il est clair que cette justification ne serait pas réfutée tant que resterait ouverte la possibilité de la causalité agentive. Par conséquent, la causalité agentive constituerait évidemment ma « position de repli » en cas de preuve du déterminisme au niveau de la physique fondamentale. Mais en l'état actuel des sciences, je ne vois pas de raison de favoriser l'un plutôt que l'autre.

L'avantage de « résistance aux résultats des sciences » présente un revers de la médaille : on pourrait reprocher à la causalité agentive d'être trop résistante aux résultats des sciences. Là où les théories à la Kane sont réfutables et pour l'instant non réfutées, l'indéterminisme émergent de la causalité agentive risquerait d'être absolument irréfutable, et d'échapper ainsi au critère popperien de scientificité. On peut partiellement répondre à cette inquiétude : en effet, dire que l'indéterminisme émergent ne repose pas sur les indéterminismes micro, ce n'est pas dire qu'il est scientifiquement totalement indétectable. Comme on l'a vu plus haut, O'Connor imagine même un scénario dans lequel des observations scientifiques obligerait à postuler une propriété émergente : c'est le scénario dans lequel on observerait un comportement *différent* des composants micro selon qu'ils sont pris dans un système de complexité n ou selon qu'ils sont pris hors d'un tel contexte d'émergence. Par exemple, il n'est pas exclu *a priori* qu'on puisse un jour observer un neurone, pris dans un système neuronal agentif, dont le fonctionnement serait contraire à ce qu'il devrait être s'il était hors d'un tel système. Une telle observation nous amènerait à postuler qu'il y a eu influence causale « de haut en bas » et éventuellement causalité agentive. Certes, une telle observation ne semble pas vraiment à notre portée en l'état actuel des neurosciences, mais il est important de noter qu'elle n'est pas exclue *a priori* et que la causalité agentive n'est pas donc exclue par principe de toute confirmation ou réfutation expérimentale.

Par rapport aux théories à la Kane, la théorie de la causalité agentive a donc un rapport très différent aux objections empiriques, et d'une manière générale aux données expérimentales. L'objection déterministe proprement dite ne peut pas fonctionner contre cette théorie car toute « preuve du déterminisme » aurait la forme de lois déterministes du niveau fondamental ou neuronal, et de telles lois déterministes sont totalement compatibles avec l'indéterminisme émergent de la causalité agentive : l'indéterminisme postulé par cette théorie se situe d'abord au niveau macro et ne repose pas sur des indéterminismes « préexistants » au niveau micro. Du même coup, l'objection d'intelligibilité empirique n'a plus lieu d'être non plus : nous avons vu que cette objection reposait sur l'intuition que les indéterminismes quantiques « n'ont rien à voir » avec la liberté, mais la causalité agentive peut tout à fait admettre cette intuition.

La seule difficulté de la causalité agentive dans son rapport aux données empiriques, c'est qu'on pourrait lui reprocher d'être une théorie trop « métaphysique », en discontinuité voire en opposition avec la démarche scientifique elle-même. De fait, ce qui permet à cette théorie d'être indépendante des résultats de la physique fondamentale, c'est un appareil métaphysique particulier, la théorie de l'émergence. Mais nous avons défendu que cet appareil n'était pas du tout incompatible avec la démarche scientifique, et qu'en outre la théorie de la causalité agentive restait ouverte, au moins en droit, à une forme d'observation empirique, bien que cette observation semble aujourd'hui loin de nos capacités scientifiques.

III. Conclusion : une « information recevable »

Tout l'enjeu de cette étude est de savoir si notre expérience du libre arbitre nous apporte ou non une certaine connaissance. Dans la première partie, nous avons vu que, avant même de savoir ce que cette expérience « nous dit », nous avons de bonnes raisons de présumer que ce qu'elle nous dit est vrai et apporte une connaissance. En un mot, nous avons de bonnes raisons *prima facie* de nous fier à son témoignage. Mais ces raisons pourraient être réfutées une fois qu'on prend en considération le contenu exact de ce que nous dit cette expérience. Dans le chapitre 4, nous avons vu que cette expérience nous représentait comme étant à la source ultime de nos actions et comme ayant des alternatives réellement possibles et indéterminées. Reste à savoir ce que nous pouvons faire d'une telle information ou d'un tel témoignage.

Lorsqu'un témoin nous apporte une information sur un certain événement, cette information peut souffrir de deux défauts très différents : d'une part, l'information peut être évidemment fausse, soit nécessairement fausse, soit rendue fausse par de meilleurs témoins ou une observation directe. Dans ce cas, je dirai que l'information est *irrecevable*. Mais lorsque l'information est *recevable*, elle peut encore avoir un deuxième type de défaut : elle peut n'être pas recevable *de ce témoin*. Par exemple, quelqu'un me dit que la voiture bleue est rentrée dans la voiture rouge parce qu'elle n'a pas freiné. Cette information est compatible avec tout ce que je sais. Elle est donc recevable : c'est une information que je *pourrais* ajouter à mon système de croyance. Mais par ailleurs, la personne qui me donne ce témoignage était absente du lieu de l'accident et n'a pas pu rencontrer d'autre témoin. Par conséquent, bien que je puisse recevoir, en principe, une telle information, je ne peux pas la recevoir *de cette source*. Je dirai donc que l'information est *mal sourcée*.

Ce que j'ai examiné dans le présent chapitre, c'est la question de savoir si l'information délivrée par notre expérience agentive était une information « recevable », c'est-à-dire si « ce que nous dit » notre expérience du libre arbitre est compatible avec ce que nous savons par ailleurs, soit *a priori* soit scientifiquement, de nous-mêmes et de nos actions. Et après avoir envisagé les différentes objections possibles, nous pouvons conclure que cette information du libre arbitre est bel et bien une information recevable, quelque chose que nous pourrions tout à fait ajouter à notre système de croyance sans entraîner ni contradiction ni incohérence. Pour recevoir effectivement l'information, il faudra s'assurer qu'elle est en outre « bien sourcée » ; c'est ce que nous étudierons dans le chapitre 6. Mais pour l'instant, nous pouvons déjà établir que notre expérience agentive ne nous dit rien d'impossible ou d'évidemment faux : ce qu'elle « prétend » être le cas est une des possibilités laissées ouvertes par notre connaissance de nous-mêmes et de nos actions, la possibilité que nos actions soient réellement indéterminées et que nous en soyons la source ultime.

Rappelons cependant que, pour établir ce résultat, nous avons dû envisager deux modèles distincts, qui tous les deux rendent concevable la notion de libre arbitre indéterministe : le modèle de la causalité événementielle (modèle à la Kane) et le modèle de la causalité agentive. Chacun de ces deux modèles a, comme on l'a vu, ses avantages et ses inconvénients. Nous verrons d'autres éléments de comparaison au chapitre suivant. Mais en l'état actuel de nos connaissances et de mes réflexions, il ne me semble pas requis ni profitable d'exclure l'un ou l'autre. Tous les deux sont, à mon sens, compatibles avec notre expérience agentive : dans la mesure où ce qui nous intéresse ici est de savoir si *cette* expérience est véridique, il est plus profitable de laisser ouvertes les différentes manières dont

cette expérience pourrait être rendue vraie par les faits. Trivialement, la disjonction de ces modèles a plus de chances d'être vraie que chacun pris séparément.

Rappelons néanmoins les points de comparaison déjà envisagés. On a vu d'abord que la causalité agentive pouvait prétendre (quoi que de manière tout à fait contestable pour l'instant) à un certain avantage phénoménologique, car la « grammaire » de l'agent comme sujet correspond sans doute à notre manière naturelle de décrire linguistiquement notre expérience agentive. Deuxièmement, on a vu que la causalité agentive était moins soumise aux aléas de la physique fondamentale et, en particulier, qu'elle pouvait très bien être défendue dans le contexte déterministe qui a précédé la physique quantique. Enfin, la causalité agentive peut faire droit à l'intuition que la liberté n'a « rien à voir » avec les arcanes des indéterminismes quantiques : le vrai niveau du libre arbitre, c'est le niveau macro, celui où chacun comprend ce qu'est être libre, pas le niveau abscons des particules élémentaires. On pourrait résumer ces trois avantages de la causalité agentive en disant qu'il s'agit d'une théorie plus « traditionnelle » : elle est plus traditionnelle dans le sens où elle essaye de s'attacher davantage à notre expérience commune, d'en rester au vocabulaire et au niveau ontologique de cette expérience, et dans le sens où elle a toujours pu être soutenue jusqu'à présent, dans tous les contextes scientifiques particuliers.

Inversement, les théories à la Kane présentent l'avantage d'être plus « scientifiques » : elles sont assurément réfutables par une découverte déterministe en physique fondamentale, mais une théorie réfutable qui n'est pas réfutée est une théorie (provisoirement) *corroborée*. La mécanique quantique peut donc être considérée comme une corroboration de ces modèles, à la différence de la théorie de la causalité agentive. Par ailleurs, ces théories à la Kane présentent l'avantage d'être métaphysiquement plus économiques : elles ne postulent aucune forme originale de causalité, aucune entité ou propriété émergente, *a fortiori* aucune âme. Tous les « ingrédients » de ces théories sont des ingrédients dont n'importe quel philosophe, si naturaliste et réductionniste soit-il, accepte déjà l'existence, et la seule question qui se pose est celle de savoir si les ingrédients sont disposés de la bonne façon ou non pour qu'il y ait de l'indéterminisme bien placé. Or Balaguer et Kane ont fait beaucoup pour montrer que, en l'état actuel de nos connaissances, *cette* question est une question tout à fait ouverte. Ce qui, encore une fois, revient à dire que l'*information* libertarienne donnée par notre expérience est une information tout à fait recevable.

Ces avantages comparés des deux théories pourraient suggérer que les préférences de chacun seront liées à la méta-philosophie de chacun : selon qu'on adopte une méta-philosophie plus traditionnelle et de sens commun, ou plus naturaliste et scientifique, on penchera respectivement dans le sens de la causalité agentive ou de la causalité événementielle. En ce qui me concerne, je ne crois pas qu'il y ait ici davantage que des « inclinations » : comme je l'ai montré, on peut tout à fait être naturaliste et accepter la causalité agentive, et inversement, on peut tout à fait vouloir défendre notre expérience phénoménale dans son sens traditionnel et estimer que ce sens traditionnel est celui de la causalité événementielle (ou compatible avec la causalité événementielle). La seule chose qui pourrait me porter à choisir un modèle, ce serait que l'autre modèle soit effectivement réfuté par une objection. Par exemple, si l'on parvenait à me convaincre que les modèles à la Kane ne répondent pas bien à l'argument de la chance, je m'engagerais en faveur de la causalité agentive. Mais pour l'instant, il ne me semble pas qu'il y ait aucune objection décisive contre l'un ou l'autre modèle.

Chapitre 6 : Absence de réfutation fondationnelle : la croyance libertarienne comme connaissance

It is one thing to argue that it must be possible to meet the [Integration] challenge. It is another thing to say how to meet it.

(Peacocke 1999, 6)

I. Réfutations par la source faible	402
I.1. Libet : la conscience d’agir comme épiphénomène retardataire	403
I.2. Wegner : les erreurs de la conscience d’agir	411
II. Réfutation par le contenu alourdi : le Problème d’Intégration.....	417
II.1. L’argument de réfutation fondationnelle et son principe épistémologique...	418
II.2. Conditions à satisfaire pour être une <i>connaissance</i>	422
II.3. Modèle d’intégration pour l’indéterminisme du libertarianisme événementiel	425
II.4. Modèle d’intégration pour l’indéterminisme de la causalité agentive	433
II.5. Conclusion : la possibilité d’une <i>connaissance</i> introspective de l’indéterminisme.....	460

Résumons les acquis des chapitres précédents. Au chapitre 2, nous avons vu que notre expérience du libre arbitre nous apportait une justification *prima facie* à croire à son contenu, en amont de toute analyse conceptuelle de ce contenu. Au chapitre 3, nous avons vu que nous étions également justifiés *prima facie* à tenir cette expérience du libre arbitre pour une source de *connaissance*. Et, si le chapitre 1 a raison de focaliser notre attention sur cette expérience, alors tout le « problème du libre arbitre » consiste à savoir si ces justifications *prima facie* rencontrent des réfuteurs ou bien si elles se préservent en justification *ultima facie*. Le « point de départ épistémologique » de la question du libre arbitre est donc le suivant : avant toute analyse du contenu de notre expérience et avant toute information empirique sur notre expérience, nous avons toutes les raisons de nous y fier, c’est-à-dire de la tenir pour une source de connaissance et de croire à son contenu de manière « indexicale ». Restent alors deux grands types de réfuteurs possibles : il se pourrait que l’analyse conceptuelle fasse apparaître pour notre expérience un contenu démontrablement faux (réfuteur direct) ; ou bien il se pourrait que diverses informations viennent réfuter la possibilité pour notre expérience d’être une source de *connaissance* (réfuteur fondationnel). Au chapitre 5, nous avons tâché de montrer que le contenu de notre expérience n’était pas démontrablement faux, bien qu’il soit relativement engageant métaphysiquement (d’après le chapitre 4, il implique l’indéterminisme causal) : l’information indéterministe de notre expérience est recevable si du moins on accepte de se fier à elle comme source de connaissance. La dernière question est donc celle de la réfutation fondationnelle : sachant que notre expérience du libre arbitre nous apporte une information recevable, et sachant que nous avons une justification *prima facie* à nous y fier, avons-nous des raisons empiriques de ne pas nous y fier ? C’est à cette question que nous répondrons dans ce dernier chapitre.

Une réfutation fondationnelle consistera donc à soutenir que notre expérience du libre arbitre ne peut pas constituer une bonne source ou un bon fondement pour procurer la

connaissance du libre arbitre. On peut arriver à une telle conclusion de deux manières différentes. Un trépied peut constituer un mauvais support pour un pot de fleurs soit parce que le trépied en lui-même est fêlé ou défaillant, soit parce que le pot de fleur est trop lourd pour ce trépied. Dans le domaine épistémologique, lorsqu'une croyance B est initialement justifiée par une source épistémique S, il existe également deux manières de faire échouer ce « support » épistémique : soit en « affaiblissant » la source épistémique elle-même (par exemple en montrant que tel témoin n'est pas fiable), soit en « alourdissant » le contenu de la croyance de telle sorte que la source épistémique ne soit plus en mesure de la supporter. Ces deux types de réfutations sont « fondationnelles » dans la mesure où elles mettent en jeu un échec de la relation de support épistémique entre la source et la croyance. Mais elles fonctionnent néanmoins de façons tout à fait différentes. J'appellerai la première « réfutation par la source faible » et la seconde « réfutation par contenu alourdi ».

Les réfutations par contenu alourdi ont été, à ma connaissance, peu étudiées en épistémologie contemporaine. Un autre exemple pourra donc aider à comprendre leur nature.

Réfutation par contenu alourdi d'une théorie physique

Imaginez que le professeur Phyléas, prix Nobel de physique, prétende avoir établi une certaine théorie physique, la théorie T. La communauté des physiciens trouve ses arguments très convaincants, et rapidement, tout le monde est convaincu que la théorie T est très probablement vraie. Compte tenu de l'autorité du professeur Phyléas et de la communauté des physiciens qui lui donne raison, nous avons toutes les raisons de croire que T est vraie.

Mais quelques années plus tard, le mathématicien Marcel, médaille Fields, parvient à montrer un résultat étonnant : il montre que si l'on admet l'équation fondamentale de la théorie T, on peut en déduire logiquement la vérité d'un certain principe mathématique, la conjecture d'Hypothrace. Comme son nom l'indique, la conjecture d'Hypothrace n'est justement qu'une *conjecture*, une hypothèse qui résiste à tous les efforts de démonstration par les mathématiciens depuis plusieurs siècles.

Devant ce résultat de Marcel, certains physiciens sont enthousiasmés : pour eux, Phyléas aurait non seulement établi une théorie physique, mais il aurait aussi, du même coup, prouvé une conjecture mathématique qui résistait depuis des siècles. Pourtant, la communauté des mathématiciens, puis rapidement la majorité des physiciens, se range à un autre avis : pour eux, le résultat de Marcel a au contraire pour conséquence de remettre en cause la théorie T. Ce résultat montre que la théorie T a un présupposé mathématique qui requiert une preuve proprement mathématique. Pour accepter la théorie T, il faudrait d'abord que la conjecture d'Hypothrace soit démontrée. Finalement, la théorie T, qui était initialement acceptée par tous, est reclassée au rang d'hypothèse, et la communauté des physiciens s'attelle à formuler une autre théorie, dont les présupposés mathématiques seraient moins lourds.

Dans ce scénario, la théorie T est réfutée fondationnellement sans que sa source épistémique soit discréditée d'aucune manière : à aucun moment l'autorité scientifique de Phyléas n'est mise en question. Si la théorie est réfutée, ce n'est donc pas parce que la source épistémique serait « affaiblie », mais parce que le contenu initialement justifié se trouve avoir des conséquences plus lourdes qu'on ne le croyait, des conséquences telles que la source épistémique (un physicien) n'a plus d'autorité pour les justifier ; pour justifier ces conséquences (mathématiques), il faudrait un mathématicien et une preuve mathématique. Une preuve physique n'est pas le genre de chose qui peut établir une conjecture mathématique comme la conjecture d'Hypothrace. Ainsi, une information qui vient « alourdir » le contenu d'une croyance initialement justifiée peut constituer une *réfutation* de ce contenu.

Dans le cas de l'introspectionnisme libertarien, on peut retrouver ces deux types de réfutations fondationnelles : réfutation par la source faible ou réfutation par le contenu alourdi. On peut défendre que notre expérience agentive constitue un mauvais support pour l'information du libre arbitre soit en montrant que l'expérience agentive est en elle-même défailante, soit en montrant que l'information du libre arbitre est, en fait, trop lourde pour être supportée par un tel fondement épistémique. C'est le deuxième type d'objection qui constitue, à proprement parler, le « problème d'Intégration » du libertarianisme. Avec ce deuxième type d'objection, nous arrivons donc au cœur du problème qui est à mon sens le plus pressant dans le débat sur le libre arbitre. Mais pour être complet, et pour vraiment montrer que la justification du libre arbitre n'est pas réfutée, il faudra également aborder le premier type de réfutation fondationnel, car il est également représenté dans la littérature.

Plus précisément, la première section de ce chapitre sera consacrée à certains résultats empiriques, notamment rapportés par Benjamin Libet et Daniel Wegner, qui semblent indiquer (ou qu'on a interprétés comme des indications) que notre expérience ou notre conscience agentive est un simple épiphénomène totalement illusoire, qui ne saurait apporter aucune information sur le déroulement réel de nos actions. Si l'on accepte ce genre d'interprétation, ce n'est pas le contenu « lourd » ou incompatibiliste de notre expérience agentive qui l'empêche de constituer une source de connaissance : notre expérience agentive est discréditée même pour les aspects plus « légers » que défendrait un auteur compatibiliste. Ce qui est ainsi réfuté, c'est donc ce que j'ai appelé au chapitre 2 « l'introspectionnisme a-compatibiliste », le fait de fonder quelque croyance que ce soit sur notre expérience du libre arbitre.

Dans la deuxième section du chapitre, je tâcherai de répondre au problème d'Intégration. C'est-à-dire que je m'efforcerai de montrer que notre expérience agentive peut procurer une *connaissance* de son contenu indéterministe. Plus précisément, je reviendrai sur la notion anti-hasard de la connaissance et je montrerai qu'on peut construire des modèles dans lesquels il y a une forme de dépendance non hasardeuse entre notre expérience indéterministe et le fait de l'indéterminisme.

I. Réfutations par la source faible

Depuis une trentaine d'années, un certain nombre d'expériences neuroscientifiques ont été menées pour étudier notre expérience agentive. Les résultats de ces expériences ont été diversement reçus et diversement interprétés, mais plusieurs scientifiques ont estimé que ces résultats établissaient scientifiquement une conclusion radicale : la volonté consciente est une pure illusion (voir le titre de (Wegner 2002) : *L'illusion de la volonté consciente*). Les études les plus importantes en la matière sont une série d'expériences sur le temps de l'intention réalisées par Benjamin Libet. Daniel Wegner, qui propose dans (Wegner 2002) une interprétation des résultats de Libet, a ajouté également d'autres études indépendantes sur les erreurs de notre expérience agentive.

Si ces auteurs ont raison, alors notre expérience agentive est absolument discréditée : cette expérience consciente serait un simple épiphénomène mental, sans aucune corrélation véritable avec la manière dont nous agissons réellement, et elle ne saurait donc en aucun cas servir de « source d'information » pour connaître les caractéristiques de notre action. Ces auteurs estiment que l'expérience consciente est sans fiabilité, même pour les aspects les moins controversés de l'action, les aspects qu'un auteur compatibiliste admettra généralement, notamment sur le simple fait de savoir si c'est moi qui agis ou non. Si on ne peut plus faire confiance à notre expérience agentive sur ces aspects anodins, il est clair qu'*a fortiori* on ne

devra pas lui faire confiance pour des contenus plus « lourds », notamment le fait que nous agissions *de manière indéterministe*.

Heureusement, il y a de bonnes raisons de penser que ces conclusions et ces interprétations sont exagérément pessimistes. Dans cette section, je résumerai les arguments principaux qui permettent de répondre à Libet et Wegner. Ultimement, il me semble que nous n'avons pas aujourd'hui de raison empirique de nous défier systématiquement de notre expérience agentive. Par conséquent, l'attitude à adopter à son égard restera la confiance *prima facie* que nous avons établie au chapitre 3, c'est-à-dire la justification *prima facie* à la considérer comme source fiable de connaissance.

1.1. Libet : la conscience d'agir comme épiphénomène retardataire

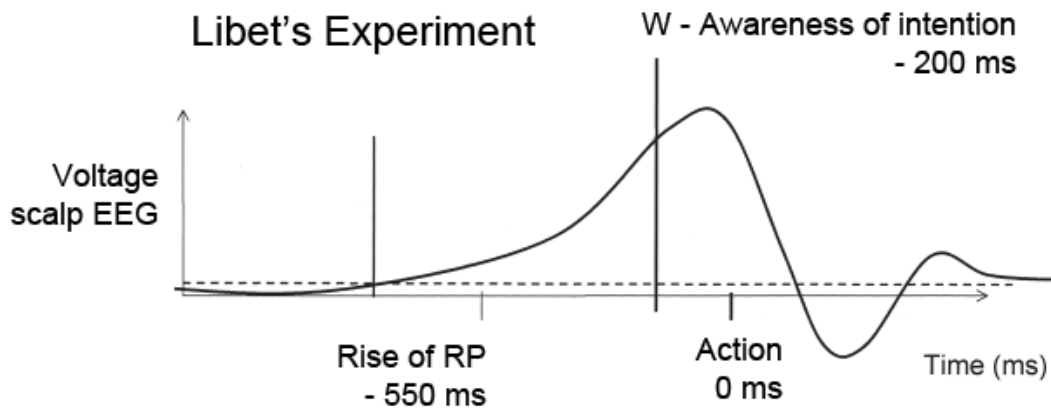
Pour bien comprendre les résultats de Libet, ainsi que les interprétations exagérées qui en ont été données, il est utile de repartir des découvertes qui ont précédé en neurosciences du mouvement volontaire. Au milieu des années 60, Kornhuber et Deecke (1965) eurent l'idée d'utiliser les techniques de l'électroencéphalogramme et de l'électromyogramme dans l'intention de repérer l'activité neuronale qui précède un mouvement volontaire. Les sujets de cette expérience devaient lever un doigt, et les appareils de mesure permettaient d'étudier avec précision le temps où l'action musculaire commençait ainsi que l'activité neuronale dans la période qui précédait cette activité musculaire. Les résultats de cette étude comblèrent toutes les espérances : juste avant le début de l'activité musculaire (environ 800 ms auparavant), l'électroencéphalogramme mit en évidence une activité cérébrale caractéristique, que Kornhuber et Deecke baptisèrent le « potentiel de préparation » (*Readiness Potential* ou *RP*). Dans l'euphorie de cette découverte, les neuroscientifiques acceptèrent unanimement que ce RP était exactement ce que l'on cherchait au départ, à savoir « un corrélat cérébral de la volonté consciente » (voir Wegner 2002, 52). Pourtant, comme le fait remarquer Wegner (*ibid.*), personne n'avait cherché à tester si ce RP neuronal correspondait ou non à quoi que ce soit de conscient, et en particulier au moment de la formation consciente d'une décision ou d'une volonté.

C'est précisément ce que voulut tester l'équipe de Benjamin Libet (Libet et al. 1983) au moyen d'un dispositif assez complexe.³⁹² Dans ces études, les sujets reçoivent l'instruction de plier un doigt de la main droite (ou leur poignet droit) quand ils le veulent. Utilisant également l'électroencéphalogramme et l'électromyogramme, ces études ont retrouvé le résultat de Kornhuber et Deecke, à savoir un potentiel de préparation (RP), dont le début est cette fois aux alentours de 550 ms avant l'enregistrement d'une première activité musculaire. Mais par ailleurs, on demande également aux sujets de regarder une horloge et de « se rappeler la position spatiale d'un point qui tourne sur l'horloge au moment où ils commencent à avoir conscience de l'intention ou de la volonté de bouger (W) » (Libet 1985, 529).³⁹³ Le but de cette deuxième instruction est bien sûr de pouvoir situer, dans la chronologie du RP et du mouvement, le « moment de l'intention consciente » (en se fiant à l'estimation qu'en font les sujets).

³⁹² Libet a mené, au cours des années 80s et 90s, un grand nombre d'expériences qui amélioraient l'expérience initiale et faisaient varier divers paramètres. Nous gardons ici les enseignements généraux de l'ensemble de ces expériences, tels qu'ils ont été réunis notamment dans (Libet 2004).

³⁹³ Cet événement (W) dont les sujets doivent tenter de situer le commencement est décrit de manière assez diverse dans les différentes expériences de Libet et al. : on demande au sujet de situer tantôt la décision de bouger le doigt (*decision*), tantôt l'intention (*intention*), mais aussi le souhait (*wish*), la volonté (*will* ou *wanting*), ou encore le désir (*urge*) – c'est d'ailleurs ce dernier terme (*urge*) que les sujets eux-mêmes semblent préférer pour décrire leur propre expérience.

Le résultat, moyenné sur un grand nombre de flexions du doigt pour chaque sujet, est le suivant : le moment estimé de l'expérience consciente W se situe en moyenne aux alentours de 200 ms avant le début (0 ms) de l'activité musculaire. Il se situe donc environ 350 ms après le début de l'activité neuronale RP. Ce résultat peut être résumé par le schéma suivant.



Après le résultat lui-même viennent les interprétations de ce résultat. Pour résumer, on peut considérer que deux grandes conclusions en ont été tirées. Tout d'abord, la conclusion de Libet lui-même : « le cerveau [...] 'décide' d'initier ou de se préparer à initier l'acte [...] avant qu'il y ait aucune conscience rapportée par les sujets qu'une telle décision s'est produite » (Libet et al. 1983, 640). Autrement dit : la conscience de la décision (W) est postérieure à la décision elle-même (que depuis Kornhuber et Deecke on situe au début du RP), ce qui veut dire que la décision elle-même, au moment où elle se produit, n'est pas quelque chose de conscient. Inversement, cela signifie que notre « expérience » de décision ou d'intention n'est pas réellement une *conscience* de cette décision puisqu'elle ne se situe pas au bon moment. Elle pourrait à la rigueur être un « souvenir » de la décision, mais elle ne se présente pas du tout ainsi : au contraire, il semble qu'il s'agisse d'une *illusion* de prendre la décision à -200 ms, alors que celle-ci a été prise à -550 ms. La première conclusion, celle de Libet, concerne donc le rapport temporel entre décision et conscience. La seconde conclusion, tirée notamment par Daniel Wegner ou Patrick Haggard, concerne le rapport causal entre intention consciente et action : puisque « l'intention consciente se produit *après* le début de l'activité cérébrale préparatoire[,] il est donc impossible qu'elle soit la cause de nos actions » (Haggard 2005, 291). L'idée est ici que la véritable cause de notre action est le potentiel de préparation neuronal, et que l'intention consciente n'a donc qu'une place d'épiphénomène, d'effet collatéral – « *a loose end* » pour (Wegner 2002, 55) – sans influence motrice propre.³⁹⁴ Là encore, notre expérience de la décision serait profondément illusoire, car cette expérience de la décision ne se présente pas seulement comme conscience (simultanée) de la décision, elle se présente également comme étant la cause de l'acte volontaire. Corrélativement, si notre intention consciente n'a aucun effet causal, cela signifie que « les véritables mécanismes causaux qui sous-tendent l'action ne sont jamais présents dans la conscience » (Wegner 2002, 97), c'est-à-dire que notre conscience est toujours « court-circuitée » (*bypassed*) par des mécanismes inconscients.

Si ces conclusions sont légitimes, s'il est vrai notamment que les expériences de Libet démontrent le court-circuitage constant des processus conscients, alors ce qui est réfuté n'est pas seulement une certaine conception forte et libertarienne de l'action libre, c'est la base de

³⁹⁴ Pour d'autres auteurs qui tirent cette deuxième conclusion dévastatrice, voir (Mele 2009, 71–72).

toute théorie de l'action libre, la plus compatibiliste soit-elle.³⁹⁵ Ainsi, notre expérience serait discréditée en amont de toute analyse compatibiliste ou incompatibiliste de l'action, c'est-à-dire que c'est l'introspectionnisme « a-compatibiliste » qui se trouverait alors réfuté.³⁹⁶

Mais ces conclusions radicales sont-elles réellement établies par les résultats empiriques de Libet *et al.* ? Si l'on accepte de décrire ces résultats avec une certaine précision conceptuelle, on peut s'apercevoir qu'il n'en est rien, ainsi que l'ont fait remarqué plusieurs philosophes, notamment (Mele 2009, chap. 3 et 4), (Nahmias 2010) ou (Roskies 2010).

On peut en réalité soulever de nombreux problèmes contre les résultats de Libet. Certaines objections ont tendance à refuser tout caractère significatif à ces résultats. Une de ces objections consiste à dire que la décision ou le choix qui est étudié dans ce dispositif expérimental n'est absolument pas représentatif du type de choix qui a des chances d'être réellement libre. Comme le fait remarquer (Mele 2009, sec. 4.3), de nombreux auteurs, par exemple van Inwagen ou Kane, estiment que le libre arbitre n'est possible que dans des situations où nous avons des raisons ou motivations significatives en faveur de deux options contraires, par opposition aux situations « de Buridan », dans lesquelles nous avons deux possibilités absolument indifférentes. Or le choix de plier ou non le doigt est clairement un choix d'indifférence ; on ne devrait donc pas s'attendre à y trouver les « traces » du libre arbitre.³⁹⁷ Une autre objection radicale consiste à remettre en question la fiabilité des estimations faites par les sujets à propos du moment W. La valeur retenue par Libet (-200 ms) n'est qu'une moyenne à partir d'estimations dont la variabilité est en fait considérable – de l'ordre de la seconde pour certains sujets, c'est-à-dire que les estimations extrêmes peuvent s'étendre d'avant le début du RP jusqu'à après le début du mouvement !³⁹⁸ Cette variabilité n'est d'ailleurs pas tellement étonnante si l'on prend en compte la difficulté objective de la tâche qui est demandée aux sujets (se rappeler *a posteriori* la position du point sur l'horloge *sans y avoir prêté attention* au moment de l'action volontaire elle-même). Enfin, on pourrait émettre des doutes encore plus fondamentaux sur le dispositif expérimental lui-même, en particulier sur l'instruction qui est donnée aux sujets de plier le doigt « lorsqu'ils le souhaitent » et de noter le moment de leur « intention / volonté / désir / souhait » conscient. Alfred Mele (2009, 34) rapporte qu'il s'est prêté au jeu d'une expérimentation de type Libet pour mieux la comprendre. Mele rapporte qu'il lui a semblé extrêmement difficile au premier abord de se souvenir d'un souhait ou d'une impulsion à plier le doigt, de telle sorte qu'il a finalement choisi d'adopter une « stratégie pour pouvoir suivre les instructions » : il a choisi de se dire mentalement (à des moments arbitraires) « Maintenant ! », comme un ordre intérieur, et de plier le doigt immédiatement en réponse à cette « décision ». Ce que révèle l'expérience de Mele, c'est que les instructions de Libet sont tout sauf limpides pour un sujet normal et que les différents sujets de l'expérience ont sans doute été obligés, pour pouvoir les suivre, de choisir une interprétation et une stratégie particulière – peut-être certains sujets ont-

³⁹⁵ Comme on l'a vu au chapitre 4 (p.295), l'hypothèse du *bypassing* est typiquement utilisée par les auteurs compatibilistes comme un archétype d'une situation qui exclut la liberté, mais que le déterminisme n'implique absolument pas.

³⁹⁶ Pour cette raison, il n'aurait pas été absurde de situer la discussion de Libet et Wegner *avant* le chapitre 4, puisqu'il s'agit d'une objection qui ne dépend pas de son résultat incompatibiliste. On aurait pu également choisir de placer cette discussion dans le chapitre des réfutations *directes* (chapitre 5) puisqu'en un certain sens, c'est bien un certain *contenu* (a-compatibiliste) de l'expérience agentive qui se trouve directement contesté. Si j'ai choisi de placer ici cette discussion, c'est qu'il s'agit d'une discussion (i) empirique (donc de deuxième partie plutôt que de première partie) et (ii) qui engage la *crédibilité* de notre expérience ou de notre conscience agentive et non pas la question du déterminisme en elle-même.

³⁹⁷ Pour une objection similaire, voir (Roskies 2010, sec. II.2.).

³⁹⁸ Pour une quantification précise de cette grande variabilité, voir (Haggard and Eimer 1999) cité et commenté par (Mele 2009, 124).

ils choisi *d'attendre* une impulsion à plier le doigt et d'agir en réponse à cette impulsion, tandis que d'autres auront choisi une stratégie proche de celle de Mele : « il y a plusieurs stratégies possibles, et [dans la version de Brass et Haggard] quatre sujets sur dix-neuf de la 'préexpérience' ont été exclus de l'expérience effective parce qu'ils 'se déclaraient incapables de suivre les instructions' (Brass and Haggard 2007, 9141) » (Mele 2009, 77).

A cause de toutes ces difficultés, on peut légitimement estimer que les conclusions générales et fracassantes de Libet et consorts mériteraient, tout au moins, de nombreuses précautions. Mais ce qui est plus important, c'est que même en admettant la pertinence du dispositif de Libet – même en admettant que son dispositif permet bel et bien d'établir avec fiabilité « le » moment de l'apparition de la conscience dans un processus de décision « typique » – même dans ce cas, les résultats de l'expérience ne justifient absolument pas les conclusions qui en ont été tirées.

Commençons par la deuxième conclusion mentionnée plus haut : les résultats de Libet permettraient d'établir que l'intention consciente n'a aucune influence causale sur le mouvement musculaire. Par conséquent, notre conscience serait systématiquement court-circuitée par l'efficacité causale de processus cérébraux non conscients (c'est-à-dire de processus cérébraux qui ne sont pas le corrélat neuronal d'un état mental). Mais comment les résultats de Libet permettraient-ils de défendre une telle conclusion ?

Pour (Haggard 2005, 291), cité plus haut, le problème est que « l'intention consciente se produit *après* le début de l'activité cérébrale préparatoire[,] il est donc impossible qu'elle soit la cause de nos actions, puisqu'une cause ne peut pas avoir lieu après son effet ». Mais ce court argument est clairement bancal : certes, l'intention consciente (W) se produit après *l'activité cérébrale préparatoire* (le RP) et ne peut donc en être la cause. Mais la question est de savoir si elle peut être la cause *de l'activité musculaire* et non pas du RP, or elle est bel et bien *antérieure* au début de l'activité musculaire. Pourquoi donc ne pourrait-elle en être la cause ? Plusieurs auteurs, notamment (Shariff, Schooler, and Vohs 2008), ont exprimé l'intuition suivante : l'intention consciente ne peut pas être la cause de l'activité musculaire parce que *c'est le RP* qui en est la cause (puisque'il arrive avant) : « sans la priorité temporelle, il est très difficile d'attribuer la causalité à l'impulsion volontaire » (Shariff, Schooler, and Vohs 2008, 186). Cette intuition suppose que le RP soit une cause *déterminante* de l'activité musculaire et considère que, si l'activité musculaire est « déjà » entièrement déterminée (à -550 ms) par le RP, alors l'intention consciente W (à -200 ms) n'a plus aucune contribution causale à « ajouter ». Ce raisonnement est, de nouveau, clairement fautif. Tout d'abord, comme on va le voir, l'expérience de Libet ne permet absolument pas d'établir que le RP soit une cause *déterminante* du mouvement musculaire. Mais même en admettant qu'il soit une cause déterminante, cela n'empêcherait absolument pas qu'il soit une cause déterminante lointaine, qui cause l'événement M (le mouvement) *en causant* un autre événement intermédiaire W qui serait à son tour la cause prochaine de M. Pour reprendre un exemple de (Mele 2009, 71), « quand l'allumage d'une mèche précède la combustion de la mèche, qui à son tour précède l'explosion d'un pétard, nous ne concluons pas que la combustion de la mèche n'a aucun rôle causal pour produire l'explosion ». Le fait d'être précédé par une cause déterminante n'empêche donc pas l'intention consciente W de faire partie à part entière de la « chaîne causale » qui mène à l'exécution du mouvement M. Wegner a d'ailleurs aperçu cette possibilité : « La position de la volonté consciente dans la série temporelle suggère peut-être que l'expérience de la volonté est un maillon dans la chaîne causale qui mène à l'action, mais en fait elle pourrait n'être même pas cela. Elle pourrait n'être qu'une effet collatéral – une de ces choses qui, comme l'action, sont causées par des événements cérébraux et mentaux antérieurs » (Wegner 2002, 55). Au regard des résultats de Libet, l'hypothèse que la volonté

mentale soit un pur « effet collatéral » plutôt qu'un « maillon dans la chaîne causale » est clairement gratuite.

On ne peut donc absolument pas dire que les résultats de Libet apportent la preuve (ni même la moindre indication) en faveur d'un court-circuitage de la volonté consciente. Même si l'on admettait que le RP est une cause *déterminante* (de l'action) qui précède la volonté consciente, un auteur compatibiliste pourrait tout à fait continuer de défendre l'existence de la liberté en intégrant ce résultat empirique : il lui suffirait de dire que la volonté consciente W est une étape fondamentale et causalement efficace dans la chaîne causale qui mène (de manière déterministe) du RP à l'action. On n'a donc pas ici une réfutation du libre arbitre « a-compatibiliste ». Peut-être a-t-on tout de même une réfutation du libre arbitre *incompatibiliste*, ce qui serait problématique tout de même compte tenu des résultats du chapitre 4 ? Pour plusieurs auteurs, notamment (Dennett 2004, 222), les conclusions dramatiques des expériences de Libet viennent entièrement d'un présupposé incompatibiliste : les expériences de Libet montreraient en effet que le libertarianisme est faux, mais cette information ne deviendrait une révélation catastrophique que pour les quelques neuroscientifiques qui ont (indûment) présupposé une conception libertarienne de la liberté. En réalité, nous allons voir que les résultats de Libet ne permettent même pas de démontrer quoi que ce soit contre le libertarianisme. Pour voir cela, il faudra savoir si les résultats de Libet font du RP une cause *déterminante* de l'action. C'est ce que nous allons voir à présent en interrogeant la nature de cet événement RP.

La première conclusion que nous avons mentionnée plus haut consistait à dire que le RP observé sur l'électro-encéphalogramme représentait le moment de la décision ou de l'intention, et que par conséquent l'événement conscient W était postérieur à la véritable prise de décision. A première vue, dire que l'action est « déjà décidée » lorsqu'on arrive au moment W semble vouloir dire que l'action est déjà *déterminée* à ce moment ; c'est pourquoi l'identification du RP comme décision ou intention (non conscientes) est particulièrement problématique pour une conception libertarienne du libre arbitre. Mais un examen attentif des résultats de Libet révèle que le RP a bien peu de chances d'être une décision ou une intention.

La première remarque importante est de comprendre pourquoi on a aussi naturellement identifié le RP avec le moment de la décision ou de l'intention. Ce que montrent vraiment les résultats de Libet (si l'on accepte la pertinence de son dispositif expérimental), c'est que l'intention consciente est précédée par « une certaine activité cérébrale ». Mais pourquoi cela devrait-il sembler surprenant. Comme le fait remarquer (Nahmias 2010, 352), « si l'on suppose que les processus conscients sont corrélés avec des processus neuronaux, on devrait s'attendre à ce que les expériences conscientes ne naissent pas de nulle part et n'importe quand ». Loin d'être un résultat étonnant, le fait que notre intention consciente soit précédée d'une certaine activité cérébrale est bien plutôt ce à quoi on aurait dû s'attendre, quelle que soit notre conception du libre arbitre. Mais alors pourquoi a-t-on identifié cette activité cérébrale antérieure comme une « pré-décision » ? Il me semble que la raison tient à l'ordre historique des découvertes rappelé au début de cette section. Initialement, ce que cherchaient les neuroscientifiques était le corrélat neuronal de la volonté consciente et ils s'attendaient à trouver ce corrélat sous forme d'une activité neuronale précédant le mouvement lui-même. Comme on l'a rappelé, la communauté scientifique crut avoir trouvé son Graal en observant le RP. Il fut tout d'abord admis qu'il s'agissait là du corrélat neuronal de l'intention consciente. Puis Libet eut l'idée de tester expérimentalement la présence effective de la conscience et s'aperçut qu'elle apparaissait *après* le (début du) RP. De « décision consciente », le RP devint donc une « décision inconsciente ». Mais imaginez que les choses aient été découvertes dans un ordre différent : imaginez qu'on ait d'abord su localiser le moment de l'intention consciente (à -200 ms) et qu'on ait *ensuite* observé une activité cérébrale à -550 ms. Il me

semble qu'on aurait simplement tiré la conclusion suivante : « nous n'avons pas (pas encore) réussi à trouver le corrélat neuronal de l'intention consciente – le corrélat de ce qui se passe à -200 ms – mais nous avons observé en revanche une certaine activité cérébrale préparatoire à -550 ms ».

A quoi correspond cette activité cérébrale préparatoire qu'est le RP, cette activité qui précède l'intention consciente d'environ 250 ms ? Si l'on accepte une certaine précision conceptuelle, il y a ici un grand nombre de possibilités – un grand nombre d'alternatives à la solution retenue par Libet *et al.*, à savoir qu'il s'agirait d'une pré-décision ou d'une pré-intention inconsciente de plier le doigt. (Mele 2009, 56) a regroupé ces différentes alternatives dans ce qu'il appelle le « groupe pré-intention proximale » (*preproximal intention group*, PPG), qui inclut notamment : « des événements cérébraux adaptés pour être des contributeurs causaux relativement proches à [un désir de plier le doigt], des préparations motrices, ou de l'imagerie motrice » (*ibid.*). Autrement dit, il y a toutes sortes de préparatifs cérébraux qui pourraient être des conditions préparatoires, nécessaires mais non suffisantes, à une prise de décision postérieure.

Avons-nous des raisons positives de penser que le RP correspond plutôt à une condition nécessaire préparatoire (à un élément du PPG de Mele) qu'à une pré-intention inconsciente ? La réponse est oui... et plusieurs de ces raisons sont apportées par Libet lui-même dans ses divers dispositifs expérimentaux. Pour résumer, on peut dire qu'il y a deux types de raisons. Le premier type est un ensemble de raisons indépendantes qui suggèrent que l'événement correspondant au RP peut avoir lieu *en l'absence* de toute intention ou décision, et qu'il ne peut donc pas *être* une intention ou une décision. Le deuxième type de raison consiste à dire qu'il ne peut pas être une intention parce qu'une intention devrait se situer beaucoup plus proche dans le temps de l'action visée (à cause de données indépendantes sur les temps de réaction).

Dans le premier type de raisons, on peut mentionner tout d'abord les données de Libet lui-même sur ce qu'il appelle les « cas de veto ». Dans certaines expériences, Libet a demandé aux sujets de se préparer à plier le doigt à un moment fixé, et à la dernière seconde de « s'opposer » ou de « mettre un veto » à ce mouvement. Dans ce dispositif, on observe un potentiel de préparation, légèrement différent du RP des expériences initiales, mais qui correspond tout à fait au RP d'autres expériences de flexion du doigt, en particulier les expériences où les sujets doivent plier le doigt à un moment annoncé à l'avance (et où ils plient effectivement le doigt), ou les expériences où les sujets plient le doigt quand ils veulent mais en « préparant » leur mouvement. C'est ce que Libet *et al.* ont appelé un RP de type I, par opposition au RP de type II observé dans le cas des flexions du doigt absolument spontanées et sans préparation. Si donc on considère que les RP de type I correspondent au même phénomène dans les différents dispositifs où ils sont observés (cas du veto, cas de la flexion à un moment annoncé, cas de la flexion préparée), alors il est clair que ce RP ne correspond *pas* à une intention ni à une décision de plier le doigt, puisqu'il n'y a aucune intention de plier le doigt dans le cas du veto : un sujet dans le cas du veto sait bien, à l'avance, qu'il ne devra pas plier le doigt, et s'il veut suivre les instructions, il n'a donc aucune *intention* de plier le doigt. A quoi correspond alors le RP de type I ? Il semble qu'il doive correspondre à quelque chose qui est *en commun* dans les trois dispositifs, soit une intention de *préparer* une flexion du doigt, soit des préparatifs cérébraux nécessaires à une éventuelle future flexion du doigt, soit de l'imagerie motrice permettant éventuellement de plier le doigt. Corrélativement, un certain nombre d'études en neurosciences ont permis d'établir que « le schéma d'activité cérébrale associé au fait d'imaginer l'exécution d'un mouvement est très similaire au schéma d'activité associé au fait de préparer l'exécution d'un mouvement » (Spence and Frith 1999, 27, cité par Mele 2009, 55). Par conséquent, il ne serait pas du tout absurde de penser que ce RP de type I correspond à l'activité d'imagerie cérébrale

qui correspond à la fois à l'imagination de l'action et, en l'occurrence, à la préparation d'un (éventuel) mouvement du doigt. Si on a des raisons indépendantes de penser que le RP de type I correspond seulement à ce genre de préparatifs d'un éventuel mouvement, pourquoi irait-on croire que le RP de type II, quant à lui, correspond à une véritable *intention* d'effectuer l'action, plutôt qu'à ce genre de préparatifs ? En particulier, sachant que le RP de type I n'est pas toujours suivi d'un mouvement effectif (à cause des cas de veto), pourquoi irait-on penser que le RP de type II, quant à lui, est toujours (ou typiquement) suivi d'un mouvement effectif ?

La première chose à noter est que les résultats de Libet ne permettent absolument pas de répondre à cette question. Dans l'expérience de Libet, il n'y a d'enregistrement de l'activité cérébrale *que* pour les intervalles de temps qui précèdent immédiatement un mouvement effectif. Pour toutes les phases sans mouvement, aucun enregistrement n'est gardé. Par conséquent, « il est possible que des RP robustes apparaissent fréquemment durant l'intervalle critique, et qu'il n'y ait de réponse [motrice] que lorsqu'une décision consciente est rapportée. Dans ce cas, le processus inconscient ne serait clairement pas une cause suffisante de l'action. Nous ne détectons pas les RP inefficaces parce que l'enregistrement des RP se fait par moyenne rétroactive [*back-averaging*] à partir de la réponse. Les RP qui ne débouchent pas sur une réponse ne seraient donc pas découverts. » (Banks 2006, 239) La méthode de moyenne rétroactive entraîne donc un « biais d'échantillonnage » massif.³⁹⁹ Les résultats de Libet laissent totalement ouverte la possibilité que les RP de type II soient de simples préparatifs causaux, non déterminants, et que le facteur causalement déterminant soit l'intention consciente elle-même. Non seulement cette possibilité est laissée ouverte, mais Libet apporte lui-même des raisons de penser que c'est effectivement ce qui se produit. En effet, Libet souligne qu'un certain nombre de sujets dans l'expérience initiale (l'expérience sans veto et sans préparation du mouvement du doigt) « ont déclaré qu'ils se rappelaient avoir ressenti, au cours de certains tests, un désir (*urge*) conscient d'agir, mais que ce désir avait été 'avorté' ou pour ainsi dire supprimé avant qu'un mouvement effectif n'ait lieu ; dans ce genre de cas, le sujet attendait simplement l'apparition d'un nouveau désir qui, une fois satisfait, constituait l'événement effectif dont le RP était enregistré » (Libet 1985, 538). Ces déclarations des sujets suggèrent donc à Libet les conclusions suivantes : dans le cas des mouvements non préparés comme dans le cas des mouvements préparés, il semble exister une possibilité de « bloquer » un désir ou une impulsion par un « veto » ; puisque dans le cas de mouvement préparé, ce désir contrarié est représenté par un RP (de type I) non suivi de mouvement musculaire, il y a toutes les raisons de penser que dans le cas des mouvements spontanés aussi on aurait pu en principe observer des RP (de type II) non suivis de mouvement musculaire, correspondant aux désirs bloqués par veto ; enfin, ceci suggère que même dans les cas où le RP a été suivi d'un mouvement musculaire, ce RP ne *déterminait* pas ce mouvement musculaire, c'est-à-dire que ce RP enregistré *aurait pu* ne pas être suivi d'un mouvement musculaire effectif. C'est pourquoi Libet considère finalement que la conscience garde toujours un « pouvoir de veto » après l'apparition du RP. Mais si c'est l'intention consciente (W) qui *détermine* réellement l'occurrence ou non du mouvement, alors il est illégitime de dire que cela est déjà « décidé » par le cerveau sous la forme du RP. C'est-à-dire que, si l'on admet comme Libet que le RP ne détermine pas le choix, il est illégitime de continuer à considérer ce RP comme une « pré-décision » ou une « pré-intention », comme le fait Libet. La décision a lieu après le RP, et ce dernier n'en est, au mieux, qu'un contributeur causal, une cause nécessaire mais non suffisante.

Les résultats de Libet ne permettent donc pas d'établir que le RP serait une « pré-décision » ou une « pré-intention », et corrélativement, ils ne permettent pas de réfuter la

³⁹⁹ Cette objection est formulée par (Mele 2009, 72 sqq), (Roskies 2010, 15–16), ou encore (Bayne 2011, 40).

théorie libertarienne de la décision ou du choix, puisqu'ils sont compatibles avec l'idée que l'issue reste causalement indéterminée jusqu'au moment de la décision consciente. Si on reprend le modèle libertarien de Robert Kane, on pourrait par exemple soutenir que le RP observé par Libet correspond à une motivation de l'agent en faveur d'une certaine action (plier le doigt), ou à un certain effort de la volonté pour prendre la décision de plier le doigt ; on peut également soutenir que cet effort de la volonté *cause* la décision consciente ultérieure de plier le doigt. L'important ici, c'est que cet effort de la volonté peut tout à fait ne causer la décision que de manière indéterministe. Par conséquent, nous n'avons là ni une réfutation de l'influence causale de la volonté consciente, ni une réfutation de l'indéterminisme de la décision consciente. Peut-être y a-t-il cependant une réfutation de *certain*s modèles libertariens ? En particulier des modèles de la causalité agentive ? On pourrait soutenir que, si le RP est *cause* antécédente de la décision consciente (W) puis du mouvement, alors on ne peut être que dans un modèle de causalité événementielle, même s'il peut s'agir d'une causalité événementielle indéterministe. Mais en réalité rien ne prouve non plus que le RP soit une *cause* à proprement parler des événements qui s'ensuivent (la décision consciente W puis le mouvement). On pourrait très bien considérer que le RP est simplement une *condition de possibilité*, un état nécessaire pour que *l'agent* puisse causer la décision, et non pas un événement qui vient causer lui-même la décision.

Pour résumer, les résultats de Libet, si on accepte l'interprétation qu'en fait Libet lui-même, ne s'opposent pas en réalité au libre arbitre, même pas à une conception libertarienne du libre arbitre, ni dans la version de la causalité événementielle, ni dans celle de la causalité agentive. Tout ce que montrent les résultats de Libet, c'est que la décision consciente a des conditions nécessaires antécédentes dans le cerveau, et que nous pouvons observer ces conditions nécessaires antécédentes. Mais cette décision consciente pourrait très bien être (et tout semble indiquer qu'elle *est* effectivement) ce qu'elle semble être pour le sens commun : le facteur causal déterminant responsable de l'action effective.

Il convient de mentionner pour finir une autre étude, proche de celles de Libet, qui semble plus problématique parce qu'elle révèle une forme de prédictibilité de la décision consciente, et donc une forme de déterminisme de cette décision. Encore une fois, un tel résultat ne serait pas forcément problématique dans une conception compatibiliste de la décision libre, mais dans la perspective libertarienne qui est la nôtre, il est important de rejeter également ce résultat apparent.

Dans l'étude menée par Soon *et al.* (2008), on demande aux sujets de presser un bouton soit avec leur index droit soit avec leur index gauche « dès qu'ils ressentent un désir (*urge*) de le faire », et de prêter attention au moment où « la décision motrice a été consciemment prise ». Les déclarations des sujets dans cette expérience font apparaître un moment de la décision consciente (W) qui correspond aux résultats de Libet (moins d'une demi-seconde). Mais un IRM fait apparaître une activité cérébrale très précoce, de 7 à 10 secondes avant le mouvement (donc bien avant le RP), et ce qui est plus important, c'est que l'observation de cette activité cérébrale antécédente permet de *prédire* quel bouton le sujet va presser ! Il semble donc que, cette fois, l'impression de *décider* ou de déterminer (au temps W) quel bouton sera pressé n'est qu'une illusion, puisque cela était déjà prédictible et déterminé 7 à 10 secondes auparavant.⁴⁰⁰

Comme le fait remarquer (Nahmias 2010, 349), une plus grande attention aux résultats de Soon permet d'éviter une conclusion trop radicale. En effet, les « prédictions » dont il s'agit ici (qui n'étaient en fait que des corrélations *a posteriori*) n'étaient pas absolument certaines, mais ne pouvaient être faites qu'avec une précision de 60% – soit 10 points de

⁴⁰⁰ L'activité cérébrale permettait aussi, dans ce dispositif, de prédire *quand* le sujet allait presser le bouton, environ 5 secondes avant l'activité musculaire.

mieux que le pur hasard. Si ce résultat montre quelque chose, ce n'est certainement pas l'existence d'un *déterminisme* en faveur de tel ou tel choix, mais plutôt l'existence de facteurs augmentant la *probabilité* de tel choix plutôt que tel autre. Par exemple, au moment où je vais prendre ma décision pour savoir sur quel bouton appuyer, il semble (compte tenu des résultats de Soon) que j'aie une probabilité objective de 60% de choisir tel bouton particulier, par exemple le bouton droit, et que cette probabilité objective soit fondée sur un certain facteur cérébral qui me donne cette *propension* à choisir le bouton droit.

Ce résultat permet-il de dire que notre expérience de la décision consciente est illusoire ? Certainement pas. Notre expérience de la décision libre suppose seulement que notre choix soit non déterminé et qu'il soit contrôlé. On a vu au chapitre 5 (p.367) que l'existence d'une probabilité objective (inférieure à 1) n'empêchait pas le choix d'être *contrôlé*. Et il est clair que l'existence d'une probabilité objective (supérieure à 0,5 mais inférieure à 1) ne donne pas au choix un caractère *déterministe*. Par conséquent, le fait qu'il existe des facteurs causaux antécédents fixant la probabilité objective de tel ou tel choix à 60% n'empêche pas que ce choix soit libre dans le sens libertarien. Dans le modèle de la causalité événementielle, on peut très bien imaginer que deux raisons concurrentes aient des « poids » différents, de telle sorte qu'il y ait, par exemple, une probabilité de 60% que la raison A cause mon action et une probabilité de 40% que la raison B cause une autre action. Dans le modèle de la causalité agentive, les raisons contraires ne jouent pas le rôle de causes, mais comme l'a montré (O'Connor 2000, sec. 5.4), on peut les considérer comme des « conditions nécessaires structurantes », c'est-à-dire des conditions qui structurent la capacité causale de l'agent (la causalité agentive) en lui conférant des propensions ou probabilités déterminées d'agir de telle ou telle manière.

Pour conclure, il n'existe pas aujourd'hui de données neuroscientifiques qui viennent discréditer notre expérience agentive. Plusieurs auteurs ont pensé que, compte tenu des expériences de Libet, l'expérience agentive consciente arrivait trop tard pour être véritablement une conscience de la décision, causalement efficace dans le processus de l'action volontaire, et qu'elle n'était par conséquent qu'un épiphénomène tardif et illusoire. Mais on a vu qu'en réalité les résultats de Libet ne montraient absolument pas cela : tout ce qu'ils montrent, c'est que la décision consciente est précédée de phénomènes préparatoires, observables dans le cerveau, qui constituent des conditions nécessaires de la décision consciente. Mais l'expérience consciente peut tout à fait être le moment véritable de la décision, c'est-à-dire le moment où est fixée l'issue du choix, jusqu'alors indéterminée.

I.2. Wegner : les erreurs de la conscience d'agir

L'intention générale de Wegner dans *The Illusion of Conscious Will* (2002) est de discréditer l'expérience de l'action en montrant qu'il y a une *dissociation*, ou une *décorrélacion*, entre l'expérience (ou la conscience) d'agir et le fait que nous agissons. Pour cela, il accumule un grand nombre de données expérimentales, et s'appuie notamment sur les données neuroscientifiques de Libet, qui peuvent suggérer, comme on l'a vu, une *décorrélacion temporelle* entre expérience et action – c'est-à-dire un retard de l'expérience d'agir par rapport au moment supposé de la décision véritable. Les autres données expérimentales sur lesquelles s'appuie Wegner sont des expériences de psychologie qui suggèrent une *décorrélacion* plus radicale encore entre expérience et action – non pas seulement un retard de l'une sur l'autre, mais une absence complète de l'expérience d'agir dans certains cas où nous agissons, ou inversement une présence de l'expérience d'action lorsqu'en fait nous n'agissons pas. Ces données cumulatives sont censées suggérer une

indépendance complète entre action et expérience de l'action, de sorte qu'on ne pourrait plus faire confiance du tout à notre expérience de l'action, même pour son contenu le plus basique et anodin (à savoir que « nous agissons »). Or si nous ne pouvons plus nous fier à notre expérience de l'action pour ce contenu anodin, il est clair que nous ne pourrions pas nous y fier, *a fortiori*, pour un contenu plus fin et plus engageant métaphysiquement (à savoir que « nous agissons *de manière indéterministe* »).

Si l'intention générale de Wegner est assez claire, son argument précis peut donner lieu à différentes interprétations, comme l'a montré Tim Bayne (2006, 177–178). Nous allons voir les deux grandes interprétations de l'argument de Wegner. Indépendamment de ses arguments à proprement parler, il est clair que Wegner tire une grande force rhétorique du caractère cumulatif de sa liste « d'erreurs » de la conscience d'agir. Il faut pourtant noter qu'il ne s'agit là que d'une force rhétorique : par le même genre de procédé, on pourrait écrire un livre qui accumule les exemples où nous avons des expériences perceptives erronées (les innombrables formes d'illusions ou d'hallucination), ou inversement les exemples où nous traitons les informations de notre environnement sans expérience perceptive (les cas de « vision aveugle » ou *blind sight*), et l'on pourrait choisir d'intituler un tel livre *The illusion of Perceptive Experience*, jetant ainsi le discrédit sur toute forme de perception. Si longue que soit la liste des illusions de la perception, cela ne fait pas de la perception *en général* une illusion. De même, une longue liste d'illusions de l'expérience d'agir peut tendre, rhétoriquement, à affaiblir notre confiance dans cette expérience, et pourtant cela ne fait pas de cette expérience en général une illusion. Avant de voir les deux versions de l'argument de Wegner, je rappellerai brièvement les expériences principales sur lesquelles il s'appuie.

Les expériences que mentionne Wegner sont de deux types : les unes sont des cas d'expérience d'agir sans action, les autres des cas d'action sans expérience d'agir.

Le premier cas de figure est représenté essentiellement par une expérience menée par Wegner lui-même, l'expérience « I spy » (Wegner and Wheatley 1999, 487–489; Wegner 2002, 74–78). L'expérience est censée montrer que pour avoir l'impression d'agir (ou l'impression d'avoir agi), il suffit d'avoir *pensé* à l'action réalisée peu de temps avant sa réalisation. Le dispositif élaboré pour montrer cela est le suivant : le sujet est invité à regarder sur un ordinateur une photo contenant un grand nombre de petits objets (dinosaur, cygne, voiture, etc.) et à promener un curseur sur cette image au moyen d'une souris surmontée d'une petite plaque carrée. Un autre sujet, qui est en fait un « complice » de l'expérimentateur, pose également ses doigts sur la même plaque, et contribue donc à mouvoir le curseur. L'expérimentateur invite les deux sujets à arrêter la souris pour sélectionner un objet lorsqu'ils le souhaitent. Les deux sujets sont équipés d'un casque audio, qui est censé émettre la même chose pour les deux sujets, à savoir de la musique parfois accompagnée de mots, qui correspondent ou non à des objets représentés à l'écran. En réalité, le complice entend dans son casque des instructions précises, et notamment, il reçoit parfois l'instruction de « forcer » un arrêt du curseur sur un objet précis. Lorsque cela se produit, le sujet entend dans son casque le nom de cet objet peu de temps avant le moment fixé de l'arrêt forcé par le complice. Dans ce genre de cas, on constate que le sujet rapporte une impression d'avoir lui-même provoqué l'arrêt, simplement parce qu'il a *entendu* le nom de l'objet peu de temps avant l'arrêt, alors qu'en réalité l'arrêt a seulement été forcé par le complice.⁴⁰¹ On a donc bien là, semble-t-il, le cas d'une impression d'agir sans qu'il y ait action de la part du sujet.

⁴⁰¹ On sait, grâce à une expérience témoin, que le simple fait d'entendre un nom d'objet dans les écouteurs n'entraîne pas, typiquement les sujets à faire un arrêt sur l'objet mentionné, lorsqu'il n'y a pas de complice pour forcer l'arrêt. Dans l'expérience où il y a sujet et complice, on a donc toutes les raisons de penser que si le complice n'avait pas forcé l'arrêt, l'arrêt n'aurait pas eu lieu, et non pas que le sujet et le complice ont conjointement contribué à l'arrêt.

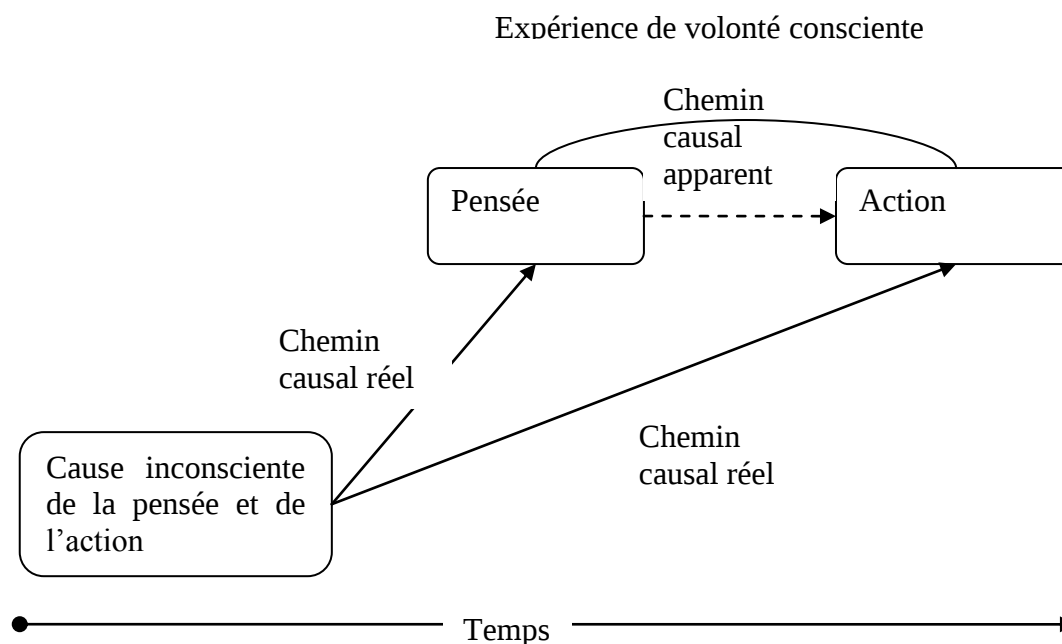
Le deuxième cas de figure est celui d'une action menée par le sujet sans qu'il y ait conscience d'agir. Wegner mentionne trois situations principales qui correspondent à ce cas de figure : les actions idéomotrices enregistrées par automatographe, la communication facilitée, et le comportement d'utilisation d'objets. Un automatographe (Wegner 2002, 123–124) est une planche montée sur trois boules, qui permet d'enregistrer avec précision les mouvements d'une main posée sur la planche. Avec ce dispositif, on a pu observer qu'un sujet normal avait tendance à effectuer inconsciemment certains mouvements de la main, en fonction de mouvements ou d'événements perçus dans son environnement : par exemple, un sujet qui devait compter les cliques d'un métronome faisait de petites allées et venues avec sa main ; un sujet à qui on demandait de penser à un objet se trouvant à sa gauche dans la pièce bougeait légèrement sa main vers la gauche ; un sujet qui regardait un objet en mouvement circulaire reproduisait un léger mouvement circulaire de la main, etc. La communication facilitée (Wegner 2002, 195–201) est une technique mise en place, à l'origine, pour permettre à des personnes souffrant d'autisme ou d'autres incapacités à parler, un moyen de communiquer en indiquant des lettres sur un clavier d'ordinateur. Cette technique fait appel à un « facilitateur » qui tient les mains du patient handicapé mais est simplement censé « aider » leur mouvement, et non pas le diriger. Cette technique a eu des résultats spectaculaires (des personnes gravement handicapées relationnellement qui devenaient capables d'écrire des poèmes) qui ont attiré beaucoup de suspicions contre les facilitateurs et leur prétention à ne pas diriger le mouvement des mains des patients. Tout porte à croire, pourtant, que nombre de facilitateurs étaient de bonne foi, et qu'ils contrôlaient ce qui s'écrivait sur l'ordinateur sans avoir conscience de le contrôler. Enfin, le comportement d'utilisation d'objets (Wegner 2002, 122) est un comportement pathologique dans lequel un patient atteint de lésion au lobe frontal accomplit de manière automatique, et sans conscience d'agir, des actions qui lui sont « suggérées » par les objets se trouvant à sa portée immédiate : par exemple, en touchant la main du patient avec un verre et une carafe d'eau, le patient prendra automatiquement la carafe pour remplir le verre ; touché par un paquet de cigarettes et un briquet, il allumera la cigarette ; « un patient à qui l'on tendait trois paires de lunettes les chaussa toutes les trois à la suite et se retrouva avec trois paires de lunettes sur le nez ». Dans chacun de ces trois cas – automatographe, communication facilitée, comportement d'utilisation – il semble donc qu'on soit en présence d'une action qui n'est pas accompagnée d'une expérience ou d'une conscience d'agir.

Tous ces cas de « décorrélations » entre action et expérience d'agir sont sans doute troublants. Mais pour savoir s'ils décrédibilisent (s'ils réfutent fondationnellement) notre expérience d'agir, encore faut-il voir précisément les arguments qui sont censés mener à une conclusion si radicale. Tim Bayne (2006, 177–178) a le mérite d'avoir montré que l'argument de Wegner pouvait être interprété de deux manières très différentes : un argument classique d'erreur et un argument d'inférence à la meilleure explication.

L'argument d'erreur s'appuie essentiellement sur les cas d'expérience agentive sans action (du type *I-spy*) et manifeste une méfiance toute cartésienne : puisque l'expérience d'agir nous a trompés une fois, nous ne devons pas (épistémologiquement) lui faire confiance. Mais de nombreux philosophes ont souligné, contre Wegner, qu'un tel argument revient à confondre fiabilité et infaillibilité : « Le fait que les expériences de l'action *puissent* être non-véridiques montre que les mécanismes responsables de l'apparition de ces expériences sont faillibles, mais cela ne montre pas qu'ils ne sont pas fiables » (Bayne 2006, 178).⁴⁰² On a vu au chapitre 2 (p.135) qu'une contrainte infaillibiliste en épistémologie était beaucoup trop exigeante et menait au scepticisme radical dans le cas de la perception ; ce qui est important,

⁴⁰² Voir aussi (Pacherie 2006, 163) et (Mele 2008, 121–122).

c'est que nos sources épistémiques soient *fiables*, et l'existence d'erreurs localisées n'empêche pas une source épistémique d'être fiable. Qu'en est-il de l'expérience d'agir ? Est-elle erronée de manière seulement exceptionnelle, dans des dispositifs complexes et rares comme celui de l'étude *I-spy*, ou bien est-elle erronée de manière fréquente voire systématique ? Il est particulièrement intéressant de voir la réponse de Wegner lui-même à cette question : « Même si la volonté consciente n'est pas un signe infaillible de notre influence causale, c'en est un assez bon signe » (Wegner 2002, 333). Le modèle proposé par Wegner est en effet le suivant : dans la majorité des cas, la cause inconsciente de l'action (qui se trouve dans notre cerveau) cause *à la fois* une représentation consciente de l'action et, peu de temps après, l'action elle-même.



Modèle de la causalité mentale apparente (voir Wegner 2002, 68)

Autrement dit, la représentation consciente de l'action est un effet corrélatif qui ne cause pas l'action, mais qui est tout de même typiquement associé à une véritable action. Parfois (dans les cas de *I-spy*) la représentation consciente de l'action est produite artificiellement par un autre mécanisme que celui de la cause cérébrale inconsciente de l'action, et dans ce cas, il y a expérience de l'action sans aucune action initiée au niveau cérébral. Mais le plus souvent, lorsqu'on a une représentation consciente juste avant l'action, c'est un bon signe qu'il y a eu une cause inconsciente commune de cette représentation et de l'action. Bien que ce modèle garde quelque chose « d'illusoire » (le chemin causal de la pensée à l'action n'est qu'apparent), il est important de noter que l'expérience d'agir y reste un « signe fiable » du fait que nous agissons. Autrement dit, l'argument de l'erreur ne peut pas fonctionner aux yeux même de Wegner : l'expérience d'agir, lorsqu'elle est présente, est une expérience certes faillible, mais tout à fait fiable quant au fait que nous agissons.⁴⁰³ Et le

⁴⁰³ Une difficulté pour interpréter Wegner vient du fait qu'il semble parfois dire que même le contenu « j'agis » est illusoire dans le cas normal : « It seems to each of us that we have conscious will. It seems we have selves. It seems we have minds. *It seems we are agents. It seems we cause what we do.* Although it is sobering and ultimately accurate to call all this an illusion » (Wegner 2002, 342 italiques ajoutées). Si l'on suit ces déclarations alarmistes au pied de la lettre, il semble que nous n'agissions jamais ; qu'il n'y ait en fait aucune différence entre les cas *I-spy* où nous avons l'illusion d'agir et ceux où nous agissons en effet. Si tel était le cas, il n'y aurait même pas lieu de parler de la deuxième catégorie de cas, ceux dans lesquels *nous agissons* sans avoir

fait de mettre en avant les cas exceptionnels où l'on peut la mettre en défaut n'a finalement aucune conséquence argumentative. Pour mieux comprendre le cœur de l'argument de Wegner, il faut s'intéresser à ce qui demeure « illusoire » ou problématique dans le modèle épiphénoménal que je viens de présenter.

La conclusion problématique à laquelle Wegner veut arriver, ce n'est pas que notre expérience d'agir serait non-fiable (pour le contenu « j'agis »), mais plutôt qu'elle est « épiphénoménale » ou sans influence causale dans le processus de l'action. Et il entend parvenir à cette conclusion au moyen d'une inférence à la meilleure explication, pour laquelle les cas importants sont les cas d'action sans expérience d'agir (et non plus les cas d'expérience d'agir sans action). En résumé, l'argument est le suivant : puisqu'il existe des situations (automatographe, communication facilitée, comportement d'utilisation) dans lesquelles l'action se produit sans expérience consciente, cela montre qu'il existe un processus causal inconscient capable à lui seul de causer l'action ; mais s'il existe un tel processus, il est plus simple ou plus économique de penser que c'est ce processus qui, d'une manière générale, cause nos actions, même lorsque celles-ci sont accompagnées d'une expérience d'agir. C'est-à-dire que l'expérience d'agir, même lorsqu'elle est présente, n'aurait aucune influence causale. En quoi cette conclusion implique-t-elle un élément d'illusion dans notre expérience d'agir ? C'est que notre expérience d'agir n'a pas seulement le contenu « j'agis », c'est-à-dire le contenu « c'est moi (ou mon cerveau), plutôt que quelqu'un d'extérieur, qui suis à l'origine causale de tel mouvement ». Pour ce contenu, notre expérience agentive, on l'a vu, est une indication fiable (un « assez bon signe ») quoique marginalement faillible. Mais notre expérience a aussi le contenu sui-référentiel selon lequel « cette expérience consciente est à l'origine causale de l'action » ; et pour ce contenu plus riche, notre expérience est *toujours* et absolument erronée (et pas seulement « peu fiable »), si du moins la conclusion épiphénoménale de Wegner est légitime.

La question est donc de savoir si l'argument de l'inférence à la meilleure explication permet d'établir le caractère « épiphénoménal » (ou causalement inefficace) de l'expérience d'agir. Si la volonté consciente est causalement inefficace, alors notre expérience agentive sera toujours erronée sur un aspect fondamental de son contenu, et donc on ne devra pas s'y fier pour les autres aspects de son contenu (notamment les aspects incompatibilistes). Mais là encore, un grand nombre de philosophes ont fait remarquer que l'argument reposait sur une généralisation pessimiste et douteuse : « Le fait qu'il y ait ces exceptions [dans lesquelles aucune volonté consciente ne correspond à l'action produite] ne montre pas que, dans les cas *normaux* de correspondance, la volonté consciente est causalement non-pertinente » (Nahmias 2002, 533). Pour pouvoir généraliser ce qui se passe dans les cas d'action automatique aux cas d'actions volontaires (ou vécues comme volontaires), il faudrait que nous ayons une raison indépendante de supposer une *homogénéité* entre ces deux situations. Mais nous n'avons pas de raison de ce type : les cas d'actions automatiques ont toutes les chances d'avoir un fonctionnement causal différent des actions volontaires normales, en dépit de ce que suggère Wegner :

Soit les automatismes sont des étrangetés sur la toile de fond de la causalité par comportement conscient dans la vie quotidienne, soit nous devons tout inverser radicalement et commencer à penser que le comportement qui se produit *avec* une expérience de volonté est d'une certaine manière le cas étrange, un ajout par rapport à

l'expérience d'agir, puisque nous n'agissons jamais ! En réalité, il semble que Wegner, dans un vocabulaire peu précautionneux, veuille dire la chose suivante : il arrive que nous agissions au sens où c'est notre cerveau qui a le contrôle d'un mouvement volontaire donné, mais ce qui n'arrive jamais, c'est que le « moi conscient » (les événements conscients) soit à l'origine causale de l'action.

un système sous-jacent plus fondamental. Depuis longtemps, on a traité les automatismes comme des exceptions [...]. La théorie de la causalité mentale apparente suggère qu'il pourrait être temps d'inverser les rôles.
(Wegner 2002, 144).

La « théorie de la causalité mentale apparente » n'est rien d'autre que *l'hypothèse* épiphénoméniste présentée plus haut. En tant que pure hypothèse, elle ne peut pas « suggérer » l'homogénéité entre actions automatiques et actions (apparemment) volontaires : elle ne fait que *modéliser* une telle éventualité. Le seul argument de Wegner semble être le suivant : si les comportements automatiques sont « le cas normal », alors les comportements avec expérience de volonté doivent être un simple ajout reposant fondamentalement sur le même fonctionnement. Il est bien possible que les comportements automatiques soient, statistiquement, plus fréquents ou plus nombreux que les comportements avec volonté consciente ; certaines études de psychologie semblent en effet révéler une telle répartition statistique.⁴⁰⁴ Mais cette répartition statistique n'implique absolument pas qu'il faille penser une homogénéité entre les deux systèmes.⁴⁰⁵ Wegner nous laisse donc sans aucune raison valable de croire à l'hypothèse (certes intéressante) qu'il construit. En fin de compte, nous en restons donc au point de départ dialectique, c'est-à-dire que nous avons toutes les raisons de penser que les actions automatiques et les actions volontaires ne sont *pas* homogènes dans leur fonctionnement causal. Les actions automatiques restent des exceptions marginales : les comportements d'utilisation sont des comportements *pathologiques* qui indiquent un mauvais fonctionnement – sans doute un problème dans la chaîne causale menant au mouvement. Les mouvements enregistrés par l'automatographe ne sont pas pathologiques, mais ils sont *infimes*, et il n'est pas clair du tout que les mécanismes de l'action automatique et inconsciente pourraient générer des mouvements plus importants chez un sujet sain. Pour toutes ces raisons, on n'a pas de raison de supposer que nos actions volontaires ont un fonctionnement causal comparable à celui des actions automatiques. Et l'on n'a donc pas de raison de rejeter la différence apparente entre les deux, à savoir que les actions volontaires sont causées *par des intentions conscientes* (ou par leurs corrélats neuronaux).

Les cas d'erreur de la conscience d'agir accumulés par Wegner ne parviennent donc pas à discréditer cette conscience d'agir – pas plus qu'un livre de psychologie ou de neurosciences accumulant les illusions de la perception ne serait une bonne raison de devenir sceptique à propos de nos sens. Le fait que nous puissions avoir, marginalement, une expérience d'agir sans agir véritablement montre seulement que cette expérience n'est pas *infaillible*, mais Wegner lui-même reconnaît qu'elle est néanmoins un signe *fiable* de l'action. Par ailleurs, le fait que nous puissions, marginalement, agir sans en avoir conscience, et par simples processus inconscients, ne montre pas que ces processus inconscients sont également à l'œuvre dans les actions volontaires quotidiennes, car il n'y a aucune raison de supposer une homogénéité entre le fonctionnement causal des actions automatiques et celui des actions volontaires normales. Notre expérience d'agir est donc non seulement un indice fiable du fait que nous agissons, mais elle a aussi toutes les chances d'être un indice fiable du fait que nous agissons « par causalité mentale », c'est-à-dire par l'influence causale d'une intention mentale sur le mouvement volontaire.

⁴⁰⁴ Voir notamment (Kahneman 2011), avec sa fameuse distinction, parmi les processus de pensée, entre « système 1 » (automatique et rapide) et « système 2 » (conscient et lent). Il semble que le cerveau favorise le système automatique aussi souvent qu'il est possible, pour gagner en rapidité.

⁴⁰⁵ Chez Kahneman, notamment, le fait que le système 1 soit plus fréquent n'enlève absolument rien à la grande différence fonctionnelle entre les deux systèmes.

Les expériences rapportées par Wegner, pas plus que les expériences de Libet, ne permettent de montrer une erreur de notre expérience agentive dans ses aspects anodins et indépendants de la question de l'incompatibilisme – l'aspect de l'action en tant que telle, et l'aspect de la « causalité mentale ». D'une manière générale, il n'existe pas, à ma connaissance, de données empiriques qui permettraient de dire que notre expérience agentive, indépendamment de la question de l'incompatibilisme, est non fiable.

II. Réfutation par le contenu alourdi : le Problème d'Intégration

Récapitulons la situation dialectique où nous nous trouvons à présent : nous avons une expérience agentive à laquelle nous sommes *prima facie* justifiés à faire confiance, c'est-à-dire que nous sommes *prima facie* justifiés à accepter son contenu et à la tenir pour une source de connaissance. Cette expérience a un contenu riche, qui contient des aspects anodins indépendants de la question du déterminisme et d'autres aspects qui semblent, et qui sont, incompatibles avec le déterminisme. Pour les aspects anodins, tout semble confirmer que notre expérience agentive est *effectivement* fiable – c'est-à-dire que rien ne vient réfuter notre justification *prima facie* à croire qu'il y a là une connaissance. Cela vient sans doute renforcer la confiance *prima facie* que nous mettons dans cette expérience, même pour les aspects plus « lourds », les aspects incompatibilistes. D'autant que nous avons vu (chapitre 5) que ces aspects incompatibilistes constituaient une information tout à fait *recevable* d'un point de vue scientifique. Nous nous trouvons donc dans la situation suivante : un certain « témoin » (notre expérience agentive) nous donne deux (groupes d') informations A et B ; on sait que ce témoin est fiable sur A, et que B est une information totalement recevable. Pourquoi ne croirait-on donc pas ce témoin à propos de B aussi bien qu'à propos de A ? Un dernier problème pourrait être le suivant : il se pourrait que B (à la différence de A) soit une information « trop lourde » pour être connue ou transmise par ce témoin. Dans le cas de l'expérience agentive, il se pourrait que les aspects incompatibilistes soient une information trop lourde métaphysiquement pour pouvoir être *connue* sur la base d'une simple expérience. Ou pour le dire autrement, si l'on croyait à l'indéterminisme causal sur la base de notre expérience du libre arbitre, cette croyance ne pourrait être, au mieux, qu'une croyance vraie « par pur hasard » et non pas une *connaissance*.

Nous arrivons donc enfin au cœur du problème d'Intégration tel que nous l'avons vu à la fin du chapitre 3. Après avoir présenté en première partie l'intuition « mooréenne » :

(IM) notre expérience du libre arbitre apporte une connaissance

puis, au chapitre 4, l'intuition analytique :

(IA) être libre suppose l'indéterminisme causal

nous allons voir à présent l'intuition épistémologique qui fera apparaître une contradiction avec ces deux premières intuitions :

(IE) notre expérience du libre arbitre ne peut pas délivrer une *connaissance* de l'indéterminisme causal.

II.1. L'argument de réfutation fondationnelle et son principe épistémologique

Plusieurs philosophes ont exprimé l'intuition (IE), et ils sont encore plus nombreux à avoir considéré cette intuition comme une évidence sans avoir pris la peine de l'explicitier ou de la défendre.⁴⁰⁶ Fischer, par exemple, écrit ceci :

Un libertarien est obligé, semble-t-il, de prétendre qu'il connaît, dans son fauteuil, que le déterminisme causal est faux ; mais comment pourrions-nous savoir à l'avance une telle thèse empirique ? (Fischer et al. 2007, 81)

Comme on le voit, cette intuition n'est pas spécifiquement une intuition anti-introspection, mais plutôt une intuition contre la connaissance *a priori* (non expérimentale) de l'indéterminisme : le problème n'est pas tant que *cette* expérience (l'expérience du libre arbitre) est trop faible pour fournir une connaissance de l'indéterminisme ; le problème est que *toute source a priori* (qu'il s'agisse d'une expérience commune, ou bien de réflexions *a priori* à partir de la moralité, comme chez Peter van Inwagen), toute source qui n'est pas l'expérimentation scientifique elle-même, sera trop faible pour pouvoir apporter l'information de l'indéterminisme causal. Van Inwagen développe bien ce point lorsqu'il envisage l'objection qu'on pourrait faire à sa théorie de la connaissance du libre arbitre à partir de la moralité :

Si tous vos arguments sont corrects, alors notre connaissance (supposée) de la responsabilité morale, associée à certains arguments *a priori*, peut constituer une bonne raison de croire que le déterminisme est faux. Mais ces choses ne sont pas le *genre* de choses qui peuvent être de bonnes raisons pour croire à l'indéterminisme. L'indéterminisme est, pour le dire franchement, une thèse qui concerne le mouvement des particules de matière dans le vide ; la thèse plus spécifique que *nous* sommes indéterminés est une thèse qui concerne les détails et les fonctionnements microscopiques des systèmes nerveux. Seules des investigations scientifiques sont pertinentes pour la vérité ou la fausseté de telles thèses. (Cette affirmation est vraie même si vous avez raison de dire qu'il est improbable que la science soit jamais capable de montrer que nous sommes indéterminés.) (van Inwagen 1983, 210)

Van Inwagen reconnaît fort honnêtement que « parfois, dans certaines dispositions d'esprit, [il] trouve cet argument très puissant » bien que « [son] opinion réfléchie soit qu'il doit être rejeté ». Autrement dit, van Inwagen partage, à première vue, l'intuition (IE) selon laquelle l'indéterminisme ne peut pas être connu autrement que de manière empirique – et donc ni par une expérience agentive commune, ni par des réflexions de sens commun sur la responsabilité morale ; mais face au trilemme entre cette intuition d'une part et d'autre part son intuition que nous savons que nous sommes libres et responsables (IM) et son intuition incompatibiliste (IA), van Inwagen est plus enclin à rejeter l'intuition épistémologique que les deux autres.⁴⁰⁷

Mark Balaguer en revanche exprime clairement que l'intuition (IE) est pour lui la plus forte :

⁴⁰⁶ Outre Fischer, van Inwagen et Balaguer, cités plus bas, voir également (Kane 1996, 239) : « We cannot know we have free will by introspection alone, or indeed on a priori grounds. [...] Empirical evidence is needed about the hidden causes of our behavior, including physical ones, and this must come about by indirect, empirical inquiry. »

⁴⁰⁷ Et on a vu, par ailleurs, qu'il est aussi plus enclin à rejeter l'intuition incompatibiliste (IA) que l'intuition de l'existence (connue) du libre arbitre (IM).

Une bonne raison de croire à l'indéterminisme devrait forcément contenir des données empiriques à propos du comportement causal des objets physiques, et ne pourrait pas être dérivée de croyances préthéoriques que nous avons à propos de notre statut d'êtres moraux. (Balaguer 2010, 140).

Balaguer reconnaît également qu'il y a, *prima facie*, des raisons d'admettre les deux autres thèses du trilemme d'intégration, pour cet argument van inwagenien à partir de la responsabilité morale, à savoir :

- (1) [Nous savons que] les êtres humains sont parfois moralement responsables.
- (2) La responsabilité morale requiert l'indéterminisme.

Balaguer soutient clairement que « chacune de ces prémisses, prise isolément, pourrait être motivée d'une manière qui est au moins assez plausible » (Balaguer 2010, 138). Mais ce qui n'est pas possible, pour lui, c'est d'admettre les deux conjointement, et cela à cause de l'intuition épistémologique (IE). La manière dont Balaguer montre l'incompatibilité entre (1) et (2) est particulièrement intéressante, parce qu'elle prend la forme d'un argument de réfutation par le contenu alourdi :

Supposez que (2) est vraie, et que nous avons découvert de bonnes raisons de croire à cette thèse, c'est-à-dire de bonnes raisons de croire que la responsabilité morale requiert l'indéterminisme [...]. Alors, ce serait une découverte qui montre que la responsabilité morale est bien plus difficile à obtenir que l'on aurait pu croire [...] et donc nous n'avons pas [ou plus] de bonnes raisons croire à (1).
(Balaguer 2010, 138–139)

Si l'on transpose à l'épistémologie introspectionniste ce que Balaguer dit ici pour l'épistémologie van inwagenienne de la responsabilité morale, on obtient exactement l'argument anti-introspection qui est au cœur de cette section :

Supposez que (IA) est vraie, et que nous avons découvert de bonnes raisons de croire à cette thèse, c'est-à-dire de bonnes raisons de croire que notre concept expérientiel de libre arbitre requiert l'indéterminisme (cf. chapitre 4). Alors, ce serait une découverte qui montre que le libre arbitre est bien plus difficile à obtenir que l'on aurait pu croire, et donc notre justification *prima facie* à croire au libre arbitre (sur la base de notre expérience) se trouverait réfutée (c'est-à-dire que (IM) serait rejetée).

Si on utilise les outils et principes épistémologiques mis en place au chapitre 3, ce problème de réfutation peut se formuler de la manière suivante :

Réfutation fondationnelle de la croyance introspective au libre arbitre

(Justification *Prima Facie*) Notre expérience du libre arbitre nous donne une justification *prima facie* à croire au libre arbitre. J^{pf}_{ELA} .⁴⁰⁸
(IA) Mais [nous reconnaissons que] le libre arbitre (tel que nous en faisons l'expérience) implique l'indéterminisme causal. $J(LA \rightarrow \text{Indet.})$.⁴⁰⁹

⁴⁰⁸ Chapitre 2, sec. II.2.c.

⁴⁰⁹ Chapitre 4, sec. IV.2.d.

(Clôture Disj.)⁴¹⁰ Si nous croyons de manière *prima facie* justifiée que nous sommes libres sur la base de notre expérience du libre arbitre $J_E^{pf}LA$ et si nous reconnaissons que le libre arbitre implique l'indéterminisme causal $J(LA \rightarrow Indet.)$, alors

(a) soit nous sommes en mesure d'avoir la croyance justifiée que nous sommes indéterminés sur la base...		(b) soit notre justification <i>prima facie</i> pour le libre arbitre se trouve par là même réfutée. $\sim J_E LA$.
(i) ... d'une justification antécédente (par une autre source) à croire que nous sommes indéterminés. $J_{S2}Indet.$	(ii) ... d'une justification déductive, en fondant notre croyance à l'indéterminisme causal sur notre expérience du libre arbitre (plus déduction). $J_E Indet.$	

Contre l'alternative (a-i) :

($\sim a-i$) Nous n'avons pas de justification antécédente à croire que nous sommes indéterminés.

Contre l'alternative (a-ii) :

(MC Disj.)⁴¹¹ Si nous croyons de manière *prima facie* justifiée à l'indéterminisme causal sur la base de notre expérience du libre arbitre (plus déduction) $J_E^{pf}Indet.$, et si nous nous interrogeons sur le caractère de *connaissance* de cette croyance $K_E Indet.$, alors

(c) soit nous sommes en mesure d'avoir la croyance justifiée que notre croyance à l'indéterminisme fondée sur cette expérience constitue une *connaissance*. $JK_E Indet.$

(d) soit notre justification déductive *prima facie* pour la croyance indéterministe se trouve par là même réfutée. $\sim J_E Indet.$

(IE) Or ($\sim c$) [nous reconnaissons que] l'expérience du libre arbitre ne peut pas constituer une source de *connaissance* de l'indéterminisme. $\sim JK_E Indet.$

($\sim a-ii$) Donc (d) nous ne sommes pas en mesure d'avoir la croyance (*ultima facie*) justifiée que nous sommes indéterminés sur la base de notre expérience du libre arbitre. $\sim J_E Indet.$

(C1) Donc ($\sim a$) nous ne sommes pas en mesure d'avoir la croyance justifiée que nous sommes indéterminés ni (i) sur une base indépendante de E , ni (ii) sur la base de E .

(C2) Donc (b) notre justification *prima facie* à croire au libre arbitre sur la base de notre expérience agentive E est *ultima facie* réfutée.

On peut reformuler cet argument de manière plus simple en remarquant que, par son usage des principes de clôture et de méta-cohérence, il s'agit en fait d'un argument « du double engagement », très proche de celui que nous avons vu au chapitre 3 (p.219).

Argument anti-introspection du double engagement

Supposons que nous possédons, de par notre expérience du libre arbitre E , une justification *prima facie* à croire au libre arbitre $J_E^{pf}LA$ et même une justification *prima facie* à croire que nous le savons $J_E^{pf}KLA$. A t_0 , nous croyons LA sur la base de notre expérience E , et comme nous n'avons aucune autre croyance occurrente sur la question, nous sommes *ultima facie* justifiés à t_0 . Mais à t_1 , nous prenons conscience que notre expérience E implique l'indéterminisme:

⁴¹⁰ Chapitre 2, p.159.

⁴¹¹ Chapitre 3, p.200.

(IA) Nous reconnaissons, après réflexion, que nous pouvons déduire validement *Indet* de *LA*, donc :

(Clôture Engag.) Pour pouvoir continuer après réflexion à croire *LA* de manière justifiée (pour n'être pas réfuté par inconséquence), nous sommes engagés à croire *Indet* de manière justifiée.

(MC Engag.) Pour pouvoir continuer après réflexion à croire *Indet* de manière justifiée (pour n'être pas réfuté fondationnellement), nous sommes engagés à croire *KIndet* de manière justifiée.

(IE) Nous ne *pouvons pas* satisfaire l'engagement de croire *KIndet* de manière justifiée.⁴¹²

(C) Donc ($\sim$ IM), nous ne pouvons pas continuer après réflexion à croire *LA* de manière justifiée.

Que penser de cet argument ? Comme van Inwagen, je reconnais volontiers qu'il s'agit là d'un argument très fort contre la croyance introspective libertarienne. Dans la mesure où j'ai admis en première partie les principes de clôture (disjonctive) et de méta-cohérence (disjonctive), la seule manière de refuser cet argument est de rejeter l'une des prémisses (IA) ou (IE) ; c'est-à-dire que cet argument fait apparaître de nouveau que le problème d'Intégration est un *trilemme*, entre (IA), (IM), et (IE). Nous avons vu la réaction de Balaguer, qui consiste à rejeter (IM) au profit d'une attitude « d'empiriste libertarien ». Nous avons mentionné au chapitre 3 la possibilité d'abandonner (IA), c'est-à-dire de devenir compatibiliste « d'ajustement » pour éviter le problème d'Intégration. Nous avons vu également l'attitude de van Inwagen qui semble ne refuser le principe épistémologique (IE) *que pour* répondre au trilemme, parce qu'il s'agit pour lui de l'intuition la plus faible. C'est ce que j'ai appelé au chapitre 3 une solution « d'Intégration forcée » ; cela revient à adapter son épistémologie à ce que l'on veut pouvoir dire à propos du libre arbitre.

Pour ma part, si je reconnais une certaine plausibilité initiale à l'intuition (IE), il ne me semble pourtant pas qu'elle soit suffisamment motivée en elle-même : il n'est pas nécessaire, à mon sens, de vouloir « sauver » (IM) et (IA) pour rejeter l'intuition (IE). Lorsqu'on prête attention à ce qu'est une *connaissance*, on s'aperçoit de manière indépendante que l'expérience du libre arbitre pourrait en fait constituer une source de connaissance. La solution que je propose est donc une solution « d'Intégration directe » et non d'Intégration forcée. C'est ce que je m'attacherai à montrer dans cette dernière section. Pour montrer que la croyance à l'indéterminisme (sur la base de notre expérience du libre arbitre) *peut* constituer une connaissance, je partirai des théories de la connaissance que nous avons défendues en première partie ; ces théories de la connaissance ont été établies de manière indépendante du problème du libre arbitre, à partir du cas de la perception sensible. Je commencerai donc par rappeler les conditions que doit satisfaire une croyance pour éviter le « hasard épistémique » et pour constituer une connaissance. Puis j'appliquerai ces conditions à la croyance libertarienne pour montrer qu'elle peut les satisfaire ; c'est-à-dire que je montrerai que l'expérience libertarienne peut dépendre de manière non hasardeuse du « fait » de l'indéterminisme libertarien. Dans la mesure où nous avons vu deux interprétations différentes du « fait » métaphysique du libertarianisme – l'interprétation agentive et l'interprétation événementialiste – il faudra aussi donner deux théories épistémologiques distinctes du lien expérience-fait. Après avoir présenté ces deux modèles épistémologiques, je rappellerai en conclusion comment ce genre de modèle entre dans la dialectique de l'introspectionnisme.

⁴¹² Il s'agit plutôt d'une conséquence immédiate de (IE) que de l'intuition (IE) elle-même : si l'on reconnaît la vérité de $\sim$ *KIndet* – intuition (IE) – alors il est évident qu'on ne peut pas croire *KIndet* de manière justifiée.

II.2. Conditions à satisfaire pour être une *connaissance*

Pour voir si la croyance introspective à l'indéterminisme peut constituer une connaissance, commençons donc par rappeler les conditions que doit satisfaire une croyance pour constituer une *connaissance*.

Nous avons vu en introduction (p.19) que le cœur analytique de la notion de « connaissance » résidait dans la contrainte anti-hasard :

(AH) S sait que *p* seulement si ce n'est pas par pur hasard que sa croyance que *p* est vraie.

Plus précisément, nous avons choisi de focaliser notre attention, parmi les différentes acceptions possibles du mot « connaissance », sur la notion qui se *réduit* pour ainsi dire à la condition anti-hasard (plus, bien sûr, les conditions de croyance et de vérité). C'est-à-dire que nous avons choisi, dans l'économie de cette étude, d'utiliser le mot « connaissance » dans le sens de « croyance vraie épistémiquement non hasardeuse », rejetant ainsi les éventuelles conditions supplémentaires (notamment les conditions internalistes de justification) qui pourraient constituer des sens de « connaissance » plus riches. Une raison de ce choix, comme on l'a vu en introduction, était de pouvoir séparer clairement ce qui relève des conditions épistémologiques externalistes de ce qui relève des conditions épistémologiques internalistes. Une autre raison apparaît plus clairement à présent. Dans le problème d'Intégration du libertarianisme, le cœur de l'objection n'est pas de dire que nous n'avons aucune *justification internaliste* pour croire à l'indéterminisme. Si tel était le problème, il ne serait pas très difficile de prétendre que nous avons une justification déductive à partir de notre justification à croire au libre arbitre. S'il est difficile de reconnaître un statut épistémologique positif à notre croyance introspective à l'indéterminisme, c'est pour des raisons proprement externalistes. C'est parce qu'il semble bien que notre expérience agentive « n'ait rien à voir » avec les *faits* qu'elle prétend décrire : l'une est un phénomène psychologique macroscopique, l'autre concerne « les détails et les fonctionnements microscopiques des systèmes nerveux » (van Inwagen 1983, 210) ou encore les « ruminations et résultats mystérieux des physiciens et cosmologistes théoriciens » (Fischer 2006, 5). Autrement dit, il semble que notre croyance introspective à l'indéterminisme, même si elle était vraie, ne pourrait être vraie que par pur hasard épistémique. Tel est le cœur du problème d'Intégration : pour résoudre ce problème, par conséquent, il faut et il suffit de montrer que l'expérience du libre arbitre peut dépendre de manière non hasardeuse de son contenu, et constituer une « connaissance » au moins dans ce sens minimal anti-hasard.

Reste à savoir ce qu'est le « hasard épistémique » et quelles sont les conditions à remplir pour qu'une croyance soit « non hasardeuse ». Nous avons eu l'occasion d'envisager plusieurs modèles anti-hasard dans les chapitres 2 et 3, mais nous avons vu que certains d'entre eux n'étaient pas acceptables parce qu'ils conduisaient au scepticisme pour la perception sensible.

Le premier modèle que nous avons vu, au chapitre 2 (II.2.c. p.135), est celui de l'infailibilisme. L'infailibilisme part de l'idée que la notion anti-hasard doit s'interpréter comme une notion essentiellement antinomique avec les possibilités d'erreur. L'intuition qui est au cœur de cette analyse est la suivante :

(PE) Si S croit quelque chose de vrai mais aurait aussi bien pu se tromper, alors le fait qu'il croit quelque chose de vrai est un pur hasard épistémique.

A partir de cette idée qu'il y a une certaine exclusion logique entre connaissance (anti-hasard) et possibilités d'erreur, une théorie qui émerge naturellement est la théorie selon laquelle la connaissance exclut *toute* possibilité d'erreur, c'est-à-dire la théorie infaillibiliste :

(Infaillibilisme) Si S sait que p , alors il n'y a aucune possibilité d'erreur pour sa croyance que p .

ou encore

(Infaillibilisme) Si S sait que p , alors dans tous les mondes possibles où S croit que p , p est vraie.

Chacun sait que cette théorie infaillibiliste de la connaissance mène directement au scepticisme : les scénarios philosophiques du Malin Génie ou des cerveaux dans des cuves nous rappellent qu'il est *possible* (au sens métaphysique) pour chaque croyance perceptive, et même pour toutes les croyances perceptives ensemble, d'être erronées. Il n'est donc pas plausible d'exiger l'infailibilité. Dans le chapitre 2, nous avons conclu que les croyances devaient certainement être *fiabiles* pour éviter le hasard épistémique et constituer des connaissances, mais qu'on ne saurait exiger qu'elles soient *infaillibles*. Ainsi nous avons déterminé plus précisément ce que *n'est pas* la condition anti-hasard (ce n'est pas une condition d'infailibilité), mais nous avons encore laissé ouverte l'interprétation de ce qu'est, positivement, la condition de *fiabilité*.

Au chapitre 3 (III.3), nous avons considéré deux interprétations positives de la condition de fiabilité : la condition de sensibilité (ou de traçage) et la condition de sûreté. Ces deux conditions ont en commun avec la notion d'infailibilité d'être des conditions « modales » (qui peuvent s'exprimer en termes de « mondes possibles »), mais ce sont des conditions modales beaucoup moins exigeantes. La condition de sensibilité (voir ch.3 p.225) interprète la dépendance fait-croyance de la condition anti-hasard comme une dépendance contrefactuelle :

(Condition de Sensibilité)

Si S sait que p , alors le conditionnel suivant est vrai :

(3) Si p n'était pas vraie, S ne croirait pas que p .

ou encore

(Condition de Sensibilité)

Si S sait que p , alors

(3) dans le monde possible le plus proche où p est fausse, S ne croit pas que p .

Cette condition peut s'interpréter, d'une certaine manière, comme une forme de rejet de possibilités d'erreur puisqu'une croyance non sensible est une croyance qui a le défaut suivant : il aurait suffi que son contenu soit faux pour que S se trouve dans l'erreur (en continuant à croire ce contenu). L'idée fondamentale de la condition de sensibilité, c'est que lorsqu'on fait varier le fait qui est cru, la croyance « suit » cette variation « à la trace », ce qui montre bien que la croyance *dépend* du fait qui est cru (même si elle n'en dépend pas de manière logique ou nécessaire, comme dans l'infailibilisme). Cette condition de sensibilité a une caractéristique très intéressante, c'est qu'elle n'est pas « close sous l'implication logique ». C'est-à-dire qu'une croyance que p peut être sensible au fait que p , mais ne *pas* être sensible au fait que q bien que $p \Rightarrow q$. L'exemple de Dretske que nous avons vu est celui de la croyance Z (ceci est un zèbre) qui implique logiquement la proposition $\sim MD$ (ceci n'est pas

une mule habilement déguisée) : ma croyance que ceci est un zèbre peut être *sensible* (si ceci n'était pas un zèbre, je ne croirais pas que c'est un zèbre) sans que la croyance $\sim MD$ soit sensible (si ceci était une mule habilement déguisée, je croirais tout de même que c'est un zèbre). Dretske voit dans cette propriété de non-clôture une solution au problème du scepticisme : puisque la sensibilité n'est pas close sous la déduction logique, on peut connaître une proposition p sans connaître ses conséquences logiques, et donc le fait que nos croyances perceptives aient des conséquences inconnaissables (la conséquence que nous ne sommes pas manipulés par un Malin Génie) n'empêche pas ces croyances de constituer des connaissances. Ce sont simplement des connaissances dont nous ne connaissons pas les conséquences logiques.

Nous avons vu cependant que cette réponse au scepticisme ne permettait pas de contrecarrer l'argument du « double engagement » : si nous voulons vraiment répondre aux arguments sceptiques par hypothèses d'illusion (scénario du Malin Génie ou des cerveaux dans des cuves), nous sommes obligés de reconnaître un certain engagement épistémique non seulement à *croire* à la négation de ces hypothèses sceptiques (par engagement de clôture), mais aussi à croire que ces négations d'hypothèses sceptiques constituent des *connaissances* et non pas des croyances purement hasardeuses (par engagement de *méta-cohérence*). Or la théorie de la sensibilité ne permet pas de dire que nos croyances anti-sceptiques (les négations d'hypothèses sceptiques) sont des connaissances, et par conséquent, elle mène aussi au scepticisme, par une voie que Dretske semble n'avoir pas aperçue.

La condition anti-hasard qui permet de considérer les croyances anti-sceptiques comme des connaissances est la condition de « sûreté » (voir ch.3 p.229), dont la formulation par Williamson reprend presque directement l'intuition anti-possibilité d'erreur évoquée plus haut :

(Condition de Sûreté)

Si l'on sait, on ne pourrait pas facilement s'être trompé dans un cas semblable.

Cette notion intuitive s'exprime de manière plus précise en termes modaux :

(Condition de Sûreté)

La croyance de S que p est sûre ssi

dans tous (ou presque tous) les mondes possibles du voisinage dans lesquels S croit que p (sur la même base épistémique que dans le monde actuel), p est vraie.

Comme on l'a vu, toute l'originalité de cette notion de sûreté épistémique est de reposer sur un « voisinage modal », un ensemble de mondes possibles suffisamment proches pour être pertinents dans l'investigation des « possibilités d'erreur ». Qu'il existe des possibilités d'erreur n'est pas en soi gênant si ces possibilités sont extrêmement éloignées du monde actuel (la possibilité d'être manipulé par un Malin Génie, la possibilité d'être un cerveau dans une cuve). La connaissance est incompatible avec les possibilités d'erreur, mais seulement avec celles qui se situent à l'intérieur de notre voisinage modal. Par conséquent, nos croyances anti-sceptiques (la croyance que je ne suis pas un cerveau dans une cuve) peuvent tout à fait constituer des connaissances au sens de croyances non hasardeuses : ce n'est pas « par pur hasard » qu'elles sont vraies, car elles n'auraient pas pu facilement être fausses ; il aurait fallu une situation extraordinairement différente de la nôtre pour qu'elles puissent être erronées.

Pour conclure, quelles conditions une croyance doit-elle satisfaire pour constituer une connaissance ? Le principe que nous suivons ici, c'est qu'il ne faut pas placer la barre des

exigences trop haut, autrement les perceptions sensibles elles-mêmes ne pourront pas l'atteindre et on sera conduit au scepticisme radical. Par conséquent, les seules conditions anti-hasard qu'on peut imposer aux croyances sont les conditions anti-hasard que satisfont les croyances perceptives. Comme les croyances perceptives ne sont pas infaillibles, on ne peut pas interpréter la condition anti-hasard comme une condition d'infaillibilité. En revanche, les croyances perceptives respectent clairement la condition de *sûreté*. Il est donc très probable que pour être non hasardeuse épistémiquement, et pour constituer une connaissance, une croyance doit être *sûre*.

La question est plus difficile pour la condition de *sensibilité* : on a vu que cette condition de sensibilité ne pouvait pas être respectée par les croyances anti-sceptiques elles-mêmes (les croyances dont le *contenu* lui-même, et non la conséquence logique, est une proposition anti-sceptique) mais qu'elle pouvait être respectée par les croyances perceptives elles-mêmes (les croyances de contenu anodin, qui ont seulement pour conséquence logique une proposition anti-sceptique). Dans la mesure où nous avons dit que les croyances anti-sceptiques devaient elles aussi être considérées comme des connaissances, il est clair qu'on ne peut pas faire de la condition de sensibilité une condition nécessaire pour *toute* connaissance : les croyances anti-sceptiques sont non hasardeuses *du simple fait* qu'elles sont sûres, et elles n'ont pas besoin d'être sensibles. Mais les croyances anti-sceptiques ne sont pas à proprement parler des croyances *perceptives* et on pourrait défendre qu'il existe une contrainte supplémentaire sur les croyances perceptives pour qu'elles soient, de manière appropriée, non hasardeuses. C'est-à-dire qu'on pourrait défendre une double contrainte anti-hasard sur les croyances perceptives :

(Double contrainte anti-hasard sur les croyances perceptives)

L'expérience perceptive que *p* peut fournir à *S* une connaissance que *p* (et éviter le hasard épistémique) ssi

- (i) l'expérience de *S* est *sensible* à l'égard de son contenu propre (elle *trace* le fait que *p*)
- (ii) la croyance de *S* dans les conséquences logiques de *p* est *sûre*.

Les croyances perceptives satisfont cette double condition : (i) leur contenu propre (voici deux mains) est sensible, et trace le fait connu ; et les conséquences logiques lourdes de ce contenu propre (je ne suis pas un cerveau dans une cuve), bien que non sensibles, sont sûres. Notez qu'on *pourrait* défendre que la sensibilité des croyances perceptives n'est pas ici pertinente, et que le fait qu'elles soient sûres (pour leur contenu propre comme pour ses conséquences logiques) est *suffisant* pour dire qu'elles sont des connaissances. Mais cela n'est pas évident, et par principe je m'attacherai donc ici à la « plus forte contrainte » que peuvent satisfaire les croyances perceptives. Or la plus forte contrainte qu'on peut poser sur les croyances perceptives sans mener au scepticisme est, à mon sens, la double contrainte que je viens de formuler. C'est donc cette contrainte que nous allons appliquer à présent à l'expérience introspective du libre arbitre, pour voir si elle peut ou non la respecter.

II.3. Modèle d'intégration pour l'indéterminisme du libertarianisme événementiel

L'objection générale à laquelle nous tentons de répondre ici est l'intuition que la croyance introspective au libertarianisme ne *peut pas* constituer une connaissance, parce qu'il ne *peut pas* y avoir une dépendance adéquate entre notre expérience du libre arbitre d'une part et d'autre part le *fait* du libre arbitre lorsque celui-ci est décrit de manière libertarienne. Mon intention sera de montrer qu'une dépendance adéquate entre fait et expérience n'est en fait pas

impossible. Et la manière classique en philosophie de montrer que quelque chose n'est pas impossible, c'est d'en donner un modèle. Cependant, l'objection prendra des formes différentes selon l'interprétation précise que l'on donne du fait métaphysique du libertarianisme, et par conséquent, la réponse (le modèle) prendra également des formes différentes selon cette interprétation.

Je commencerai par envisager le fait libertarien tel qu'il est décrit et interprété dans la théorie événementielle de Robert Kane ou Mark Balaguer. Après avoir rappelé la description métaphysique du *fait* libertarien dans cette théorie, je montrerai que l'expérience du libre arbitre peut dépendre de ce fait de manière non hasardeuse, selon la théorie anti-hasard présentée dans la section précédente.

En reprenant l'analyse donnée au chapitre 5 (p.390), les conditions métaphysiques du libertarianisme événementiel peuvent être résumées de la manière suivante :

Fait métaphysique du libre arbitre d'après le modèle de Kane

Lorsqu'un agent *S* fait un choix difficile, ses deux efforts concurrents sont représentés physiquement par deux réseaux neuronaux récurrents, connectés entre eux. Cette configuration neuronale particulière présente un fonctionnement chaotique susceptible d'amplifier des indéterminismes microscopiques. Par ailleurs, des indéterminismes microscopiques existent au niveau des neurones individuels, si bien que l'issue des contributions causales des deux réseaux neuronaux est réellement indéterministe. (voir Kane 2005, 133–139)

L'essentiel, dans ce modèle, c'est la structure de « double réseau neuronal récurrent » qui fonctionne comme « amplificateur chaotique » des indéterminismes microscopiques. Il est peut-être utile de rappeler ici qu'il ne s'agit que d'un *modèle*, c'est-à-dire d'une construction théorique visant à montrer la *possibilité* (et la concevabilité) du libre arbitre indéterministe. Un auteur libertarien n'est pas engagé à soutenir que les choses se passent *exactement* comme cela : on pourrait donner d'autres modèles du libertarianisme, qui différeront soit sur des détails, soit même sur des points importants (un autre modèle, sensiblement différent, est bien sûr le modèle agentif que nous allons reprendre un peu plus loin).

Tel est donc le fait à connaître. L'expérience du libre arbitre, quant à elle, est un événement mental dont on peut supposer qu'elle est réalisée ou représentée par un certain état ou par certains événements neuronaux. La question est de savoir si cette expérience (ou sa réalisation neuronale) peut dépendre adéquatement du fait décrit dans le modèle.

L'objection d'intégration, appliquée au modèle de Kane, prendra sans doute la forme suivante :

Objection d'Intégration au modèle de Kane

L'état neuronal (ou les événements neuronaux) qui réalisent l'expérience du libre arbitre ne peuvent pas détecter la présence ou l'absence d'indéterminismes au niveau microscopique (quantique), car ils se situent à une échelle bien supérieure. Par conséquent, l'expérience du libre arbitre ne peut pas *dépendre adéquatement* du fait métaphysique qu'elle représente, quand bien même ce fait serait actuel. C'est-à-dire qu'elle ne peut être véridique, au mieux, que par pur hasard épistémique.

Nous avons donc à présent le fait, l'expérience, et le problème de dépendance entre les deux. Nous allons voir à présent s'il est possible de concevoir entre l'un et l'autre une dépendance épistémique adéquate. Dans ce qui suit, je défendrai une réponse positive. Mais il est intéressant de noter que les principaux défenseurs de ce modèle libertarien (Kane et Balaguer)

répondent négativement à cette question : pour eux, ce modèle « scientifique » du libertarianisme ne peut être établi et connu que sur une base empirique, et non pas sur la base de notre expérience du libre arbitre. La position « introspectionniste événementielle » n'a, à ma connaissance, aucun défenseur aujourd'hui. Il me semble que cette absence dialectique repose sur une erreur épistémologique, comme on va le voir.

Pour voir si notre expérience libertarienne peut dépendre de manière non hasardeuse du fait décrit par Kane, il faut repartir de la condition anti-hasard décrite dans la section précédente. Nous avons vu que, dans le cas des expériences perceptives, il fallait distinguer entre le contenu propre de l'expérience et les conséquences logiques de l'expérience, les premières satisfaisant une forme de dépendance anti-hasard plus forte que les secondes, à savoir la sensibilité (ou le traçage). Dans le cas de l'expérience du libre arbitre, on peut également distinguer entre le contenu propre de l'expérience et ses conséquences logiques. Le contenu propre de l'expérience du libre arbitre est de la forme suivante :

Contenu propre de l'expérience du libre arbitre :

Cette action (que je suis présentement en train d'accomplir) est libre (de manière indéterministe).

Quant à la thèse indéterministe en général, elle est logiquement *impliquée* par ce contenu propre (tout comme « ceci n'est pas une mule déguisée » est impliqué par « ceci est un âne ») mais elle n'est pas *le* contenu propre de l'expérience. Compte tenu de la « double contrainte » anti-hasard que nous avons posée, la question d'intégration pour le modèle de Kane sera double :

(i) Question de sensibilité : L'expérience de telle action libre peut-elle dépendre de manière sensible du fait que « cette action (que je suis présentement en train d'accomplir) est libre (de manière indéterministe) » ?

(ii) Question de sûreté : Notre croyance dans l'indéterminisme comme thèse générale est-elle *sûre* à l'égard du fait de l'indéterminisme ?

Si la réponse à chacune de ces questions est positive, alors notre expérience du libre arbitre indéterministe constituera une source de connaissance selon les mêmes standards que les expériences perceptives. Si la réponse à ces questions *peut* être positive (si on peut modéliser une réponse positive), alors cela réfutera l'idée pessimiste que notre expérience du libre arbitre indéterministe ne *peut pas* constituer une source de connaissance. C'est la deuxième conclusion, la conclusion de possibilité, que je m'efforcerai d'établir ici.

Commençons par la question la plus simple, la question de sûreté. Si nous croyions au libre arbitre sur la base de notre expérience, et que nous en déduisions que nous vivons dans un monde indéterministe, cette croyance indéterministe serait-elle *sûre* à l'égard du fait de l'indéterminisme ? La réponse est très certainement « oui », si l'on prête attention à la définition modale de la sûreté, c'est-à-dire qu'une croyance est sûre si elle ne peut être fausse que dans des monde possibles très éloignés du nôtre, des mondes possibles qui se situent en dehors de notre voisinage modal. Mais il est clair que si le monde actuel est un monde dont les lois de la nature sont indéterministes, ce qui semble bien être le cas, alors les mondes possibles déterministes sont des mondes extraordinairement différents du nôtre. C'est-à-dire que, du point de vue de notre monde actuel indéterministe :

(Sûreté Indéterministe) L'indéterminisme est vrai dans tous les mondes possibles de notre voisinage modal.

et donc

(Sûreté de la croyance indéterministe) Notre croyance indéterministe est telle que dans tous (ou presque tous) les mondes possibles du voisinage dans lesquels nous la croyons (sur la même base épistémique que dans le monde actuel), elle est vraie.

Quelques remarques peuvent aider à bien comprendre cette conclusion.

Premièrement, comme nous l'avons vu au chapitre 5, il n'est pas absolument *évident* que nous vivons dans un monde indéterministe ; il existe des interprétations déterministes des résultats de la physique quantique. Or si nous sommes, en fait, dans un monde déterministe, il est clair que l'indéterminisme se trouvera finalement être *faux* dans tous les mondes de notre voisinage modal. Mais cette difficulté est indépendante du problème épistémologique que nous traitons ici. Il s'agit d'une objection *directe* au libertarianisme, que nous avons déjà envisagée au chapitre 5. L'objection épistémologique consiste à dire que *même si elle était vraie*, la croyance indéterministe ne pourrait pas être une connaissance, car elle ne serait vraie que par pur hasard ; or ce que nous venons justement de montrer, c'est que *si elle est vraie* (si on se place du point de vue d'un monde indéterministe), alors elle satisfait la condition anti-hasard de sûreté. Cette conclusion conditionnelle est tout ce dont nous avons besoin pour répondre à l'objection proprement épistémologique.

Deuxièmement, l'idée que les mondes possibles déterministes sont très éloignés modalement des mondes indéterministes (et que les premiers n'appartiennent donc pas au « voisinage modal » des seconds) pourrait être remise en question. Une manière de mettre en question cette idée serait de dire la chose suivante : la question du déterminisme a longtemps été, et reste dans une certaine mesure, une question très difficile à trancher, ce qui montre que nous avons beaucoup de mal à « discriminer » entre mondes indéterministes et mondes déterministes. Les premiers ne doivent pas être si radicalement différents des seconds, puisque nous pouvons facilement « douter » si nous sommes dans l'un ou dans l'autre, nous pouvons facilement « confondre » l'un avec l'autre. Cette objection repose sur une mauvaise compréhension du caractère fondamentalement *externaliste* de la notion de sûreté. Comparons avec le cas des hypothèses sceptiques : nous avons vu que les croyances anti-sceptiques peuvent être appelées des connaissances parce qu'elles sont sûres, c'est-à-dire qu'elles ne sont fausses que dans des mondes « très éloignés ». Mais évidemment, ces mondes très éloignés du point de vue *objectif* ou *externaliste* sont aussi très difficiles à distinguer du monde normal du point de vue *subjectif* ou *internaliste*. En eux-mêmes, les mondes où il n'existe que mon âme et un Malin Génie sont objectivement très différents du monde où il existe une terre, de nombreux humains et mon corps, mais du point de vue internaliste il est impossible de discriminer entre les deux. De la même façon, le fait de souligner que nous avons du mal à discriminer entre mondes déterministes et mondes indéterministes n'est qu'une remarque internaliste, qui ne nous apprend rien sur le caractère objectivement semblable d'un monde déterministe et d'un monde indéterministe.

Troisièmement, si on s'intéresse à la question proprement externaliste de savoir si les mondes déterministes sont « semblables » ou « très éloignés » des mondes indéterministes, la réponse semble bien être que les deux types de monde sont objectivement très dissemblables, et ce notamment à cause du caractère fondamental des lois de la nature. C'est un principe souvent invoqué en métaphysique modale que « les [contrefactuels qui reposent sur des] lois de la nature sont des caractéristiques du monde plus fondamentales que les faits particuliers » (Flint 1998, 141; voir Hasker 1989, 47). Toutes choses égales par ailleurs, un monde possible

qui diffère du nôtre par un fait particulier (mais où toutes les lois sont les mêmes que dans le nôtre) est moins éloigné du nôtre qu'un monde possible qui diffère du nôtre par une loi de la nature différente.⁴¹³ Or un monde déterministe et un monde indéterministe différeront probablement par un très grand nombre de lois, et surtout par une caractéristique très fondamentale de ces lois. Par conséquent, il semble très raisonnable de supposer qu'un monde déterministe sera, toutes choses égales par ailleurs, très éloigné modalement d'un monde indéterministe. Ne pourrait-il pas y avoir des mondes déterministe et indéterministe qui soient, individuellement, proches l'un de l'autre, même si l'on admet que statistiquement et généralement le groupe des mondes indéterministes est éloigné des mondes déterministes ? C'est possible, mais ce ne serait pas très problématique pour la condition de sûreté telle que nous l'avons définie : pour qu'une croyance soit sûre, il n'est pas nécessaire qu'elle soit vraie dans *tous* les mondes possibles du voisinage, mais seulement dans *presque tous* les mondes du voisinage. Par conséquent, pour que la croyance indéterministe soit sûre, il suffit que la grande majorité des mondes du voisinage soient indéterministes ; si quelque monde déterministe se retrouve dans notre voisinage modal, ce n'est pas très grave du moment que la très grande majorité des mondes déterministes est trop éloignée pour être modalement pertinente.

Par conséquent, à la question de sûreté, la réponse est très certainement « oui » : indépendamment du problème d'Intégration, indépendamment de toute motivation que l'on peut avoir à sauver le libre arbitre, il est très probable que notre croyance indéterministe sera une croyance épistémiquement *sûre*. Certains estimeront peut-être que la sûreté n'est pas suffisante, que c'est une condition trop facile à satisfaire. C'est pourquoi nous allons passer à présent à la condition de sensibilité (pour le contenu propre de l'expérience). Mais quoi qu'il en soit, il semble difficile de nier que la croyance indéterministe, en elle-même, est sûre.

Passons à la question de sensibilité : notre expérience (et notre croyance) que « cette action est libre (de manière indéterministe) » peut-elle dépendre de manière sensible de son contenu, peut-elle « tracer » la vérité de son contenu ? Pour répondre à cette question, il faut, là aussi, repasser par la définition modale de la sensibilité :

(Condition de Sensibilité)

(3) Si p n'était pas vraie, S ne croirait pas que p .

c'est-à-dire :

(3) Dans le monde possible le plus proche où p est fausse, S ne croit pas que p .

La question de sensibilité pour l'expérience libertarienne est donc de savoir si les propositions suivantes sont vraies ou peuvent être vraies :

(Sensibilité de l'expérience libertarienne)

Si cette action n'était pas libre de manière indéterministe, je n'aurais pas l'expérience qu'elle l'est

c'est-à-dire

Dans le monde possible le plus proche où cette action n'est pas libre de manière indéterministe, je n'ai pas l'expérience qu'elle l'est.

⁴¹³ Flint (*ibid.*) souligne que *certaines faits* sont sans doute plus fondamentaux que n'importe quelle loi de la nature. Cela permet peut-être de remettre en question le principe absolu selon lequel toute loi de la nature est plus fondamentale que tout fait ; mais cela ne remet pas forcément en cause le principe général selon lequel, toutes choses égales par ailleurs et en moyenne, les différences de faits « éloignent » moins modalement que les différences de lois.

Mais quel est le monde le plus proche dans lequel *cette action* n'est pas libre de manière indéterministe ? En se rappelant que le monde « actuel » (le monde de référence) est celui du modèle de Kane (c'est-à-dire un monde dans lequel cette action est indéterministe parce qu'un double réseau neuronal sert d'amplificateur chaotique aux indéterminismes quantiques), on peut envisager deux candidats au titre de « monde le plus proche » où cette action est déterministe.

W1 : cette action repose sur le même type de double réseau neuronal que dans le monde actuel, et ce double réseau neuronal fonctionne également comme système chaotique. Seulement, le fonctionnement des particules élémentaires est strictement déterministe, donc il n'y a aucun indéterminisme microscopique à « amplifier », et le système dans son ensemble fonctionne par conséquent de manière purement déterministe.

W2 : comme dans le monde actuel, les lois de la physique quantique sont indéterministes, mais cette action ne repose pas sur un double réseau neuronal faisant fonction d'amplificateur quantique. Par conséquent, il y a bien des indéterminismes quantiques dans les composants microscopiques de cette action, mais ces indéterminismes sont annulés au niveau macroscopique de l'action elle-même.

Chacun de ces deux mondes possibles est un monde dans lequel cette action ne s'est pas déroulée de manière indéterministe. Mais lequel est *le plus proche* du monde actuel. La réponse est simple si l'on se réfère à ce que l'on vient de voir sur le caractère fondamental des lois de la nature : le monde W1 est clairement un monde très éloigné du monde actuel puisqu'il en diffère par un passage à des lois déterministes. Le monde W2 en revanche ne diffère du nôtre que par un fait local, le fait que cette action repose sur un certain type de réseau neuronal plutôt que sur un autre type de réseau neuronal. Cette possibilité est certainement une possibilité très proche, d'autant plus proche qu'elle est très probablement réalisée dans le « monde actuel » du libertarien. En effet, même si l'on est libertarien et qu'on admet que certaines actions reposent sur des réseaux neuronaux récurrents tels que Kane les décrit, on admettra certainement que *d'autres* actions ne sont pas libres dans ce sens et reposent sur des réseaux neuronaux plus simples qui n'amplifient pas les indéterminismes microscopiques.

La question de sensibilité revient donc à se demander ceci : est-il possible que la proposition suivante soit vraie ?

(Sensibilité de l'expérience libertarienne)

Si cette action reposait sur un système neuronal non-chaotique (et qui par conséquent annule les indéterminismes microscopiques), je n'aurais pas l'expérience d'indétermination que j'ai dans le monde actuel.

Ou pour le dire autrement, en termes de « traçage », la question de sensibilité revient à se demander ceci :

Question de traçage pour l'expérience libertarienne

Est-il possible que notre expérience du libre arbitre trace de manière fiable la différence entre les réseaux neuronaux qui sont des amplificateurs chaotiques et ceux qui annulent les indéterminismes quantiques ?

Et la réponse à cette question semble bien être « oui ». Autant il est probable (comme le fait remarquer l'objecteur) que notre expérience du libre arbitre ne peut pas tracer

l'indéterminisme de tel événement quantique (parce que notre expérience est d'un niveau trop élevé pour pouvoir tracer quelque chose de microscopique), autant la différence entre une structure de double réseau neuronal récurrent et une structure de réseau neuronal simple ou non récurrent est un fait structurel qui se situe au niveau neuronal lui-même, et donc au bon niveau pour que l'expérience du libre arbitre (ou son corrélat neuronal) ait avec lui une certaine dépendance. C'est-à-dire qu'il *pourrait* exister un certain système neuronal dans le cerveau dont le comportement serait différent selon que le processus décisionnel passe par un réseau neuronal récurrent double, ou par un autre type de réseau neuronal. Et ce système neuronal pourrait être responsable de la production (ou de l'absence) de l'expérience du libre arbitre. Bien sûr, nous n'avons pas, pour l'instant, de preuve que cela se passe ainsi ; l'important ici est de remarquer que, contrairement à l'objection *d'impossibilité*, cela est possible et modélisable.

Modèle de détection de réseaux chaotiques

Chaque fois que notre processus décisionnel est réalisé par une structure neuronale chaotique, un certain réseau neuronal, dédié à cette tâche, détecte ce fait et produit l'expérience du libre arbitre. Et on sait par ailleurs que, dans tous (ou presque tous) les mondes possibles du voisinage, une délibération est indéterministe si et seulement si elle est réalisée par une structure neuronale chaotique. Donc dans le monde possible le plus proche dans lequel « cette action n'est pas libre de manière indéterministe », le réseau décisionnel n'est pas chaotique, et donc nous n'avons pas l'expérience que « cette action est libre de manière indéterministe ».

Dans ce modèle, nous avons donc à la fois :

(i) Notre expérience du libre arbitre trace la vérité de son contenu propre (cette action est libre de manière indéterministe).

et

(ii) Quant à son implication logique, que « l'indéterminisme est vrai », elle est épistémiquement sûre (vraie dans tous, ou presque tous, les mondes possibles proches).

Contrairement à ce que soutenait l'objection initiale, il n'est donc pas impossible que l'expérience du libre arbitre dépende de manière non-hasardeuse du fait libertarien tel que le décrit le modèle événementiel.

D'où venait alors l'objection d'intégration ? Pourquoi *semble-t-il* que l'expérience du libre arbitre soit épistémiquement hasardeuse à l'égard du fait libertarien ? La formulation que nous avons donnée à l'objection peut aider à comprendre l'origine de cette intuition : l'objection soulignait que « l'état neuronal qui réalise l'expérience du libre arbitre ne peut pas *détecter* la présence ou l'absence *d'indéterminismes* au niveau microscopique (quantique) ». Le terme « détecter » suggère fortement une notion de traçage épistémique. Ce que l'objection souligne, c'est donc que notre expérience du libre arbitre ne peut pas être *sensible à* (ou ne peut pas *tracer*) le fait de l'indéterminisme général. Autrement dit, nous n'avons pas le conditionnel suivant :

(Sensibilité à l'indéterminisme)

Si j'étais dans un monde déterministe, je n'aurais pas cette expérience d'action libre indéterministe.

En effet, dans le monde possible le plus proche où les lois sont entièrement déterministes, il est tout à fait possible que nous ayons les expériences libertariennes que nous avons dans le monde actuel. Donc l'objection a *raison* de souligner qu'il n'y a pas de sensibilité de l'expérience du libre arbitre à l'égard de la thèse générale du libre arbitre. Là où l'objecteur se trompe, c'est lorsqu'il croit que l'expérience du libre arbitre *doit* être sensible à l'indéterminisme (général) pour éviter le hasard épistémique. Cela serait le cas uniquement si le principe épistémologique suivant était vrai :

(Principe de sensibilité aux implications logiques)

L'expérience que *p* ne peut être une source de connaissance (ne peut éviter le hasard épistémique) que si cette expérience est sensible à la vérité et à la fausseté de *p* et de ses implications logiques.

Mais on a vu que ce principe était clairement faux : mon expérience que « ceci est un âne » peut m'apporter une connaissance, même si elle n'est pas sensible au fait que « ceci n'est pas une mule déguisée » ; mon expérience que « j'ai deux mains » peut m'apporter une connaissance, même si elle n'est pas sensible au fait que « je ne suis pas un cerveau dans une cuve ». La condition de sensibilité ne peut tout simplement pas être imposée aux conséquences logiques de nos expériences épistémiques, sous peine de mener au scepticisme. Cela ne veut pas dire que la condition de sensibilité n'a aucune pertinence : elle s'applique sans doute au contenu propre de nos expériences. Mais pour les conséquences logiques, la seule condition pertinente pour éviter le hasard épistémique est la condition de sûreté. Là où une épistémologie anti-sceptique exige d'appliquer une condition de dépendance forte (sensibilité) au contenu propre et une condition de dépendance faible (sûreté) aux implications logiques, l'erreur de l'objecteur consiste donc à vouloir étendre la condition de dépendance forte aux conséquences logiques.

Ce diagnostic sur l'erreur de l'objection d'intégration permet aussi de mieux comprendre la conclusion *empiriste* qui lui est classiquement associée, c'est-à-dire l'idée que *seuls les résultats des sciences* permettront de trancher la vérité ou la fausseté du libertarianisme. Si la science, et en particulier la physique fondamentale, est considérée ici comme seule source d'information légitime, c'est probablement parce qu'on estime que ses résultats pourraient être *sensibles* à la vérité ou à la fausseté de la thèse déterministe générale. L'expérience des siècles passés montre que certains résultats expérimentaux peuvent amener la communauté scientifique à favoriser une interprétation déterministe ou indéterministe. Bien sûr, ces conclusions déterministes ou indéterministes sont plutôt la conclusion d'arguments inductifs complexes que la simple lecture d'une expérimentation isolée, mais en simplifiant la situation, on pourrait imaginer que la communauté des physiciens parvienne un jour à construire un *experimentum crucis* tel que

Si le résultat de l'expérience est A, alors probablement nous vivons dans un monde généralement indéterministe ;

Si le résultat de l'expérience est B, alors probablement nous vivons dans un monde généralement déterministe.

Une telle expérience constituerait alors un indicateur *sensible* à la vérité ou à la fausseté de la thèse générale de l'indéterminisme. Et si la science peut apporter ce genre d'indications à propos de l'indéterminisme, alors il est légitime de dire qu'elle constitue une meilleure source d'information que notre expérience du libre arbitre relativement à la thèse du libre arbitre, puisqu'elle satisfait une condition de dépendance épistémique plus forte que la simple sûreté. Mais cela ne signifie pas que l'expérience du libre arbitre soit, en elle-même, insuffisante :

lorsque le témoignage de X prévaut sur le témoignage de Y, cela n'empêche pas que le témoignage de Y soit entièrement recevable en l'absence de témoignage de X. Du reste, nous avons vu que la science contemporaine, loin de donner tort à notre expérience indéterministe, semblait plutôt lui donner raison, si l'on suit l'interprétation majoritaire de la physique quantique.

Pour conclure, si l'on adopte un modèle événementiel du libertarianisme, il est tout à fait possible de concevoir une dépendance anti-hasard adéquate entre notre expérience libertarienne et son contenu. Il est possible, en effet, que notre expérience soit sensible à son contenu propre (cette action est libre de manière indéterministe) et sûre relativement à l'implication logique du déterminisme. Ou, pour le dire autrement, il est possible que notre expérience trace fidèlement l'occurrence d'un fait (les doubles réseaux neuronaux récurrents) qui est lui-même sûrement relié à l'indéterminisme de telle action libre. Ce qu'on ne peut (probablement pas) avoir, c'est un traçage de l'indéterminisme comme thèse générale, mais on ne saurait exiger un traçage des conséquences logiques sans mener au scepticisme radical.

II.4. Modèle d'intégration pour l'indéterminisme de la causalité agentive

Passons à présent au modèle de la causalité agentive. La situation épistémologique est sensiblement différente pour ce second modèle du libertarianisme. En effet, nous avons vu que pour le modèle événementiel, aucun auteur ne défendait une connaissance du libertarianisme par expérience, et il nous a donc fallu construire un modèle original pour défendre la connaissabilité du fait libertarien événementiel. Dans le cas de la causalité agentive en revanche, ses défenseurs (Reid, C.A. Campbell, O'Connor) ont le plus souvent adopté une démarche introspectionniste et ont parfois explicitement défendu la capacité de notre expérience libertarienne à constituer une source de *connaissance*. Nous pourrions donc nous appuyer ici sur les indications de nos prédécesseurs.

Si les défenseurs de la causalité agentive ont davantage soutenu l'épistémologie introspectionniste, c'est sans doute parce que cette épistémologie n'est pas confrontée à une difficulté aussi pressante que dans le cas du modèle événementiel. Pour celui-ci, la grande difficulté, comme on vient de le voir, est une difficulté « d'hétérogénéité des niveaux ontologiques » : l'expérience du libre arbitre se situe au niveau macroscopique et mental tandis que le fait libertarien se situe au niveau microscopique des particules élémentaires. La dépendance épistémique entre les deux pouvait donc sembler difficile à assurer. Mais dans le cas du modèle agentif, ce problème d'hétérogénéité des niveaux ne se pose pas. En effet, il s'agit ici d'établir une dépendance épistémique entre l'expérience agentive (toujours de niveau macroscopique et mental) avec un fait de pouvoir agentif qui se situe aussi au niveau macroscopique de l'agent en tant que substance consciente. Pour rappel, le fait libertarien qu'il s'agit de connaître peut être résumé de la manière suivante :

Fait métaphysique du libre arbitre d'après le modèle agentif

Le corps (ou le cerveau) humain est une substance macroscopique complexe telle que, lorsque ses composants microscopiques atteignent un certain niveau et un certain type appropriés de complexité organisationnelle, une propriété macroscopique émergente (non structurelle) apparaît, à savoir un pouvoir d'action causale « de haut en bas »⁴¹⁴ qui

⁴¹⁴ Voir chapitre 5, sec. II.3.

est un pouvoir de l'agent lui-même.⁴¹⁵ Ce pouvoir de l'agent a une certaine « structure de probabilités », c'est-à-dire que les raisons qui s'offrent à l'agent le disposent à causer telle action avec une probabilité P1, telle autre action avec une probabilité P2, etc., et ces probabilités sont (ou peuvent être) inférieures à 1, de telle sorte que l'exercice effectif du pouvoir agentif n'est pas prédéterminé.⁴¹⁶

Autant le modèle de Kane, avec ses indéterminismes quantiques et ses réseaux neuronaux récurrents, incitait à consulter les données des sciences empiriques comme meilleure source d'information possible, autant le modèle agentif, avec son pouvoir émergent qui est irréductiblement un pouvoir de l'agent, incite naturellement à consulter l'agent lui-même et son expérience agentive. Autrement dit, la nature même du modèle agentif offre de bonnes ressources pour résoudre le problème d'Intégration avec une épistémologie introspectionniste. C'est un point que C.A. Campbell a aperçu avec une particulière netteté :

Il est clairement illégitime pour le Déterministe de chercher à discréditer son adversaire au motif que sa thèse dépend ultimement de rien de plus qu'une « simple impression ». Elle en dépend *effectivement*, mais ce n'est pas un défaut. Cela est dû à la nature même de ce qui est défendu. Si la liberté *existe*, alors c'est sur cette base, et sur aucune autre, qu'on peut en connaître l'existence. (Campbell 1931, 138)

Pour Campbell, les données de l'expérience sont non seulement la *meilleure* source d'information qu'on puisse espérer pour un tel fait métaphysique, elles sont même la *seule* source d'information possible, à l'exclusion de toute donnée expérimentale :

La seule indication [*evidence*] positive qui *puisse* jamais être apportée pour le libre arbitre doit être, à cause de la nature de la chose, du type de « l'expérience immédiate » ou du « sentiment ». (Campbell 1931, 137)

La manière dont Campbell développe cette intuition souligne l'adéquation fondamentale entre une expérience épistémique « interne » à l'agent et un fait métaphysique qui ressortit également de l'intériorité de l'agent :

Chercher un sens au « libre arbitre » dans un point de vue externe est absurde. Cela revient à le chercher d'une manière qui nous assure qu'on ne le trouvera pas. Si l'on admet qu'une activité créatrice est au moins *possible* (et je ne connais aucune raison de la rejeter *a priori*), il n'y a qu'une seule et unique façon dont nous pouvons espérer la saisir, et c'est du point de vue *interne* de la participation directe. [...] Il faudrait être un critique très doctrinaire de l'introspection pour refuser d'en faire usage lorsque, dans la nature même de la chose, aucun autre moyen de saisie n'est disponible. (Campbell 1951, 462–463)

Pour ma part, je n'irais pas jusqu'à dire que l'expérience agentive est l'*unique* source d'information possible qui permettrait d'indiquer l'existence d'un pouvoir agentif libertarien. En effet, je peux *imaginer* (dans un futur assez spéculatif) que des données expérimentales révèlent ou au moins suggèrent l'activité d'un pouvoir émergent « de haut en bas » dans certaines décisions humaines.⁴¹⁷ Mais ce qui m'intéresse ici, c'est que l'expérience agentive

⁴¹⁵ Voir chapitre 5, sec. I.1.b.

⁴¹⁶ Voir chapitre 5, sec. I.2.b.

⁴¹⁷ Le scénario serait le suivant : (i) une certaine partie P du cerveau humain serait telle que, si l'on calcule le fonctionnement prédictible de P sur la base du fonctionnement de ses parties (d'après les lois qui régissent ces

constitue en tout cas, à première vue, une *très bonne* source d'information compte tenu de la nature de la causalité agentive. Autrement dit, quand on aborde la question d'Intégration avec un modèle agentif du fait libertarien on ne part *pas* avec une difficulté initiale d'hétérogénéité des niveaux entre fait métaphysique et expérience épistémique ; au contraire, on part dans la situation idéale d'une congruence des niveaux.

Cette remarque devrait nous donner bon espoir de pouvoir relever le défi de l'Intégration, mais elle ne nous donne pas encore une théorie explicite de la manière dont nous connaissons le pouvoir agentif par expérience. Or comme le dit Christopher Peacocke, cité en exergue de ce chapitre, « c'est une chose de défendre qu'il doit être possible de relever le défi [de l'Intégration]. C'est autre chose de dire comment le relever » (1999, 6). Campbell, malheureusement, n'apporte pas vraiment d'élément pour répondre à cette question du « comment ». C'est elle que nous allons aborder dans cette section.

La question est donc de savoir comment notre expérience du libre arbitre pourrait nous donner une connaissance de la causalité agentive. Comment une telle connaissance est-elle possible et de quel genre de connaissance pourrait-il s'agir ? Un début de réponse est apporté par Timothy O'Connor :

Dans la formation délibérée d'une intention, l'apparition de mon intention ne me semble pas simplement se produire à l'issue de ma délibération ; il me semble que je m'expérimente moi-même comme produisant directement l'intention. [...] Cette apparente perception de la causalité pourrait être erronée, bien sûr ; notre expérience ne peut pas constituer le point de départ d'une preuve démonstrative du libre arbitre. Mais par le même raisonnement, notre perception apparente des objets physiques ordinaires pourrait aussi (dans le sens épistémique) être erronée, et pourtant peu de gens considèrent cela comme une raison de ne pas croire que nous les percevons.
(O'Connor 2000, 124)

Il est intéressant de comparer ce texte avec un texte parallèle un peu plus ancien, où O'Connor fait de nouveau « appel à la manière dont les choses nous semblent lorsque nous agissons » :

De telles expériences pourraient, bien sûr, être entièrement illusoires, mais n'avons-nous pas la présomption légitime, en l'absence de fortes raisons contraires, que les choses sont à peu près telles qu'elles nous apparaissent ? Je ne m'attarderai pas davantage ici sur cette question épistémologique, puisque je m'intéresse ici à la métaphysique descriptive, mais je noterai que le scepticisme à propos de la véridicalité de ces expériences a de nombreux isomorphes qui, si on les accepte, semblent mener à un diagnostic extrêmement pessimiste de notre connaissance du monde, un diagnostic que la plupart des philosophes refusent.
(O'Connor 1995, 199)

Dans chacun des deux textes, on remarque une même stratégie faillibiliste : l'expérience agentive *pourrait* être erronée ou illusoire, mais la simple possibilité d'erreur ne constitue pas un problème de hasard épistémique. Si la simple faillibilité était un hasard épistémique problématique, alors toutes nos croyances perceptives devraient être exclues du champ des connaissances et on tomberait dans un scepticisme radical. Ce point argumentatif est

parties hors de tout système complexe), on obtiendrait une certaine prédiction de comportement C1 ; et (ii) la partie P se trouverait, dans les conditions d'expérimentation, impliqué dans le système complexe d'un processus décisionnel ; enfin (iii) le comportement observé de P dans ce contexte complexe serait C2, distinct de C1, ce qui révélerait l'influence causale irréductible (*top down*) du contexte complexe.

important et légitime, mais il est en lui-même strictement négatif : certes la faillibilité de l'expérience agentive *n'exclut pas* que cette expérience soit fiable (ou adéquatement dépendante du fait décrit), mais cette remarque ne permet pas non plus de dire que cette expérience est positivement fiable.

Deux éléments positifs sont présentés par O'Connor en faveur de la crédibilité de notre expérience. Dans le texte de 1995, O'Connor mentionne la « présomption légitime » que donne toute « apparence » en faveur de son contenu, « en l'absence de fortes raisons contraires ». Cette remarque nous est familière depuis le chapitre 2 : les expériences doxastiques confèrent une justification *prima facie*, qui devient justification *ultima facie* en l'absence de réfuteurs. Mais il est important de noter que cette remarque est d'ordre strictement *justificatif* ou internaliste : elle concerne le poids des raisons à l'intérieur de notre système de croyances et non pas la dépendance externaliste entre fait et croyance. Elle n'apporte donc aucune réponse au problème d'intégration et laisse ouverte la possibilité d'une réfutation de second ordre de notre expérience : en effet, s'il s'avérait que notre expérience agentive n'est pas ou ne peut pas être fiable (sur le plan externaliste), cela constituerait certainement une « forte raison contraire » à l'acceptation de son contenu.

Le texte de 2000 apporte une deuxième indication positive, plus pertinente pour notre problème, lorsqu'il parle de « *perception* apparente de la causalité ». En effet, le concept de « perception » est une espèce du concept de « connaissance », et il implique donc logiquement une forme de dépendance externaliste adéquate entre l'expérience perceptive et le contenu perçu. On pourrait reprocher à O'Connor, dans ce texte, de ne donner aucune indication sur la *manière* dont nous percevons la causalité, mais il faut noter que le terme même de « perception » donne un début de théorisation de la manière dont nous connaissons. Une connaissance *par perception* n'est pas n'importe quel type de connaissance ; la dépendance fait-croyance d'une perception est un certain type particulier parmi les différentes formes de dépendances fait-croyance qui peuvent compter comme connaissances. En particulier, dans une connaissance perceptive, la croyance dépend du fait de manière « *occurentielle* »⁴¹⁸ : lorsque je perçois visuellement la chute d'une pierre, la chute de la pierre est un événement *occurent*, tout comme mon expérience visuelle, et la dépendance épistémique qui permet d'avoir une connaissance visuelle est la dépendance entre ces deux occurrences (l'occurrence de l'expérience étant, typiquement, simultanée ou légèrement postérieure à l'occurrence de l'événement perçu). Nous avons déjà eu l'occasion de souligner que l'épistémologie introspectionniste était fondamentalement une épistémologie *occurentielle*. En choisissant l'expérience introspective comme source de justification interne, nous avons donc posé du même coup une contrainte sur le type de dépendance externaliste qu'il faudra défendre entre fait et croyance. Et O'Connor vient confirmer ce point lorsqu'il décrit notre connaissance des pouvoirs agentifs comme une connaissance *perceptive*.

En somme, si O'Connor ne donne pas un modèle d'Intégration complet, un modèle de la manière dont nous connaissons la causalité agentive par expérience, il donne une première caractérisation du genre de modèle qu'il faudra rechercher, à savoir un modèle perceptif.

Notre question d'Intégration prend donc à présent la forme suivante : comment pouvons-nous *percevoir* la causalité agentive ? Lorsque la question est posée en ces termes, on est immédiatement confronté à un problème traditionnel, à savoir la célèbre objection de Hume selon laquelle il nous est impossible de percevoir la causalité en général. Autrement dit, si le problème d'Intégration pour la causalité agentive ne présente pas la difficulté initiale d'une hétérogénéité des niveaux, elle rencontre tout de même une objection, car elle entre dans le problème plus vaste qui est celui de l'Intégration de la causalité en général. Cependant,

⁴¹⁸ Voir chapitre 4, p.310.

d'un point de vue dialectique, il n'y a sans doute pas lieu d'être exagérément inquiet de cette objection si l'on considère qu'il s'agit d'une objection très générale contre la connaissance de toute forme de causalité, et que cette objection générale est rejetée aujourd'hui par de nombreux auteurs : ainsi, pour Helen Beebe, « l'argument de Hume [...] dans une perspective contemporaine, n'est pas du tout convaincant » (Beebe 2009, sec. 2). Nous pourrions donc nous appuyer sur les réponses apportées à l'argument général de Hume pour défendre la connaissabilité de la causalité agentive en particulier. Ce type de démarche est également suggéré par O'Connor (1995, 199), qui cite Madden et Harré, défenseurs de la perception de la causalité non agentive, pour défendre par analogie la perception de la causalité agentive. Cette stratégie peut se prévaloir d'une impressionnante liste d'auteurs qui, à différentes époques jusqu'à ces dernières années, ont contesté la *doxa* humienne. Parmi les défenseurs de l'observabilité de la causalité, on peut mentionner notamment : John Locke (1690 II, ch. 21, §4), Thomas Reid (1788, 36), William James (1912, pt. VI), A. N. Whitehead (1923, pt. II), C. J. Ducasse (1926), Elisabeth Anscombe (1971), Harré et Madden (1975, 49–67) ; et dans les trente dernières années : Evan Fales (1990, chap. 1), Nancy Cartwright (1993, 426), David Armstrong (1962, 23; 1968, 96–98; 1999, 211–216), David Lewis (2004, 75–76), Helen Beebe (2003; 2009), ou encore Stephen Mumford (2011, chap. 9).⁴¹⁹

Cependant, on ne peut pas se contenter de constater ce mouvement général de confiance dans l'observabilité de la causalité. Il faut aussi pouvoir l'appliquer au cas particulier de la causalité agentive. Or le problème est que ces différents auteurs défendent des conceptions de la causalité et des modèles de perception radicalement différents. Par exemple, l'emploi analogique qu'O'Connor fait de la théorie de Harré et Madden n'est pas sans poser de problèmes, dans la mesure où ces auteurs défendent la perception des pouvoirs causaux *non agentifs* en rejetant *explicitement* une épistémologie des pouvoirs causaux agentifs. En répondant à l'objection de Hume, il conviendra donc de distinguer plusieurs niveaux de réponse. Certains auteurs (Lewis, Beebe) défendent qu'on peut percevoir la causalité, mais en gardant une conception humienne (régulariste) de la causalité. Ceci est loin d'être suffisant pour notre projet : il nous faudra défendre en effet la perceptibilité de la causalité dans une conception *singulariste* et non régulariste (voir Anscombe, Fales, Armstrong, Ducasse), et plus précisément dans une conception *néo-aristotélicienne* de la causalité (voir Harré et Madden, et Mumford), c'est-à-dire qu'il nous faudra défendre la perceptibilité de la causalité comme exercice de *pouvoirs causaux*. Enfin, soutenir que l'on perçoit l'*exercice* d'un pouvoir causal, ce n'est pas encore dire que l'on perçoit le *pouvoir* causal lui-même en amont de son exercice, ni que l'on perçoit le caractère *indéterministe* de ce pouvoir causal (ou la concurrence indéterministe de deux pouvoirs causaux contraires). On peut donc distinguer cinq niveaux de plus en plus précis dans le projet d'épistémologie du libre arbitre agentif : il faut défendre (i) la perception de la causalité, (ii) d'une causalité *singulariste*, (iii) d'une causalité singulariste par *pouvoirs causaux*, (iv) la perception des *pouvoirs* eux-mêmes et pas seulement de leur *exercice*, (v) enfin la perception de pouvoirs causaux *non déterministes*. Parmi les auteurs mentionnés plus haut, bien peu nous accompagneront jusqu'au dernier niveau. À ma connaissance, seul Thomas Reid a vraiment apporté une épistémologie complète du libertarianisme agentif, et comme nous le verrons, sa solution comporte des conséquences métaphysiques difficiles à admettre. Il nous faudra donc construire une solution nouvelle.

⁴¹⁹ Il est à noter que c'est surtout dans ces trente dernières années que la *doxa* humienne a réellement chancelé ; en 1993, Goldman pouvait encore la considérer comme une évidence consensuelle : « Comme nous le savons depuis Hume, les connections causales entre événements ne peuvent pas être directement observées ; et elles ne peuvent pas non plus être connues par introspection » (1993, sec. 1).

Repartons de l'objection de Hume : pourquoi serait-il impossible de percevoir la causalité ? L'objection de Hume comporte deux parties : il s'attache à montrer que nous ne pourrions tirer une expérience de la causalité (i) ni de l'observation d'événements extérieurs causant d'autres événements, (ii) ni de l'introspection de notre propre activité par laquelle nous produisons certains effets par notre volonté. Ces deux sources étant les seules candidates pour apporter une perception de la causalité, Hume en conclut qu'une telle perception est absolument impossible.

En ce qui concerne l'impossibilité d'observer une causalité extérieure, la conviction de Hume semble reposer sur deux éléments distincts (voir Beebe 2009, sec. 2), premièrement une conviction phénoménologique, deuxièmement un argument d'imprédictibilité. L'argument phénoménologique est le suivant :

Quand nous regardons hors de nous vers les objets extérieurs et que nous considérons l'opération des causes, nous ne sommes jamais capables, dans un seul cas, de découvrir un pouvoir ou une connexion nécessaire, une qualité qui lie l'effet à la cause et fait de l'un la conséquence infaillible de l'autre. Nous trouvons seulement que l'un suit l'autre effectivement, en fait. L'impulsion de la première bille de billard s'accompagne du mouvement de la seconde. Voilà tout ce qui apparaît aux sens *externes*.
(Hume 1983, 130)

Si nous n'avons pas du tout d'expérience, pas de phénoménologie, de la causalité alors il est clair que nous ne pouvons pas *percevoir* la causalité – c'est-à-dire la connaître au moyen d'une impression que nous en avons. Une expérience de la causalité ne serait pas suffisante pour qu'on puisse parler de perception – puisqu'une perception est une expérience *véridique* et *fiable* – mais elle est clairement *nécessaire*. Le deuxième argument est l'argument d'imprédictibilité :

A la première apparition d'un objet, nous ne pouvons jamais conjecturer quel effet en résultera. Mais si le pouvoir ou l'énergie d'une cause pouvait se découvrir par l'esprit, nous pourrions prévoir l'effet, même sans expérience, et nous pourrions, dès l'abord, nous prononcer avec certitude à son sujet, par la seule force de la pensée et du raisonnement. (*ibid.*)

Ce deuxième argument est indépendant du premier : quand bien même nous aurions une impression intermédiaire entre la cause et l'effet, par exemple une impression sensible de « force », nous n'avons pas d'impression qui nous permette de *prédire* l'effet avant qu'il se produise, c'est-à-dire que nous n'avons pas d'impression qui soit intrinsèquement tournée vers la réalisation nécessaire d'un effet, comme un pouvoir causal doit être intrinsèquement disposé à produire nécessairement son effet. Donc quand bien même nous trouverions autre chose que la simple juxtaposition de la cause et de l'effet, ce troisième terme ne serait pourtant pas une expérience de la cause, parce qu'il n'offrirait aucune prédictibilité.

La seconde partie de l'objection humienne concerne l'exercice de la volonté. Dans l'exercice de la volonté, Hume estime que l'on ne peut pas avoir non plus une expérience de la cause. Son argument consiste essentiellement à montrer que le cas de la volonté n'est pas sensiblement différent de celui de l'observation de causes extérieures. Pour Hume, observer la causalité de notre volonté consisterait à observer qu'un certain événement de volition a causé un certain événement physique de mouvement corporel. Mais là encore, tout ce que nous observons réellement, c'est chacun de ces événements et leur succession temporelle, et non pas le lien qui permet de passer de l'un à l'autre.

Le mouvement du corps suit le mouvement de la volonté. Nous en avons conscience à tout moment. Mais les moyens qui en permettent la réalisation, l'énergie qui permet à la volonté d'accomplir une opération aussi extraordinaire, nous sommes si éloignés d'en avoir une conscience immédiate qu'ils doivent échapper pour toujours à nos recherches les plus diligentes. (Hume 1983, 132)

L'impossibilité d'observer le lien entre volonté et mouvement corporel est même plus radicale que l'impossibilité d'observer le lien causal entre deux boules de billard. En effet, dans le cas des boules de billard, nous voyons la contiguïté spatiale des deux boules au moment du choc ; nous voyons tous les événements de la « chaîne causale », même si nous ne pouvons pas observer le lien causal lui-même. Dans le cas de la volonté, nous n'observons même pas tous les événements qui se trouvent entre la volonté et le mouvement corporel lui-même : influx nerveux, mouvements de fibres musculaires, etc. qui sont certainement des intermédiaires essentiels du « trajet causal » et échappent totalement à notre observation commune.

Voyons à présent comment on peut répondre aux arguments de Hume pour défendre la perception de la causalité, et plus particulièrement de la causalité agentive libre.

L'élément le plus faible de l'argument de Hume est certainement l'argument phénoménologique, c'est-à-dire l'idée selon laquelle notre expérience du choc de boules de billard, par exemple, ne nous présenterait *que* la contiguïté spatiale et temporelle des deux mouvements (le mouvement de la première boule s'approchant de la seconde, s'arrêtant à son contact, puis le mouvement de la seconde) sans aucune impression supplémentaire de lien causal entre les deux. Hume se prête ici à un exercice de phénoménologie en première personne, et comme le fait remarquer Beebe (2009, sec. 2), le défenseur de l'expérience causale « peut tout simplement refuser [cette affirmation phénoménologique] ». C'est précisément ce que fait Anscombe qui reproche à Hume d'avoir tout bonnement exclu de son enquête ce qui était sous ses yeux : « [Hume], sûr de lui, nous lance le défi de 'montrer un cas dans lequel l'efficacité peut être découverte par l'esprit, et où ses opérations sont évidentes à notre conscience ou à notre sensation'. Rien de plus simple : le fait de couper, de boire, de ronronner, n'est-il pas une 'efficacité' ? » (1971). Pour Anscombe, il est tout simplement évident que nous pouvons *voir* les ciseaux couper une feuille, *voir* le chat en train de boire son lait, ou en train de ronronner. Si nous pouvons voir cela, c'est que notre expérience perceptive est quotidiennement remplie de concepts causaux. En somme, à l'affirmation phénoménologique que Hume avance « dans un fauteuil », Anscombe répond par l'affirmation contraire, ce qui semble dialectiquement légitime. Mais ce qui a définitivement réfuté l'affirmation de Hume, ce sont les efforts pour déterminer *empiriquement* l'existence ou l'absence d'une phénoménologie causale. Les travaux d'Albert Michotte (1954) ont été décisifs sur cette question. Michotte a présenté à des sujets un court film dans lequel un carré vert s'approche d'un carré rouge, s'arrête à son contact, après quoi le carré rouge part dans la direction opposée. Dans la mesure où il s'agit d'un film, il n'y a évidemment aucun lien causal entre le mouvement des deux (images de) carrés. La question était de savoir comment les sujets percevaient cette séquence. Les résultats de Michotte sont très clairs : les sujets ont l'expérience que le carré vert *fait partir* le carré rouge, qu'il *produit son mouvement*. L'expérience devient encore plus claire lorsqu'on la compare à d'autres expériences dans lesquelles Michotte a fait varier la distance entre les deux carrés au moment de l'arrêt du carré vert, et la durée entre l'arrêt du carré vert et le départ du carré rouge : les sujets continuent de rapporter une expérience de causalité lorsque la distance reste faible et que le « temps de réaction » reste court, mais rapportent une simple expérience de coïncidence lorsque la distance et le temps de réaction augmentent sensiblement. Autrement dit, il y a bien une expérience supplémentaire à la simple expérience de coïncidence spatio-temporelle, une

expérience proprement *causale*. L'affirmation phénoménologique de Hume est donc un cas de philosophie dans un fauteuil que l'expérimentation contemporaine semble infirmer. On pourrait être tenté d'objecter que les sujets de Michotte ne rapportaient pas réellement leur *expérience*, mais plutôt des *croyances* étroitement associées à cette expérience. A cela, on peut faire tout d'abord la réponse de David Lewis :

On sait bien qu'il est difficile de tracer la démarcation entre ce qui est vrai d'après l'expérience perceptive à elle seule, et ce qui est vrai d'après un système de croyances formé en partie par l'expérience perceptive et en partie par des croyances préalables. La chaussure s'approche et touche la balle, et immédiatement la balle s'envole vers le but. Est-ce que je vois une chose causer l'autre ? Ou est-ce que je l'infère à partir de ce que je vois, et de ma connaissance d'arrière-plan sur les choses du monde ? Je ne sais pas, et je ne sais pas comment trancher la question. Donc je ne suis pas en mesure de nier que, dans un tel cas, je suis en contact perceptif avec un cas de relation causale.
(Lewis 2004, 74–75)

Lewis, qui est pourtant néo-humien dans sa conception de la causalité, considère donc que nos vécus causaux *pourraient* être de vraies expériences perceptives et pas seulement des croyances. On peut sans doute aller plus loin et trancher positivement dans le sens de l'expérience perceptive si l'on suit le critère « d'encapsulation » rappelé par Beebe (2009, sec. 3.1.). L'idée est la suivante : si une certaine donnée appartient strictement à notre arrière-plan de croyances, alors il sera sensible à la variation de nos croyances ; et inversement, s'il appartient proprement au contenu perceptif, il devrait manifester une certaine indépendance vis-à-vis des variations de croyance. En vertu du caractère « modulaire » de la perception, le contenu proprement perceptif est « encapsulé », ou encore « impénétrable » par les variations de croyances. C'est ce critère qui permet de décrire le cas des lignes de Müller-Lyer comme un cas d'illusion perceptive : même lorsque nous *savons* que les deux lignes ont la même longueur, notre expérience perceptive continue de nous les représenter comme ayant des longueurs différentes. Or il se trouve que les expériences rapportées par Michotte sont justement impénétrables à la croyance dans le sens pertinent : en effet, les sujets de l'expérience *savent* que les carrés à l'écran n'ont aucun effet causal l'un sur l'autre, puisqu'il s'agit purement d'images sur un écran. Mais leur expérience de production causale est indépendante de cette connaissance d'arrière-plan. On peut donc considérer qu'il s'agit bel et bien d'une *expérience* causale et non pas seulement d'une croyance ajoutée par après à l'expérience proprement dite.

L'argument phénoménologique de Hume est donc largement rejeté aujourd'hui : d'un point de vue phénoménologique, il semble bien que nous ayons une phénoménologie causale dans les cas typiques que décrit Hume, les cas d'observation visuelle.

Mais Hume dispose d'un deuxième argument pour rejeter l'expérience causale extérieure, à savoir l'argument de prédictibilité : quand bien même il y aurait une expérience supplémentaire aux expériences purement spatio-temporelles, cette expérience ne saurait être, d'après Hume, une expérience *du pouvoir causal* ou de l'efficacité causale parce qu'elle ne nous permet jamais de *prédire* l'effet. Or si nous avions une connaissance du pouvoir causal, nous devrions pouvoir observer « en lui » l'effet qu'il peut réaliser. On voit immédiatement que ce deuxième argument repose sur une certaine conception de la causalité comme impliquant un *pouvoir causal* – c'est-à-dire sur une conception « néo-aristotélicienne » de la causalité. Ainsi, des néo-humiens comme Beebe ou Lewis pourraient répondre de la manière suivante : cet argument d'imprédictibilité permet effectivement de montrer l'impossibilité de percevoir certains types de causalité, notamment de la causalité néo-aristotélicienne, mais de toutes les façons la causalité qui existe dans le monde est d'un autre type (par exemple,

régulariste), et ce type de causalité, nous pouvons en faire l'expérience. Cette réponse ne serait évidemment pas satisfaisante pour notre projet présent, qui suppose justement une perception des pouvoirs agentifs néo-aristotéliens : elle reviendrait à accepter le niveau (i) de la réponse anti-humienne (la perception de la causalité « en général », ou « dans quelque sens de causalité ») mais en rejetant comme impossibles les autres niveaux (la perception de la causalité néo-aristotélienne, de la causalité agentive, et de la causalité libre).

Pour passer d'une perception de la causalité *en général* (telle qu'elle est défendue par Lewis et Beebe) à une perception de la causalité *néo-aristotélienne*, il faudra donc répondre à l'argument de la prédictibilité. Et pour cela, il semble nécessaire d'abandonner le paradigme de l'observation visuelle, celui de Michotte, Anscombe, Ducasse, ou Beebe, pour prendre en compte la perception *tactile* de la force ou du pouvoir causal, ainsi que le font Fales, Armstrong ou Mumford. Après avoir brièvement expliqué l'intérêt de ce nouveau paradigme par rapport au paradigme visuel, j'exposerai la version qu'en donne Fales en soulignant en quoi elle peut constituer une source de connaissance dans le sens fiabiliste ; enfin, je montrerai comment cette théorie répond à l'objection humienne de prédictibilité.

Pourquoi s'attacher à l'impression tactile de la force (celle que je subis lorsque quelqu'un m'appuie sur le front, par exemple) plutôt qu'à l'impression visuelle (lorsque je vois quelqu'un appuyer sur un objet et le mettre en mouvement) ? Et quel est l'intérêt de ce paradigme pour rendre compte de la conception *néo-aristotélienne* de la causalité ? On peut comprendre cet intérêt en partant d'un argument de Beebe (2009, sec. 3) : on a vu que, pour Beebe, la perception de la causalité était possible même avec une conception régulariste humienne de la causalité, c'est-à-dire avec une conception dans laquelle la vérité de « A cause B » dépend d'un grand nombre de faits éloignés dans le temps et dans l'espace, lors desquels un événement du même type que A a été suivi d'un événement du même type que B. Cette théorie présente une difficulté évidente : comment pourrait-on *percevoir* visuellement une telle régularité alors qu'elle dépend de fait éloignés dans le temps et dans l'espace ? Il est clair que ces faits éloignés ne peuvent pas être représentés dans nos *stimuli visuels* ; seule la succession présente de A et de B sera représentée dans nos *stimuli visuels*. Mais Beebe fait observer que, dans toute théorie plausible de la perception visuelle, le contenu perçu doit dépasser le contenu strictement représenté dans les *stimuli visuels*, autrement nous ne pourrions jamais percevoir un dé, mais seulement ses faces visibles, ni un chat, mais seulement le profil qu'il nous présente, etc. Or c'est bien un dé ou un chat que nous percevons visuellement, et non pas des faces de dé et un profil de chat. Le fait que nos stimuli visuels ne nous présentent pas *tout* le dé signifie simplement que nous ne sommes pas en *contact direct* (*acquaintance*) avec le dé dans son entier, mais cela ne signifie pas que nous ne le *percevons* pas. De même, le fait que nos *stimuli visuels* ne nous présentent pas tous les faits de successions qui fondent la régularité causale ne veut pas dire que nous ne *percevons* pas la régularité causale (du moment que celle-ci fait bien partie de notre contenu phénoménal) ; cela veut seulement dire que nous ne sommes pas en *contact direct* avec cette régularité causale. Autrement dit, la perception visuelle de la causalité reste compatible avec une théorie régulariste parce qu'elle n'implique pas un *contact direct* avec la causalité. Mais la situation est différente, justement, pour la perception *tactile* de la causalité : en effet, comme le fait remarquer Beebe (2009, sec. 4.2), « il peut sembler plus évident dans ce cas que la force ou la pression est un objet de 'conscience directe' et non pas *simplement* le contenu de l'expérience, entendu au sens large ; et les Humiens [...] doivent, bien sûr, refuser cela. » Si nous pouvons avoir, ici et maintenant, un *contact direct* avec un fait de causalité, c'est qu'un fait de causalité n'est pas constitué de régularités dispersées dans l'espace et dans le temps, mais qu'il s'agit d'un fait singulier, intrinsèquement réalisé ici et maintenant. Autrement dit, la perception tactile de la causalité, à la différence de la perception visuelle, est indissociable

d'une conception *singulariste* de la causalité. Dans la mesure où la conception néo-aristotélicienne de la causalité appartient à la famille des conceptions singularistes, il est donc clair que le meilleur paradigme pour en donner une épistémologie sera la perception tactile.⁴²⁰

L'expérience tactile de la force ou de la pression a été admirablement décrite par Evan Fales (1990, chap. 1 §4).⁴²¹ Fales commence par donner une description phénoménologique du contenu de cette expérience. Lorsque quelqu'un appuie de sa main sur mon front, j'ai une expérience de force appliquée à mon front qui a, au moins, les six caractéristiques suivantes. (1) Cette expérience a une *localisation spatiale* : je ressens cette force *sur mon front*. (2) Elle a une *intensité (magnitude)* : selon que la personne appuie plus ou moins sur mon front, je ressens une pression plus ou moins forte. (3) Elle a une *direction dans l'espace* : la personne appuie soit vers l'arrière, soit vers la gauche, soit vers la droite. (4) Elle peut être *distinguée* d'autres sensations de force : par exemple, si deux personnes appuient en même temps sur mon front, l'une vers la droite, et l'autre vers la gauche, je peux distinguer ces deux expériences de force. (5) Les différentes forces ressenties donnent lieu à une *algèbre*, c'est-à-dire qu'elles peuvent s'additionner les unes aux autres en fonction de leur direction et de leur intensité : dans l'exemple précédent où deux personnes m'appuient en même temps sur le front, j'aurai par exemple l'expérience que la force vers la droite et la force vers la gauche s'annulent parce qu'elles ont des directions opposées et la même intensité. Ces cinq premières caractéristiques sont celles que Fales appellent les caractéristiques « vectorielles » de la sensation de force. (6) Une sixième caractéristique est indispensable pour rendre compte de l'*asymétrie* de l'expérience causale ; en effet, la relation cause/effet est fondamentalement une relation asymétrique, or les caractéristiques vectorielles mentionnées jusqu'à présent ne permettent pas d'en rendre compte : pour un même « vecteur » de force, on pourrait interpréter que la pression est causée par le front ou bien par la main. Pour cette dernière caractéristique, Fales parle de l'expérience de *production* : j'ai l'expérience que c'est la main qui « produit » ou cause cette pression (et non pas mon front). Dans l'expérience de la pression, nous avons donc une expérience phénoménale de la force causale ou du pouvoir causal. Mais cela ne suffit pas encore pour parler de *perception* : l'expérience phénoménale pourrait être illusoire ou véridique par pur hasard. Fales (1990, 12) souligne que les caractéristiques vectorielles de notre expérience correspondent *exactement* aux caractéristiques des vecteurs de force tels qu'ils sont décrits par la physique : un vecteur de force a (1) un point d'application, (2) une intensité, (3) une direction, (4/5) et peut se composer avec d'autres vecteurs de force dans un calcul des forces en présence. Cette étroite

⁴²⁰ De manière étonnante, Harré et Madden, défenseurs de la causalité néo-aristotélicienne, en ont donné une épistémologie fondée sur la perception visuelle (1975, 49–67). Il semble y avoir plusieurs raisons à cela. Tout d'abord, la théorie de la perception tactile des forces qui s'appliquent sur mon corps est étroitement liée, comme on le verra plus loin, à la théorie de la perception de *ma* force par laquelle je résiste à cette pression extérieure. Or Harré et Madden accordent beaucoup d'importance à *critiquer* toute théorie qui ferait des pouvoirs de l'agent (expérimentés en première personne) un paradigme des pouvoirs causaux. Pour eux, donner la première place à l'expérience des pouvoirs de l'agent, c'est s'engager à une théorie inférentialiste des pouvoirs extérieurs (nous attribuons des pouvoirs causaux aux choses extérieures par analogie avec notre pouvoir agentif) et cette théorie mène à un panpsychisme inacceptable. Pour Harré et Madden, il est donc très important de défendre que l'observation des causes peut être une observation *strictement extérieure*. Deuxièmement, Harré et Madden n'ont pas perçu l'intérêt épistémologique de s'intéresser à la perception tactile parce qu'ils ne semblent pas avoir réfléchi à la question de fiabilité externaliste. Lorsqu'ils défendent positivement les mérites épistémiques de notre expérience causale, leurs remarques sont d'ordre strictement *justificatif* ou internaliste : « our direct experience of the action of causal powers is *prima facie* perfectly good evidence itself for what it reveals about our world unless there is counter-evidence » (Harré et Madden 1975, 56). Rien n'est proposé pour expliquer comment l'occurrence de telle expérience visuelle pourrait dépendre fiablement de l'occurrence de tel pouvoir causal. Or, c'est de ce point de vue externaliste que le paradigme tactile présente des avantages évidents.

⁴²¹ Armstrong, qui avait brièvement esquissé une telle analyse dans (Armstrong 1968), a par la suite repris largement les résultats de Fales (Armstrong 1999, sec. 14.6).

corrélation entre la chose extérieure (la force physique réelle) et les données de notre expérience phénoménale suggère fortement que la seconde *dépend* de la première et qu'il ne s'agit pas simplement d'une coïncidence hasardeuse. Fales s'en tient là pour défendre les mérites épistémiques de notre expérience, mais je crois qu'on peut aller plus loin lorsqu'on pose la question dans des termes explicitement fiabilistes.⁴²²

On peut aisément donner un modèle intelligible dans lequel notre expérience de la force dépend fiablement du pouvoir causal qui est perçu. Dans l'exemple de la pression sur le front, supposons que je résiste volontairement à cette pression mais que j'expérimente dans cette pression un pouvoir *de déplacer ma tête vers l'arrière* (pouvoir de la main qui m'appuie sur le front). Ce pouvoir n'est pas actuellement exercé parce que j'applique une force dans le sens contraire pour maintenir ma tête, et donc ma tête ne se déplace pas actuellement vers l'arrière. Comment donc puis-je percevoir ce pouvoir causal, alors qu'il n'est même pas exercé ? La réponse est la suivante : la même propriété P qui constitue un pouvoir causal *de déplacer ma tête vers l'arrière*, constitue également le pouvoir causal *d'exercer une pression sur mon front*, et c'est cette propriété P qui, actuellement, cause directement la pression physique sur mon front puis indirectement une impression sensible dans mon cerveau. Donc mon impression d'un pouvoir causal de déplacer ma tête est causée par la propriété P qui constitue réellement ce pouvoir. En termes contrefactuels, on peut dire que *si* la main n'avait pas le pouvoir causal de déplacer ma tête vers l'arrière, *alors* c'est que cette main n'aurait pas la propriété P, et dans ce cas elle ne pourrait pas non plus exercer de pression sur mon front, ce qui veut dire que je n'aurais pas l'impression phénoménologique que j'ai (l'impression d'un pouvoir de déplacer ma tête vers l'arrière). Donc, on peut dire plus directement que

Si la main n'avait pas le pouvoir causal de déplacer ma tête vers l'arrière, je n'aurais pas l'impression d'un pouvoir de déplacer ma tête vers l'arrière.

Mon impression de pouvoir causal est donc *sensible* à son objet, bien que ce pouvoir causal ne soit pas actuellement exercé.

Voyons enfin comment la perception tactile des pouvoirs causaux permet de répondre à l'objection humienne de l'imprédictibilité. Hume considère que (i) si nous pouvions faire l'expérience d'un pouvoir causal ou d'un lien de connexion nécessaire, nous pourrions prédire avec certitude l'effet qui s'ensuivra, sans avoir à attendre d'observer l'effet lui-même. Or (ii) une telle prédiction est, d'après lui, impossible. A la suite de Fales, je défendrai que les deux prémisses de cet argument sont problématiques. On va voir que (i) l'observation d'un pouvoir causal ne suppose pas une prédiction *certaine* ou nécessaire, et que (ii) une certaine forme de « prédiction » (de connaissance de l'effet antécédente à sa manifestation) est bel et bien possible.

Le premier problème de l'argument de Hume est le présupposé qu'un pouvoir causal doit entraîner son effet avec *nécessité*. Ce présupposé est erroné pour deux raisons distinctes. Tout d'abord, comme on l'a vu au chapitre 5 (p.356), le lien causal peut très bien être un lien *indéterministe*. Par exemple, un atome peut très bien avoir le pouvoir causal de se décomposer au temps *t*, et n'être pas déterminé à se décomposer à *t* car les lois de la physique quantique laissent ouvertes les deux options. Dans ce cas, si l'on connaît à *t-1* le pouvoir causal de cet atome, cela ne nous permettra pas pour autant de prédire *avec certitude* l'effet qui s'ensuivra

⁴²² Il est important de noter également que Fales ne défend *pas* une conception néo-aristotélicienne de la causalité : pour lui, le vocabulaire de la force causale doit être ramené, ontologiquement, à des relations nomologiques de second ordre entre propriétés, et non pas à des pouvoirs causaux ontologiquement premiers et irréductibles. Cependant, il est clair que tout ce qu'il dit de la perception tactile des forces peut s'appliquer aussi bien, et même mieux, à une métaphysique dans laquelle les pouvoirs causaux sont le fondement ontologique irréductible des faits de causalité.

(bien qu'on puisse prédire *quels effets* sont possibles). Deuxièmement, il y a une raison plus simple et plus radicale pour laquelle un pouvoir causal n'est pas connecté de manière nécessaire à son effet, c'est qu'un pouvoir causal peut être *empêché* par la présence d'autres pouvoirs causaux (voir Fales 1990, 14). Par exemple, le pouvoir causal de la main, de déplacer ma tête vers l'arrière, peut être empêché ou contrecarré par le pouvoir causal de ma tête, de se maintenir en place en exerçant une force vers l'avant. Même si nous vivions dans un monde purement déterministe (sans mécanique quantique), la connaissance d'un pouvoir causal ne permettrait pas de prédire avec certitude les effets à venir ; c'est seulement la connaissance de *tous* les pouvoirs causaux d'un système qui permettrait, par un calcul de composition des forces, de prédire les effets.⁴²³

Donc la première prémisse de Hume est mal fondée : il se pourrait très bien que nous observions directement des pouvoirs causaux et que nous soyons pourtant incapables de prédire avec certitude et nécessité les effets. Cela veut-il dire qu'il faut abandonner toute idée de connexion logique entre le pouvoir causal et l'effet, et toute idée de « prédiction » ? Non, et l'on peut donc donner une version plus faible de la prémisse de Hume : dans la mesure où un pouvoir causal est un état *intentionnel* (il est le pouvoir *de produire tel ou tel effet*), il existe une forme de « connexion logique » entre un pouvoir et son effet, dans le sens où l'effet est le contenu intentionnel du pouvoir. Si donc nous pouvions observer directement un pouvoir, nous pourrions *nous attendre* à un certain *effet possible* déterminé, c'est-à-dire, nous pourrions nous attendre à ce que le pouvoir soit actualisé (s'il n'est pas empêché et si le hasard indéterministe tourne dans le sens de son actualisation).

Un effet ne s'ensuit pas toujours de sa cause typique [...] La notion de cause n'est pas quelque chose qui entraîne automatiquement et immédiatement la nécessité. [...] Même les vecteurs résultants ne représentent jamais plus qu'une tendance vers un certain effet. Elle peut bien produire cet effet, et de fait *on serait justifié à s'attendre à ce qu'elle le fasse*. Mais la connexion entre une configuration causale et son effet est irréductiblement dispositionnelle. (Mumford and Anjum 2011, 175, italiques ajoutés)

Est-ce le cas ? L'observation d'un pouvoir causal nous permet-elle de *nous attendre* à un certain effet déterminé, avant l'actualisation de cet effet ? D'après Hume, non :

La position de Hume est assez claire. Il pense que, pour autant que nous puissions le savoir par expérience, *n'importe quel* événement imaginable pourrait s'ensuivre, ou ne pas s'ensuivre, de l'expérience de force. Par conséquent, toute restriction significative du champ de possibilités comptera comme une prédiction pertinente. (Fales 1990, 18)

Après avoir rappelé qu'il serait difficile, expérimentalement, de mettre en évidence une telle prédiction, Fales montre que nous avons toutes les raisons d'admettre la possibilité de principe d'une telle prédiction. En effet, l'expérience de la force contient en elle-même, c'est-à-dire dans son contenu tel que nous l'avons décrit plus haut, tout ce qui est nécessaire pour restreindre le champ de nos attentes à un effet déterminé, indépendamment de l'observation effective de cet effet. Dans l'exemple de la pression sur le front, l'expérience de pression permet de « prédire » que, si je relâchais mon effort de résistance, (1) c'est ma tête qui se

⁴²³ C'est dans ce second sens que Mumford et Anjum (2011, sec. 8.2) parlent de la causalité (ou de la « dispositionnalité » dans leur vocabulaire) comme « en deçà de la nécessité » : ce qu'ils ont à l'esprit, ce n'est pas qu'une cause pourrait causer ou ne pas causer son effet *dans les mêmes circonstances* (donc de manière strictement indéterministe), mais qu'un pouvoir causal « peut toujours être empêché, masqué ou bloqué » (*idem*, p.177) par l'ajout d'un mécanisme de blocage dont les pouvoirs causaux seraient opposés à l'exercice du premier pouvoir.

déplacerait, (2) elle se déplacerait avec une certaine vitesse, (3) elle se déplacerait vers l'arrière. Pourquoi cette expérience de pression permet-elle de prédire un tel mouvement alors qu'elle n'est pas connectée à ce mouvement de manière *nécessaire* ? Parce que son contenu, en particulier les trois premières caractéristiques vectorielles de (1) localisation, (2) intensité et (3) direction, correspondent exactement aux caractéristiques vectorielles de ce mouvement déterminé. Il n'est donc pas nécessaire que ce mouvement ait effectivement eu lieu et ait été observé pour qu'il soit une possibilité saillante à laquelle nous nous attendons : le contenu de notre expérience de la force suffit à nous attendre (sans nécessité) à ce mouvement déterminé.

Pour résumer, l'expérience tactile de la pression ou de la force permet de défendre la perception de la causalité dans un sens beaucoup plus riche que l'expérience visuelle. En effet, elle permet de défendre une connaissance de la causalité dans un sens *singulariste* de la causalité, et plus précisément dans le sens *néo-aristotélicien* des pouvoirs causaux. En outre, elle permet de défendre la perception des pouvoirs causaux même lorsque ceux-ci sont *non actualisés* et même lorsqu'ils sont des pouvoirs menant de manière *indéterministe* à leur effet. Ce modèle épistémologique est donc très proche de ce dont nous avons besoin pour défendre la connaissabilité du pouvoir d'agir autrement de manière indéterministe. Malheureusement, il manque encore un élément essentiel, à savoir le caractère « agentif » du pouvoir causal : il n'est pas évident du tout que nous puissions observer nos propres pouvoirs de la même façon que nous observons les pouvoirs qui s'exercent sur notre corps.

Pour résoudre notre problème d'intégration du libre arbitre, il faut donc aborder un niveau supplémentaire de l'analyse, à savoir une connaissance causale qui n'est plus seulement la connaissance de pouvoirs causaux néo-aristotéliens, mais surtout de pouvoirs causaux *agentifs*. En discutant du pouvoir causal de l'agent, et de sa perceptibilité, nous répondrons évidemment à la dernière objection de Hume, celle selon laquelle les pouvoirs causaux ne peuvent pas être observés dans l'activité volontaire. A ce niveau d'analyse, nous ne pouvons plus nous appuyer sur Harré et Madden, qui refusent explicitement une épistémologie des pouvoirs *agentifs*, ni sur Armstrong ou sur Fales, qui ne mentionnent le cas de l'agentivité qu'en passant, et sans l'étudier. Je n'ai trouvé que deux auteurs qui aient essayé de donner une épistémologie des pouvoirs agentifs : Stephen Mumford et Thomas Reid. Je partirai de l'épistémologie de Mumford, dont on va voir qu'elle ne permet pas d'aller jusqu'au bout de l'analyse, c'est-à-dire jusqu'à une épistémologie de l'agentivité *libre* ; mais elle offre un bon point de départ pour poser le problème et tenter de le résoudre.

L'épistémologie de Mumford s'appuie sur une intuition dont on trouve l'indication chez Armstrong et chez Fales, l'intuition d'un lien étroit ou d'une corrélation entre l'expérience passive d'une force qui agit sur moi, et l'expérience active par laquelle je mets en oeuvre ma propre force.

Nous sommes capables de distinguer dans la perception entre l'agentivité active de notre côté, et la réception passive d'une force, entre une force qui s'applique sur nos corps et la résistance de nos corps. [...] La sensation de force est identifiée tout d'abord, dans la mesure où nous sommes passifs, avec le sentiment de la pression ; et dans la mesure où nous sommes actifs, avec la tension musculaire.

(Fales 1990, 12–13)

La force subie et la force exercée par l'agent sont donc étroitement corrélées du point de vue conceptuel, dans le sens où elles sont en quelque sorte l'inverse l'une de l'autre. Malheureusement, Fales s'en tient à cette correspondance conceptuelle, et a cru pouvoir approfondir la description de l'une (la force subie) en laissant de côté l'autre (la force agentive). Mais Mumford a montré qu'en réalité la force subie et la force agentive ne

pouvaient pas réellement être séparées l'une de l'autre du point de vue réel. Lorsque quelqu'un appuie son doigt sur mon front, mon expérience passive de la pression sur mon front (et du pouvoir de mouvoir ma tête vers l'arrière) est inséparable de mon expérience active de résister à cette pression (par l'effort musculaire de mon cou) pour maintenir ma tête. Mumford prend encore deux autres exemples :

Vous marchez par un temps de vent violent, par exemple, et vous pouvez ressentir ce que le pouvoir du vent pourrait vous faire. Par proprioception, vous sentez que vous devez résister contre le vent et vous pencher, sinon il pourrait vous faire tomber. Mais par la même expérience, vous sentez que vous pouvez y résister en plaçant adéquatement votre masse corporelle et en ajustant la tension dans certains muscles. [...] Vous êtes le patient dans ce cas – vous subissez un changement – mais vous pouvez aussi être actif et y résister, en maintenant l'équilibre des forces.

Dans d'autres cas vous êtes plus agent que patient. Vous voulez soulever un lourd fardeau, par exemple, en exerçant le pouvoir de vos muscles. Par proprioception, vous sentez vos muscles à l'oeuvre, et vous sentez l'effort qui est requis pour réussir la tâche. Mais vous sentez aussi que le poids résiste à votre action. [...]

Ces deux exemples montrent que la distinction agent-patient n'est pas si tranchée. La plupart des interactions causales auxquelles nous participons se trouvent sur une échelle du plus ou moins actif ou passif. Les deux exemples contiennent un certain mélange. (Mumford and Anjum 2011, 210)

Ainsi, ce que nous percevons, ce n'est jamais simplement une force externe qui s'applique sur nous, ou bien une force agentive que nous exerçons, c'est plutôt un certain équilibre de forces, passive et active, qui s'exercent l'une contre l'autre ou se résistent l'une l'autre.

Cette corrélation étroite entre force subie et force agentive permet de transposer au pouvoir agentif les éléments d'épistémologie que nous avons vus pour l'expérience de la force externe. De même que la force externe est perçue au moyen d'une sensation particulière (la sensation de pression), de même il existe une sensation qui permet de percevoir la force agentive en exercice (la proprioception de l'effort musculaire). De même que la sensation de pression était (1) localisée spatialement, (2) d'une certaine intensité, (3) dotée d'une direction dans l'espace, (4) distincte d'autres forces et (5) composable avec d'autres forces au sein d'une « algèbre », de même, l'expérience de l'effort musculaire est localisée dans un certain muscle, d'une certaine intensité, dotée d'une direction dans l'espace, et l'on peut bien sûr la distinguer d'autres forces exercées par l'agent, mais surtout des forces qui s'appliquent sur lui, et avec lesquelles s'opère une forme de composition « algébrique », lorsqu'une force agentive résiste à une force externe. Pour les pouvoirs causaux externes, nous avons vu que l'expérience de la pression pouvait être décrite comme l'expérience d'une *potentialité de mouvement* – un mouvement qui aurait justement les cinq caractéristiques vectorielles que nous venons de rappeler – et donc qu'il y avait bel et bien l'expérience d'un *pouvoir* comme tel, qui nous permet de « nous attendre » à un certain mouvement, même lorsque celui-ci n'a pas lieu parce que le pouvoir en question est contrecarré par un autre pouvoir. On peut dire exactement la même chose pour les pouvoirs agentifs : les caractéristiques vectorielles de l'effort musculaire permettent de représenter un certain « mouvement possible », même si ce mouvement n'a pas lieu parce que la force agentive exercée est contrebalancée par une force externe. Par exemple, lorsque quelqu'un appuie sur mon front, je sens non seulement que la pression exercée sur mon front est un pouvoir « de mouvoir ma tête vers l'arrière » si je n'y résistais pas, mais je sens aussi que la force de résistance que j'exerce avec mon cou est un pouvoir qui entraînerait le mouvement de ma tête vers l'avant si l'on cessait brutalement d'exercer une pression contraire sur mon front. Cette intentionalité du pouvoir causal, cette

orientation vers un certain résultat, qui peut être connue même lorsque le résultat est empêché, est ce que Mumford appelle la modalité de « dispositionalité ». Cette caractéristique de dispositionalité se retrouve donc aussi bien dans les pouvoirs externes que dans les pouvoirs agentifs, ainsi qu'il le souligne au moyen des deux exemples vus précédemment :

Il y a une disposition définie dans l'action du vent, orientée vers un certain résultat, mais ce résultat n'est pas nécessaire parce que vous avez la possibilité d'y résister. [...] [Lorsque vous vous efforcez de soulever un lourd fardeau], il y a une disposition définie orientée vers un certain résultat, et votre action est une tentative pour accomplir ce résultat, mais il est clair qu'elle peut être contrecarrée et cela montre que la modalité dans ces cas de causalité agentive est une modalité dispositionnelle. (Mumford and Anjum 2011, 210)

Cette « dispositionalité » permet d'apporter une réponse à l'argument de Hume contre l'observation interne de la causalité. On a vu que Hume appliquait au cas de l'action volontaire la même objection d'imprédictibilité qu'il avait soulevée contre l'observation de la causalité externe : on ne peut pas plus *prédire* le mouvement corporel à partir de la seule volonté antérieure qu'on ne peut prédire le mouvement résultant d'une boule de billard à partir du mouvement d'une autre boule qui est censé en être la cause. Pourtant nous avons vu que, dans le cas de la causalité externe, l'expérience de la pression permettait de « s'attendre à » un certain mouvement vers lequel le pouvoir causal est disposé, même si ce mouvement n'est pas (ou pas encore) réalisé ; de même, dans le cas de la causalité agentive, l'expérience de l'effort musculaire permet de se représenter un résultat, même si ce résultat est (pour l'instant) empêché.⁴²⁴

Mumford nous apporte donc une théorie intelligible de la perception du pouvoir causal de l'agent, à travers l'expérience de l'effort. Cette théorie permet clairement d'aller plus loin qu'une simple perception de la causalité en général, ou qu'une perception des pouvoirs causaux en général, mais malheureusement, elle ne permet pas d'aller jusqu'au bout de l'analyse dont nous avons besoin pour l'intégration du libertarianisme. En effet, les pouvoirs agentifs tels que les décrit Mumford ne sont pas des pouvoirs *libres* dans un sens incompatibiliste ; ils sont entièrement compatibles avec un fonctionnement entièrement déterministe de l'action humaine. Certes, Mumford souligne que la modalité de « dispositionalité » n'est pas la modalité de « nécessité » et qu'un pouvoir causal ne *détermine* pas la réalisation du résultat auquel il dispose. Mais on a vu plus haut le sens qu'il fallait donner à cette absence de détermination : si le pouvoir de mouvoir la tête vers l'arrière ne détermine pas le mouvement de la tête, c'est seulement parce qu'il peut être contrecarré par la présence d'un pouvoir contraire. Mais lorsqu'on connaît l'ensemble des pouvoirs causaux en présence, cela n'exclut pas qu'il y ait un déterminisme complet du résultat par l'état antérieur du système. Il ne suffit pas de parler de causalité agentive, ou de pouvoirs causaux agentifs, pour avoir un concept libertarien de l'action. Il existe en effet des versions compatibilistes de la théorie de la causalité agentive.⁴²⁵ Ce qu'il nous faut, c'est offrir une épistémologie pour une version proprement incompatibiliste de la causalité agentive. Quelle est la spécificité d'une théorie indéterministe des pouvoirs agentifs ? C'est évidemment le fait que l'agent n'est pas déterminé à exercer son pouvoir agentif. Par exemple, lorsque je marche contre un vent violent, j'ai le pouvoir de résister au vent en me penchant, mais je peux aussi bien exercer ce pouvoir que ne pas l'exercer. Ou pour le dire autrement, j'ai à la fois le pouvoir de résister au

⁴²⁴ En réalité, la réponse de Mumford à Hume est plus complexe que cela, et met l'accent sur le fait qu'on ne peut pas séparer *temporellement* la cause de l'effet (Mumford and Anjum 2011, sec. 9.6). Mais cet aspect de la réponse de Mumford ne me semble pas pertinente pour mon propos.

⁴²⁵ Voir en particulier (Markosian 1999; Markosian 2012), (Nelkin 2011) et (Pereboom à paraître).

vent, et le pouvoir de ne pas y résister, et je n'exercerai qu'un seul de ces deux pouvoirs ; l'autre restera non-exercé.

La grande difficulté, pour une épistémologie de la causalité agentive libertarienne, sera de rendre compte de la connaissance de ce pouvoir non-exercé. Dans la théorie de Mumford, les pouvoirs agentifs sont perçus *lorsqu'ils sont exercés* : mon pouvoir de résister au vent s'exerce à travers mon effort musculaire, et je perçois cet effort musculaire ; mon pouvoir de soulever le fardeau s'exerce dans les muscles de mon bras, et je perçois l'effort de ces muscles (*même* si finalement le pouvoir de lever le fardeau est empêché ou contrecarré par la masse trop importante du fardeau, mon pouvoir, bien qu'empêché, est en exercice dans l'effort musculaire, et c'est cet exercice que je perçois). Percevoir un pouvoir causal *en exercice* ne pose pas de grande difficulté théorique ; en revanche, percevoir un pouvoir causal *qui n'est pas exercé* présente clairement un problème. Le problème est qu'un pouvoir non-exercé est quelque chose de simplement potentiel, de non actuel, et qu'une connaissance perceptive ne semble pas possible pour quelque chose de non actuel. Ce problème a été aperçu par plusieurs philosophes, notamment par J.S. Mill :

La conscience me dit ce que je fais ou ressens. Mais ce que je suis capable de faire n'est pas un objet de conscience. [...] Nous sommes conscients de ce qui est, pas de ce qui sera ou peut être. Nous ne savons jamais que nous sommes capables de faire quelque chose, si ce n'est parce que nous l'avons fait, ou fait quelque chose d'égal et de similaire. (Mill 1865, chap. 26, p. 50)

On peut faire observer que ce problème n'apparaît pas uniquement pour les conceptions libertariennes du libre arbitre : en effet, Christopher Peacocke rencontre le même problème à propos de son concept compatibiliste, parce que son concept de libre arbitre cherche à préserver la notion modale des alternatives possibles. Or c'est bien l'aspect modal des alternatives qui entre en conflit avec une théorie causale de la perception :

Il y a un obstacle initial qui doit être levé pour donner une théorie causale [...] de la connaissance de la liberté sous une forme aussi directe que celle que l'on trouve dans les théories causales de la perception. L'existence apparente de possibilités d'action alternatives – l'impression de la liberté – est un phénomène actuel. Or seul ce qui est actuel peut causalement expliquer ce qui est actuel. (Peacocke 1999, 336–337)

De façon étonnante, Peacocke ne résout pas le problème. Il exprime même des doutes sur la possibilité de le résoudre de manière pleinement satisfaisante et de donner une épistémologie pour notre expérience commune de la liberté. Il semble au contraire pencher dans le sens d'une position « empiriste compatibiliste » : les conditions de vérité de l'expérience du libre arbitre seraient compatibles avec le déterminisme, mais néanmoins trop exigeantes pour être connues réellement autrement que par une investigation scientifique.⁴²⁶ La seule indication

⁴²⁶ « I doubt that the case for or against our actual freedom in any given area—as opposed to its possibility—can be decided simply on the basis of obvious truths and philosophical theory, however resourceful. Freedom is neither a mere practical commitment, nor something incoherent and unachievable. To suppose that we can establish or refute claims about our freedom in a given area from a combination of obvious truths and *a priori* theory is to fail to do justice to the strength of the notion of freedom. » (Peacocke 1999, 339) On pourrait se poser la question : si Peacocke estime que l'intégration épistémique est impossible pour le libre arbitre, pourquoi a-t-il recours à un « ajustement » compatibiliste de la notion de libre arbitre ? Pourquoi ne pas en rester à une notion incompatibiliste ? Apparemment, Peacocke reproche au libertarianisme non pas tant d'être inconnaissable que d'être *inintelligible*. Ce qu'il veut assurer en définitive, c'est la possibilité de connaître l'existence du libre arbitre, d'une manière ou d'une autre, fût-ce scientifiquement, tandis qu'un concept *inintelligible* ne peut bien sûr être instantié ni *a fortiori* connu : « [My compatibilist] account of freedom seems to be an account of the

qu'il donne pour une éventuelle résolution du « problème de la perception de pouvoir non actuel », c'est un bref argument analogique avec la théorie de la perception de la causalité telle qu'elle a été défendue à partir de Michotte – c'est-à-dire la stratégie générale que nous sommes en train de poursuivre dans cette section.⁴²⁷ Peacocke ne pourra donc pas nous apporter beaucoup d'aide pour résoudre ce problème pour une théorie libertarienne.

Le seul auteur, à ma connaissance, qui ait essayé de résoudre réellement le problème de la perception du pouvoir non actuel, et ce dans le cadre d'une théorie libertarienne, est Thomas Reid. Bien avant Mill, Reid avait formulé le problème :

La conscience est ce pouvoir de l'esprit par lequel il a une connaissance immédiate de ses propres opérations. Le pouvoir n'est pas une opération de l'esprit, et donc n'est pas un objet de conscience. (Reid 1788, 8)

A ce problème bien formulé, Reid a apporté une solution que nous allons voir à l'instant. Malheureusement, cette solution a des conséquences métaphysiques que bien peu de philosophes seraient disposés à accepter aujourd'hui. C'est pourquoi je m'efforcerai de chercher d'autres solutions. Une deuxième solution sera construite à partir de la théorie de l'effort de C.A. Campbell. La solution de Reid et la solution (inspirée) de Campbell auront en commun de ne pas remettre en question le principe fondamental qui est à l'origine du problème, à savoir que seul l'*exercice* du pouvoir est susceptible d'être perçu, et que notre connaissance du libre arbitre repose donc, ultimement, sur la perception d'un pouvoir en exercice. En définitive, je défendrai pourtant une troisième solution, qui refuse ce principe, et repose sur la perception de pouvoirs *duaux hors exercice*.

Commençons par l'épistémologie reidienne du libre arbitre. Comment connaissons-nous notre *pouvoir* agentif, et en particulier notre pouvoir *des alternatives*, si ce pouvoir ne peut être aucunement « perçu » ni connu par « conscience », si « nous ne sommes conscients que de l'opération [de l'esprit], et [que] le pouvoir reste derrière la scène » (*ibid.*) ? Puisque Reid insiste clairement sur le fait que « le pouvoir n'est l'objet d'aucun de nos sens extérieurs, ni même un objet de conscience », on pourrait s'attendre à ce qu'il donne une théorie rationaliste, ou *a priori*, de l'origine du concept de pouvoir. Et pourtant Reid veut garder une théorie empiriste dans laquelle :

la notion même, ou l'idée, de pouvoir actif et de cause efficiente est dérivée de l'exercice volontaire par lequel nous produisons des effets ; et si nous n'étions pas conscients d'un tel exercice, nous n'aurions aucune notion de cause ou de pouvoir actif. (Reid 1788, 204)

Comment comprendre ce mystère : le pouvoir ne peut être connu par conscience (ou perception) de l'action, mais il est néanmoins *dérivé* de la conscience de l'action ? On peut comprendre ce que Reid a à l'esprit grâce à ce qui suit le passage cité un peu plus haut :

Nous ne sommes conscients que de l'opération [de l'esprit], et le pouvoir reste derrière la scène ; et *bien que nous puissions légitimement inférer le pouvoir à partir de l'opération*, il faut se rappeler que l'inférence ne relève pas du domaine de la conscience mais de la raison. (Reid 1788, 8, italiques ajoutés)

metaphysics of freedom which makes freedom an intelligible possibility, and makes the distinction between the cases in which an agent is free to do something, and those in which he is not, an empirical matter. » (*ibid.*)

⁴²⁷ Puisqu'un énoncé causal implique (typiquement) des vérités contrefactuelles, alors la théorie de la perception de la causalité est une théorie qui montre la possibilité de percevoir un contenu contrefactuel. Voir (1999, 338).

Si le pouvoir est *dérivé* de notre conscience de l'action, c'est donc dans le sens d'une *inférence* : si *S* produit une certaine opération de l'esprit, alors il s'ensuit logiquement que *S* avait le *pouvoir* de produire cette opération. L'épistémologie reidienne du libre arbitre est donc une épistémologie en deux étapes : premièrement, conscience de l'opération de l'esprit, deuxièmement, inférence du pouvoir. Pour la première étape, Reid décrit plus précisément son contenu :

Dans certains mouvements de mon corps et directions de mon esprit, je sais, non seulement qu'il doit y avoir une cause qui a le pouvoir de produire ces effets, mais que je suis cette cause. (Reid 1788, 30)

Le contenu immédiat de la conscience d'agir s'exprime dans les termes de la théorie de la causalité agentive :

(1) Je suis la cause de cette opération.

De ce contenu immédiat peut ensuite être *inféré* légitimement :

(2) J'ai le *pouvoir* de causer cette opération.

Jusque là, nous avons une théorie de la perception du pouvoir en exercice, comparable à la théorie de Mumford. Comment Reid passe-t-il alors à la connaissance du pouvoir non exercé ? C'est-à-dire, comment va-t-il rendre compte de la connaissance de

(3) J'ai le pouvoir de *ne pas* causer cette opération.

La réponse de Reid se trouve dans sa théorie des pouvoirs causaux. Pour Reid, tout pouvoir causal est un pouvoir *agentif*, et tout pouvoir agentif est un pouvoir *des alternatives*. Il s'agit pour lui non pas d'une vérité contingente (ou extensionnelle), mais d'une vérité *conceptuelle* : notre *concept* même de cause vient de notre activité agentive, et ne peut donc être qu'un concept de cause agentive, et de cause agentive libre :

Le pouvoir actif, dont nous pouvons avoir une notion distincte, ne peut être présent que dans des êtres dotés d'intelligence et de volonté.

Le pouvoir de produire un effet implique le pouvoir de ne pas le produire.

(Reid 1788, 29)

Si le concept de pouvoir est tel que « pouvoir causer X » implique logiquement « pouvoir ne pas causer X », alors il est clair qu'à partir de la proposition

(2) J'ai le pouvoir de causer cette opération.

on peut directement déduire la proposition

(3) J'ai le pouvoir de ne pas causer cette opération.

Reid offre donc un modèle épistémologique dans lequel nous avons conscience immédiate de l'action, ou de l'opération, qui est l'exercice d'un pouvoir causal, et à partir de la conscience de l'exercice, nous pouvons déduire l'existence du pouvoir exercé, puis celle du pouvoir non

exercé. Le pouvoir non exercé, s'il n'est pas objet de conscience, est néanmoins déduit à partir d'un contenu connu par simple conscience.

Que penser de cette épistémologie du libre arbitre ? D'un point de vue strictement épistémologique, il me semble qu'elle est tout à fait satisfaisante : elle permet de rendre compte, à partir de la « perception » (ou de la conscience) de l'action effective, d'une connaissance d'un pouvoir agentif dual, et donc indéterministe. La grande difficulté de cette théorie réside dans les conséquences métaphysiques de la notion de « pouvoir causal » sur laquelle elle repose. Dire que tout pouvoir causal est nécessairement un pouvoir agentif, c'est dire que tout processus causal qui se déroule dans l'univers est l'activité d'un agent ; la conséquence, admise par Reid, est une métaphysique proche de l'occasionalisme malebranchiste, dans laquelle les processus causaux naturels qui ne sont pas des actions des agents humains (l'action de l'acide sur le métal, de la météorite sur le sol lunaire, etc.) sont tous, sans exception, des actions immédiates de l'agent divin. Cette conséquence métaphysique est difficile à admettre, même pour un philosophe théiste ; la plupart des philosophes théistes considèrent en effet que Dieu confère aux causes secondes inanimées un pouvoir causal qui leur est propre, un pouvoir qui n'est donc pas un pouvoir *agentif*. Si l'on peut défendre légitimement que le concept de cause agentive fait partie de notre expérience de sens commun (contre les partisans du « tout événementiel »), il est clair que la causalité non-agentive fait *aussi* partie de la manière dont nous nous représentons spontanément les événements du monde : lorsque je perçois l'influence causale d'une boule de billard sur une autre, j'ai bel et bien l'expérience d'une activité causale qui n'est pas l'activité d'un agent. Je ne m'attarderai pas ici sur les arguments que Reid avance pour défendre cette théorie purement agentive du pouvoir causal (pour cela, je renvoie à (Yaffe 2004, chap. 1 & 2) qui reconstruit ces arguments historiques de manière très claire) ; les conséquences métaphysiques de cette théorie me semblent suffisamment peu plausibles pour la compromettre sérieusement. Notre concept de pouvoir causal a fort peu de chance de contenir *analytiquement* la notion de pouvoir agentif, et en particulier de pouvoir libre ou dual, et par conséquent, en percevant l'exercice de notre pouvoir, nous ne pouvons pas *déduire* qu'il s'agit d'un pouvoir agentif libre.⁴²⁸

Thomas Reid a le mérite d'avoir proposé ce qui est, à ma connaissance, la première épistémologie complète du libertarianisme – une théorie dans laquelle le « pouvoir non exercé », et donc l'indéterminisme libertarien, peut être *connu* sur la base de notre expérience commune agentive. Mais il a dû pour cela admettre des conséquences métaphysiques trop lourdes. Nous allons voir à présent deux autres solutions d'intégration qui n'ont pas de telles conséquences.

La deuxième solution d'intégration est inspirée de la théorie de « l'expérience de l'effort » telle qu'elle est décrite par C.A. Campbell. Chez Campbell, il n'y a pas à

⁴²⁸ Une stratégie pour continuer à défendre la solution Reidienne pourrait consister à séparer l'implication « agentive » de l'implication « duale », en disant ceci : notre concept de pouvoir causal n'est peut-être pas nécessairement le concept d'un pouvoir agentif, mais cependant, il implique nécessairement le pouvoir de ne pas causer l'effet considéré. Après tout, pour obtenir une épistémologie *de l'indéterminisme*, c'est-à-dire pour passer de (2) à (3), c'est seulement de cette implication que nous avons besoin. Dans une telle stratégie, on n'aurait pas besoin de dire que toute la causalité dans l'univers est causalité agentive, mais on devrait dire néanmoins que toute causalité est *indéterministe* : à chaque fois qu'une substance inanimée cause un événement, elle aurait en même temps le pouvoir de *ne pas* causer cet événement. Certes, on peut défendre qu'il y a dans l'univers de la causalité indéterministe non agentive – par exemple dans le cas quantique, on peut défendre que telle particule a le pouvoir causal de se désintégrer *et* le pouvoir causal de *ne pas* se désintégrer. Mais il est clairement contraire à la science de soutenir que *toute* causalité naturelle est une causalité indéterministe. Par ailleurs, cette conception entièrement indéterministe de la nature n'est clairement pas celle que veut défendre Reid : s'il défend que toute causalité est indéterministe ou duale, c'est clairement parce qu'il le modèle du choix agentif à l'esprit – les deux implications (agentivité et dualité) sont pour lui inséparables (voir (Yaffe 2004, chap. 2)).

proprement parler de solution au problème épistémologique d'intégration de la causalité agentive.⁴²⁹ Mais la manière dont il décrit notre phénoménologie agentive peut suggérer une solution originale au problème épistémologique. Voici la description qu'il donne de l'expérience du libre arbitre, qui pour lui réside fondamentalement dans l'expérience de l'effort :

Dans l'acte qui requiert un effort, il n'y a pas seulement, de manière implicite, l'assurance absolue que le soi ici et maintenant est à l'origine de l'effort : il y a aussi l'assurance absolue que le soi pourrait se retenir de faire l'effort, et également [...] qu'il pourrait faire un effort plus grand, ou un moindre effort, avec les différences correspondantes dans l'acte concret de la volonté dans chaque cas.
(Campbell 1931, 135–136)

Dans cette description phénoménologique, on retrouve la distinction entre l'expérience de la source (le soi est « à l'origine » de l'effort) et l'expérience des alternatives (le soi « pourrait se retenir », etc.). Mais ce qui nous intéresse particulièrement, c'est que l'expérience des alternatives est ici rattachée à une expérience des *degrés* possibles de l'effort (le soi « pourrait faire un effort plus grand ou un moindre effort »). Ainsi, le pouvoir agentif, le pouvoir « de faire un effort », est perçu comme susceptible de divers degrés d'exercice. Les alternatives de l'agent ne se réduisent donc pas à « exercer son pouvoir agentif » ou « ne pas exercer son pouvoir agentif » ; même dans l'hypothèse où le pouvoir d'effort est exercé, il demeure une variété de degrés d'exercice qui auraient assurément des conséquences différentes. Par exemple, face à une tentation très forte de s'emporter contre un automobiliste inattentif, un effort moral de faible intensité pourra s'avérer insuffisant pour résister à la tentation. Au moment où je fais cet effort de faible intensité, je fais l'expérience que je *pourrais* faire un effort plus grand pour « atteindre mon devoir » (*to rise to duty* chez Campbell).

En quoi cette description phénoménologique peut-elle nous aider pour résoudre le problème d'intégration du pouvoir agentif libre ? L'idée est la suivante : dans la mesure où c'est le *même* pouvoir d'effort qui est susceptible de divers degrés d'exercice, on peut défendre qu'en le percevant *en exercice*, je perçois cette variabilité de degrés qui le caractérise. Ainsi, la perception des « alternatives possibles » n'a pas besoin de reposer sur la perception d'un *autre* pouvoir qui resterait inexercé (le pouvoir de « ne pas faire d'effort moral ») ; les alternatives sont inhérentes au pouvoir même qui est perçu en exercice. Cette théorie évite donc la difficulté de postuler la perception d'un pouvoir inactuel : le seul pouvoir qui est perçu est bel et bien actuel et en exercice.

En l'état, cette solution n'est pas complète, car elle repousse le problème à un autre niveau : certes le pouvoir qui est perçu est un pouvoir actuel et en exercice, mais dans la mesure où l'on dit qu'il y a perception des *degrés alternatifs* de ce pouvoir, certains de ces degrés doivent bien constituer des *possibilités* inactualisées, si ce n'est des *pouvoirs* hors exercice. Le défenseur de cette solution devra donc expliquer comment nous pouvons, en percevant un pouvoir actuel, percevoir en même temps les différentes *possibilités inactualisées* qui pourraient être réalisées par ce pouvoir. En repoussant ainsi le problème, on gagne beaucoup du point de vue dialectique, car le problème tombe alors dans la catégorie plus générale de l'épistémologie de la modalité. On peut reprendre ici un exemple de Peacocke :

⁴²⁹ Du point de vue épistémologique, Campbell ne donne pas d'autre indication que la remarque citée plus haut, à savoir que la source introspective (l'expérience *de l'agent*) est certainement *la meilleure source possible* pour découvrir un objet tel que le pouvoir *de l'agent*.

Considérez le cas d'un vase au bord d'une table. Nous savons qu'il pourrait facilement tomber. Nous savons cela parce que, étant donné sa position actuelle relativement au bord de la table, des conditions initiales légèrement différentes juste avant ce moment [...] auraient mené à ce que le vase soit renversé (par des vibrations, un courant d'air, ou des passants). Ici, la position actuelle du vase sur la table est capable d'interagir avec nous causalement. (Peacocke 1999, 337)

Autrement dit, les propriétés catégoriques et actuelles d'un objet nous placent parfois en position de connaître des vérités modales au sujet de cette chose (ce qui *pourrait* lui arriver, ce qui lui arriverait *si* ceci ou cela se passait, etc.). Pourquoi ne pourrait-il pas en être de même de la perception du pouvoir d'effort ? En percevant l'effort que nous produisons, nous pourrions peut-être percevoir une certaine caractéristique catégorique (son degré) qui nous placerait en position de connaître que d'autres événements sont possibles (d'autres degrés d'exercice du même pouvoir).

Comment connaissons-nous, d'une manière générale, les possibilités d'un objet sur la base de ses propriétés catégoriques perceptibles ? Il existe différentes « épistémologies de la modalité » qui pourraient être appliquées (peut-être avec une plausibilité inégale) au cas de la perception de l'effort agentif et de sa variable de « degré ». Ce pourrait être, par exemple, au moyen d'un « principe de recombinaison », tel qu'on le trouve chez Lewis (1986, 87–92) ou chez Armstrong (1989).⁴³⁰ Pour développer de manière complète ce modèle d'intégration, il faudrait approfondir cette question. Et je ne vois pas de raison décisive d'être pessimiste sur la possibilité de parvenir à un modèle complet.⁴³¹

Mais en ce qui me concerne, d'autres raisons m'amènent à préférer un autre modèle. Le modèle « à la Campbell » situe l'expérience incompatibiliste des alternatives dans l'expérience de l'effort (et de ses degrés) ; pour ma part, ainsi que je l'ai exprimé dans le chapitre 4, il me semble beaucoup plus probable que notre concept de libre arbitre vienne de notre expérience *du choix* (que celui-ci requière ou non un effort). Or dans l'expérience du

⁴³⁰ Pour un excellent résumé des différentes épistémologies de la modalité, notamment de l'épistémologie combinatorialiste de Lewis et Armstrong, reprise par Peacocke (1999, chap. 4), voir (Vaidya 2011).

⁴³¹ Une difficulté qu'on pourrait soulever est la suivante : les épistémologies de la modalité auxquelles je viens de faire référence, notamment l'épistémologie combinatorialiste (cf. note précédente), sont des épistémologies de la « possibilité métaphysique » et non pas de la « possibilité physique ». Or pour qu'il y ait libre arbitre, il est fondamental que les alternatives disponibles soient possibles physiquement, et pas seulement métaphysiquement. La connaissance du libre arbitre doit donc être la connaissance que des alternatives sont physiquement possibles, c'est-à-dire compatibles avec les lois de la nature. Certes, on peut défendre que la perception du degré d'effort permet de concevoir la possibilité *métaphysique* d'autres degrés du même effort, mais comment pourrait-on percevoir que ces autres degrés sont compatibles avec les lois de la nature ? Par comparaison, on peut défendre que la perception d'un nuage se déplaçant vers l'est nous rend capables de concevoir la possibilité métaphysique d'un mouvement du même nuage en sens inverse, vers l'ouest, mais cela ne nous permet pas de dire s'il y a ou non une loi de la nature qui *détermine* le mouvement du nuage vers l'est. Pour cela, il faudrait que nous connaissions les lois de la nature qui régissent les mouvements des nuages. D'une manière générale, l'objection consisterait à dire la chose suivante : pour savoir qu'une alternative est compatible avec (toutes) les lois de la nature, il faut connaître (toutes) les lois de la nature, ce qui ne peut se faire que sur la base d'une connaissance scientifique. Autrement dit, la seule épistémologie possible pour la possibilité physique (par opposition à la possibilité métaphysique) serait une épistémologie scientifique. Cette objection serait à approfondir mais elle ne me paraît pas décisive, parce que sa conclusion est beaucoup trop forte pour être plausible : s'il fallait connaître *toutes* les lois de la nature avant de pouvoir prétendre connaître que ceci est physiquement possible, alors nous n'aurions tout simplement *aucune* connaissance de la possibilité physique, car à aucun moment de l'histoire des sciences nous ne pouvons prétendre avoir une connaissance de toutes les lois de la nature. Et par conséquent, la notion même de possibilité physique deviendrait totalement inapplicable. Il semble donc qu'il faille accepter une certaine forme de connaissance de la possibilité physique qui précède la connaissance scientifique de toutes les lois, et selon toute vraisemblance cette connaissance de la possibilité physique reposera au moins en partie sur notre connaissance de possibilités métaphysiques. Voilà pourquoi j'ai proposé, dans le corps du texte, de fonder notre épistémologie des degrés possibles de l'effort sur l'une des épistémologies de la modalité (métaphysique).

choix, notre rapport au pouvoir agentif est très différent de ce qu'il est dans l'expérience de l'effort. En effet, lorsque je considère plusieurs options comme « ouvertes », ou pouvant être actualisées par moi, je me représente un pouvoir agentif, ou plutôt *deux ou plusieurs* pouvoirs agentifs, en amont de leur exercice. Je ne crois donc pas qu'on doive chercher à faire l'économie d'une épistémologie des pouvoirs *hors exercice*, comme le font les solutions de Reid et de Campbell. Du point de vue simplement doxastique, la *croyance* que nous avons tel et tel pouvoirs semble bien être formée (au moins dans certains cas) *avant* que ces pouvoirs soient exercés ou actualisés. Par conséquent, si c'est bien la croyance de sens commun que l'on veut « intégrer » (que l'on veut appeler une connaissance), il faut montrer que nous pouvons connaître l'existence de pouvoirs non exercés, notamment dans l'expérience du choix où nous sentons en nous deux ou plusieurs pouvoirs agentifs.

Le troisième modèle d'intégration de la causalité agentive sera donc un modèle de perception des pouvoirs hors exercice. Nous nous retrouvons confrontés à l'objection fondamentale : comment pourrions-nous *percevoir* quelque chose qui est simplement potentiel / inactualisé ? La réponse à ce problème reposera sur le fait que les pouvoirs causaux, même lorsqu'ils restent purement potentiels et inexercés, n'apparaissent pas n'importe comment dans le cours des événements, mais qu'ils reposent (en vertu de certaines lois de la nature) sur une base catégorique déterminée qui, elle, peut être perçue. Voyons comment cette réponse peut s'appliquer au cas des pouvoirs agentifs.

Ainsi qu'on l'a vu au chapitre 5 (sec. II.3), un pouvoir agentif est une propriété émergente d'un système macroscopique qui apparaît « lorsque ses composants microscopiques atteignent un certain niveau et un certain type appropriés de complexité organisationnelle ». Par exemple, si dans une situation donnée, Alice a le pouvoir agentif de dire la vérité, c'est que son corps se trouve, de manière catégorique, dans un certain type de configuration organisationnelle (C1) qui fait émerger ce pouvoir ; et si Alice a, en même temps, le pouvoir agentif de mentir, c'est que son corps se trouve également dans le type de configuration organisationnelle (C2) requis pour l'émergence de cet autre pouvoir. Alice ne pourra pas exercer ces deux pouvoirs en même temps, et l'un d'eux se trouvera donc nécessairement inexercé. Par ailleurs, au moment où Alice délibère, ou à l'instant qui précède immédiatement son choix, aucun des deux pouvoirs n'est exercé. Pourtant, à cet instant, la propriété C1 et la propriété C2 sont déjà des propriétés catégoriques du corps d'Alice. Et c'est en vertu de lois de la nature (des « lois d'émergence ») que la propriété C1 fait émerger le pouvoir de dire la vérité, et C2 le pouvoir de mentir. Nous avons là les éléments fondamentaux pour pouvoir construire une théorie de la connaissance, au sens qui a été rappelé à la section II.2., c'est-à-dire une théorie de la sensibilité du contenu de l'expérience et de la sûreté de ces implications.

Commençons par la condition de sensibilité. L'objectif est ici de montrer que lorsqu'Alice a l'expérience de pouvoir dire la vérité (et en même temps l'expérience de pouvoir mentir), cette expérience peut être sensible au fait qu'elle a effectivement le pouvoir de dire la vérité (et la deuxième expérience peut être sensible à son pouvoir de mentir). Je me concentrerai d'abord sur une seule des deux expériences, la seconde obéissant au même modèle. Pour voir comment l'expérience de pouvoir dire la vérité est sensible au pouvoir agentif de dire la vérité, le plus simple sera de partir de la condition de « double sensibilité » qu'est la condition de traçage :

(Condition de Traçage)

La croyance vraie de S que *p* « trace » le fait que *p* si et seulement si :

(3) Si p n'était pas vraie, S ne croirait pas que p .

(4) Si p était vraie (dans des conditions contrefactuelles proches), S croirait que p .⁴³²

Il sera plus simple de commencer par le deuxième conditionnel, le conditionnel de « sensibilité au vrai ». Peut-on donner un modèle de la croyance d'Alice telle que :

(4A) S'il était vrai que *Alice a le pouvoir agentif de dire la vérité* (dans des conditions contrefactuelles proches), alors Alice aurait la croyance qu'elle peut dire la vérité.

ou encore, en termes de mondes possibles :

(4A') Dans les mondes proches où Alice a le pouvoir agentif de dire la vérité, Alice a aussi la croyance qu'elle peut dire la vérité.

Ce conditionnel pourrait-il être vrai ? Lorsqu'on y réfléchit, il me semble que non seulement ce conditionnel *pourrait* être vrai, mais que c'est même la possibilité la plus naturelle et la plus plausible. Voici pourquoi : c'est un principe bien connu de philosophie de l'action qu'il est irrationnel d'entreprendre une action lorsqu'on ne croit pas être en mesure de la réaliser ; c'est-à-dire qu'un agent *rationnel* ne peut pas tenter de réaliser une action s'il croit ne pas pouvoir la réaliser ; et s'il ne tente pas de la réaliser, alors *a fortiori* il ne peut pas non plus la réaliser ; par conséquent, lorsqu'un agent rationnel ne croit pas pouvoir réaliser une action, cet état doxastique lui enlève le pouvoir agentif de réaliser l'action. Autrement dit, la croyance de pouvoir dire la vérité est une condition nécessaire ou une composante (pour un agent rationnel) du pouvoir agentif de dire la vérité. Cette corrélation que je viens d'exprimer au niveau des états mentaux (l'état mental de croyance et l'état mental de pouvoir agentif) doit se retrouver également au niveau des états physiques qui les réalisent : dire que la croyance de pouvoir dire la vérité est une composante du pouvoir agentif correspondant, c'est dire que la réalisation neuronale de cette croyance fait partie de la « base d'émergence » microscopique du pouvoir agentif. Le pouvoir agentif de dire la vérité émerge sur une certaine configuration physique $C1$ (qui concerne en grande partie le cerveau), et cette configuration inclut notamment certaines caractéristiques neuronales qui réalisent, au niveau mental, la croyance de pouvoir dire la vérité. Ou, pour le dire autrement, la croyance de pouvoir dire la vérité et le pouvoir agentif correspondant sont « coréalisés » : c'est (partiellement) le même état physique qui réalise les deux états mentaux. Pour être précis, il faut se rappeler que cette coréalisation ne vaut que pour un agent *rationnel* : on peut concevoir que des agents irrationnels, tout en étant persuadés de ne pas pouvoir faire A , essayent néanmoins de faire A , et éventuellement parviennent effectivement à faire A – ce qui reviendrait à dire qu'ils avaient, en fait, le pouvoir agentif de faire A . Donc, dans le cerveau d'un agent irrationnel, on peut concevoir que la croyance de pouvoir faire A soit réalisée par une certaine configuration neuronale $X1$, totalement indépendante de la configuration neuronale $C1$ sur laquelle émerge le pouvoir agentif de faire A ; et par conséquent, un agent irrationnel pourrait avoir $C1$ sans avoir $X1$. Mais le cerveau d'un agent rationnel et celui d'un agent irrationnel ne sont pas configurés de la même manière : une même croyance peut être réalisée chez l'un et chez l'autre par des configurations neuronales très différentes. L'important pour nous, c'est que lorsqu'un cerveau a la bonne configuration (la configuration « rationnelle »), nous avons toutes les raisons de penser que la réalisation physique de la croyance d'Alice fait partie intégrante de la base physique qui permet à son pouvoir agentif d'émerger. Cela est important parce que la

⁴³² Pour rappel, la condition de sensibilité, *stricto sensu*, ne concerne que le conditionnel (3), c'est-à-dire la « sensibilité au faux », tandis que la condition de traçage est à la fois une sensibilité au faux et une sensibilité au vrai. Voir chapitre 3, p.225.

rationalité d'un agent est certainement un paramètre qui variera peu lorsqu'on considérera des mondes possibles proches : si Alice est un agent rationnel dans le monde actuel, alors elle l'est très certainement aussi dans les mondes proches du monde actuel. Et par conséquent, dans les mondes proches dans lesquels Alice a le pouvoir agentif de dire la vérité, elle a aussi la croyance qui, chez un agent rationnel, est coréalisée avec ce pouvoir, c'est-à-dire la croyance de pouvoir dire la vérité. Autrement dit, la croyance d'Alice « j'ai le pouvoir de dire la vérité » est très certainement sensible à la *vérité* de son contenu : si le contenu de cette croyance est vrai, alors Alice a cette croyance. On a donc bel et bien :

(4A) S'il était vrai que *Alice a le pouvoir agentif de dire la vérité* (dans des conditions contrefactuelles proches), alors Alice aurait la croyance qu'elle peut dire la vérité.

Mais cette croyance est-elle également sensible à la *fausseté* de son contenu ? Pour cela, il faut vérifier le conditionnel suivant :

(3A) S'il n'était pas vrai que *Alice a le pouvoir agentif de dire la vérité*, alors Alice n'aurait pas la croyance qu'elle peut dire la vérité.

ou encore, en termes de mondes possibles :

(3A') Dans le monde possible le plus proche où Alice n'a pas le pouvoir agentif de dire la vérité, Alice n'a pas non plus la croyance qu'elle peut dire la vérité.

Ce conditionnel peut-il être vrai ? En s'appuyant sur la réponse que nous avons faite pour le conditionnel (4A'), on peut voir qu'il est tout à fait possible que (3A') soit vrai également. Pour (4A'), nous avons dit qu'un certain type d'agents, les agents *rationnels*, étaient tels qu'ils ne pouvaient pas avoir le pouvoir agentif sans avoir la croyance de pouvoir, parce que la réalisation neuronale de cette croyance (dans une certaine configuration « rationnelle » du cerveau) est inséparable de la base d'émergence du pouvoir agentif. Inversement, pour (3A'), on peut imaginer qu'un certain type d'agents, disons les agents *lucides*, sont tels que normalement ils ne peuvent pas avoir la croyance qu'ils peuvent faire un certain choix lorsqu'ils n'ont pas le pouvoir agentif correspondant. Un agent « lucide » est donc un agent qui, normalement, ne croit pas pouvoir faire ce qu'en fait il ne peut pas faire. Là encore, cette corrélation pourrait s'expliquer par une forme de coréalisation dans le cerveau : il se pourrait que, pour une certaine configuration cérébrale (la configuration « lucide »), la base d'émergence du pouvoir agentif de faire un certain choix (mentalement) se trouve coréalisée avec la réalisation physique de la croyance de pouvoir, de telle sorte qu'il soit impossible d'avoir la réalisation physique de la croyance sans que la base d'émergence se trouve actualisée. A la différence de la configuration « rationnelle », je ne dirais pas que nous avons toutes les raisons de supposer que les agents normaux ont une configuration cérébrale « lucide ». Mais il s'agit certainement d'une *possibilité*, et c'est tout ce dont nous avons besoin pour construire un *modèle* de croyance sensible.

En réunissant les remarques que nous venons de faire sur la sensibilité au vrai et la sensibilité au faux, on obtient un modèle dans lequel la croyance de pouvoir faire un choix X « trace » son objet, c'est-à-dire le pouvoir agentif.

Modèle de traçage de la croyance au pouvoir agentif par coréalisation croyance-pouvoir

Chez des agents normaux (rationnels et lucides), la configuration du cerveau vérifie la propriété suivante : la réalisation neuronale de la croyance de pouvoir faire un choix X

et la base d'émergence du pouvoir agentif de faire le choix X sont si étroitement corrélés (constituent des états physiques si indissociables) que les deux conditionnels suivants sont vrais

(3A') Dans le monde possible le plus proche où cet agent n'a pas le pouvoir agentif de faire le choix X, [l'agent a toujours une configuration cérébrale lucide et donc] il n'a pas non plus la croyance coréalisée de pouvoir faire le choix X.

(4A') Dans les mondes possibles proches où cet agent a le pouvoir agentif de faire le choix X, [l'agent a toujours une configuration cérébrale rationnelle et donc] il a également la croyance coréalisée de pouvoir faire le choix X.

Pour l'instant, nous n'avons parlé que d'un seul pouvoir agentif, par exemple le pouvoir pour Alice de dire la vérité, et nous venons de voir qu'il était *possible* que l'occurrence de ce pouvoir agentif soit tracée par la croyance d'avoir ce pouvoir. Mais la croyance d'avoir un certain pouvoir n'est pas encore une croyance *au libre arbitre*, si du moins on rejette la proposition Reidienne selon laquelle tout pouvoir est un pouvoir des alternatives. Pour donner un modèle de la connaissance du libre arbitre, il faut donc combiner deux croyances, correspondant à deux pouvoirs agentifs contraires :

Modèle de double traçage des pouvoirs agentifs

Un agent normal (rationnel et lucide), lorsqu'il est confronté à un choix entre « faire X » et « faire Y », vérifie les propriétés suivantes :

- la réalisation neuronale de la croyance « je peux faire le choix X » est corrélée à la base d'émergence du pouvoir agentif de faire le choix X, et

- la réalisation neuronale de la croyance « je peux faire le choix Y » est corrélée à la base d'émergence du pouvoir agentif de faire le choix Y,

de telle sorte que les deux conditionnels suivants sont vrais :

(3double) Si cet agent n'avait pas les deux pouvoirs agentifs alternatifs (faire X ou faire Y), il n'aurait pas non plus les deux croyances correspondantes (de « pouvoir faire X » et de « pouvoir faire Y ») ni la croyance conjonctive de « pouvoir faire soit X soit Y »,

(4double) Si cet agent avait les deux pouvoirs agentifs alternatifs (dans des conditions contrefactuelles proches), il aurait les deux croyances correspondantes, et potentiellement la croyance conjonctive de « pouvoir faire soit X soit Y ».

Dans ce modèle, la croyance d'avoir deux pouvoirs agentifs contraires trace de manière fiable le fait d'avoir réellement deux pouvoirs agentifs contraires. Si donc le « fait libertarien » est constitué métaphysiquement par la concurrence de deux pouvoirs agentifs contraires, alors il est possible de donner un modèle épistémologique dans lequel nous pouvons connaître ce fait de manière sensible.

Cela nous permet-il également de connaître l'indéterminisme ? A cause de l'engagement de clôture (et de l'engagement de méta-cohérence), il est fondamental qu'on puisse également connaître l'indéterminisme parce qu'il s'agit d'une conséquence logique du « contenu propre » de la croyance au libre arbitre telle que nous venons de la décrire. En effet, si Alice a le pouvoir agentif réel de dire la vérité, et le pouvoir agentif réel de mentir, alors rien ne détermine si elle va exercer le premier pouvoir ou le second, donc on se trouve dans une situation proprement indéterministe. Etant donné que la croyance au double pouvoir agentif implique logiquement l'indéterminisme, on peut être amené à déduire la seconde de la première : sur la base de son expérience agentive de pouvoir faire deux choses contraires, Alice peut déduire qu'elle se trouve dans un univers non déterministe. Mais sa croyance déduite pourra-t-elle être une connaissance ?

On peut se demander tout d'abord si cette croyance peut être *sensible* à la vérité de l'indéterminisme. Là, la réponse est très certainement non. C'est-à-dire qu'on a très certainement le conditionnel suivant :

(~3D) Si l'indéterminisme n'était pas vrai (si Alice se trouvait dans un monde déterministe), Alice aurait tout de même formé la croyance à l'indéterminisme, et ce de la même manière (sur la base de ses croyances aux deux pouvoirs agentifs contraires).

ou encore

(~3D') Dans le monde déterministe le plus proche du monde actuel, Alice a encore les deux croyances aux pouvoirs agentifs contraires, et en déduit la croyance indéterministe.

Pourquoi ? Pour voir cela, il faut essayer de déterminer quel est le monde déterministe w_1 le plus proche du monde w_0 où Alice a deux pouvoirs agentifs contraires (qui émergent sur des configurations microscopiques C1 et C2, et qui sont générateurs d'indéterminisme) ainsi que les deux croyances correspondantes. Le monde déterministe le plus proche est un monde dans lequel la situation initiale d'Alice est la même (C1 et C2), mais où les lois de la nature déterminent un unique futur pour cette situation initiale. C'est-à-dire que c'est un monde où les lois de la nature sont fondamentalement différentes de celles qui valent dans w_0 . Si nous cherchions le monde le plus proche dans lequel ce choix d'Alice est déterminé, alors nous pourrions sélectionner un monde possible qui ne diffère de w_0 que par des faits catégoriques : il s'agirait certainement d'un monde dans lequel Alice ne se trouve pas dans C1 et C2 mais plutôt dans une situation initiale distincte C3 qui, sans changer les lois valant dans w_0 , entraînent un unique futur possible. Mais ce monde (appelons-le w_2) n'est *pas* un monde déterministe, c'est-à-dire un monde dont les lois de la nature permettent de calculer un unique futur pour n'importe quelles conditions initiales. C'est simplement un monde dans lequel *un certain choix* est déterminé. Si donc nous recherchons le monde *déterministe* le plus proche de w_0 , nous sommes obligés de considérer un monde possible dont les lois de la nature sont distinctes de celles de w_0 , et si les lois de la nature sont distinctes de celles de w_0 , il n'est pas nécessaire en revanche de faire varier les conditions initiales dans lesquelles se trouvent Alice. Ainsi, le monde déterministe le plus proche de w_0 est un monde où Alice se trouve dans les mêmes conditions C1 et C2 que dans w_0 , mais où ces conditions initiales entraînent (en vertu des lois de la nature) un unique choix possible, par exemple mentir. Par conséquent, dans w_1 , Alice n'a pas réellement le pouvoir agentif de dire la vérité ; elle a seulement (au mieux) le pouvoir agentif (unilatéral) de mentir. Donc la configuration C1, qui dans w_0 faisait émerger le pouvoir agentif de dire la vérité, ne fait pas émerger ce même pouvoir dans w_1 . C'est-à-dire que w_1 diffère de w_0 non seulement par ses lois diachroniques, mais aussi par ses lois d'émergences, qui comme les lois diachroniques sont des lois de la nature.

Reste à savoir si, dans w_1 , Alice a la croyance de pouvoir dire la vérité (bien qu'elle n'ait pas réellement ce pouvoir). Nous avons proposé que, pour être fiable, la croyance de pouvoir dire la vérité soit réalisée au niveau cérébral par la même configuration neuronale C1 sur laquelle émerge le pouvoir agentif lui-même ; donc, dans w_0 , les lois psychophysiques sont telles que la configuration C1 réalise au niveau mental la croyance de pouvoir dire la vérité (et C2 la croyance de pouvoir mentir). Puisque C1 (ainsi que C2) est réalisée également dans w_1 , il s'ensuit que Alice aura la croyance de pouvoir dire la vérité (et celle de pouvoir mentir) dans w_1 à la seule condition que les lois psychophysiques soient les mêmes dans w_1 que dans w_0 . Est-ce le cas ? Il semble bien que oui : en effet, pour passer de w_0 à un monde déterministe, on est obligé de modifier les lois diachroniques et les lois d'émergence, mais il n'est nullement besoin de changer *en outre* les lois psychophysiques (qui sont elles aussi des

lois de la nature) ; pour atteindre le monde déterministe *le plus proche*, il semble qu'il faille faire un minimum de modifications possible aux lois de la nature. Donc, le monde déterministe le plus proche de w_0 est probablement un monde qui partage avec w_0 non seulement les mêmes conditions initiales (C1 et C2) mais aussi les mêmes lois psychophysiques (parce qu'il n'est pas *nécessaire* de les modifier pour obtenir un monde déterministe). Et par conséquent ce monde w_1 sera un monde dans lequel Alice a la croyance de pouvoir dire la vérité et la croyance de pouvoir mentir (qui sont réalisées par C1 et C2 en vertu des lois psychophysiques de w_0) et pourra en déduire la croyance indéterministe. Autrement dit,

(~3D') Dans le monde déterministe le plus proche du monde actuel, Alice a encore les deux croyances aux pouvoirs agentifs contraires, et en déduit la croyance indéterministe.

Donc cette croyance indéterministe, que l'on peut déduire de la double croyance à deux pouvoirs agentifs contraires, n'est pas sensible à la fausseté de son contenu.

Comme pour l'intégration du modèle événementialiste, c'est la théorie de la sûreté épistémique qui nous permet ici de résoudre le problème, et de considérer la croyance indéterministe comme une connaissance bien qu'elle ne soit pas sensible à son contenu. En effet, pour les mêmes raisons que nous avons vues plus haut (p.427), l'indéterminisme, s'il est vrai dans un monde possible, est certainement vrai également dans la grande majorité (peut-être la totalité) des mondes de son voisinage modal. Et par conséquent, si Alice croit à l'indéterminisme dans un monde indéterministe, sa croyance est sûre ; elle ne pourrait pas facilement être fausse. Si par ailleurs cette croyance sûre est obtenue par déduction compétente à partir d'une croyance sensible à son contenu (la croyance au double pouvoir agentif, dont nous venons de voir qu'elle pouvait être sensible), alors nous avons toutes les raisons de considérer cette croyance sûre comme une *connaissance*, parce qu'elle a tous les mérites épistémiques que l'on peut attribuer à nos croyances anti-sceptiques.

Pour résumer, nous venons donc de voir que la croyance au libre arbitre peut tout à fait être épistémiquement fiable si l'on adopte un modèle agentif du libre arbitre : contrairement à ce que défendait Hume, une forme de « perception » de notre pouvoir causal agentif – et même de notre *double* pouvoir causal agentif – est possible. Pour finir cette question de l'intégration de la causalité agentive, voyons plus précisément comment le modèle que nous venons de donner répond à l'objection humienne contre la perception du pouvoir agentif. On se rappelle que l'objection humienne consiste à dire que nous ne pouvons pas percevoir dans la volition une disposition à produire tel mouvement plutôt que tel autre, pour les mêmes raisons que nous ne pouvons pas percevoir dans le mouvement de la boule blanche une disposition à produire tel mouvement dans la boule de couleur qu'elle va frapper. Pour connaître la conséquence possible, il faut l'avoir préalablement observée. Pour répondre à cette objection, il faut premièrement remarquer un problème fondamental dans l'identification du pouvoir agentif : pour Hume, le pouvoir agentif se situe dans le lien de corrélation entre nos volitions et nos mouvements corporels. C'est ce que j'ai appelé au chapitre 4 (p.323) la question de l'*efficacité*. L'efficacité est bien sûre une partie importante du contrôle que nous avons sur nos actions, mais c'est une partie évidemment compatible avec le déterminisme, et qui n'est donc pas le pouvoir agentif que nous cherchons dans un modèle libertarien de la causalité agentive. Le pouvoir agentif, dans un modèle libertarien, ne se situe pas dans le lien (déterministe) entre volition et mouvement, mais dans la possibilité de produire différentes volitions à l'issue d'un *choix*. Il se situe, pour ainsi dire, en amont de la volition. Ce qui doit être perçu, ce n'est donc pas le pouvoir d'une volition de produire un mouvement, mais le pouvoir de l'agent de produire une volition. Cette clarification suffit-elle à supprimer le

scepticisme humien ? Sans doute pas : face à cette réponse, Hume répondrait certainement qu'il nous est tout autant impossible de percevoir « dans l'agent » une disposition à produire tel ou tel choix (telle ou telle volition) avant de l'avoir vu produire effectivement un tel effet. Et comme on l'a vu, il y a là un véritable problème de « perception du non actuel » : un lien perceptif doit reposer sur un lien causal, or ce qui est simplement potentiel ne peut avoir d'efficacité causale. La solution apportée par le troisième modèle d'intégration de la causalité agentive est la suivante : si, au moment du choix, les deux volitions qui sont au pouvoir de l'agent sont effectivement de simples potentialités, cependant l'existence des deux pouvoirs agentifs concurrents émerge sur une base catégorique qui, quant à elle, est tout à fait actuelle. Il n'y a donc aucun problème de perception pour cette base catégorique, et les pouvoirs qui émergent sur cette base sont si étroitement corrélés à cette base que notre croyance au pouvoir peut être *sensible* à la présence ou à l'absence de ces pouvoirs. Peut-on alors parler au sens étroit de *perception* des pouvoirs ? Ce qui est certain, c'est qu'il y a dans ce cas une croyance fiable par traçage occurrence des pouvoirs agentifs. Savoir s'il s'agit ou non d'une perception me semble ensuite une question largement terminologique. La seule conclusion importante, c'est que notre croyance au libre arbitre agentif peut constituer une connaissance, selon les critères de connaissance rappelés à la section II.2.

On peut donc donner un modèle satisfaisant dans lequel le libertarianisme agentif fait l'objet d'une connaissance, c'est-à-dire un modèle dans lequel la croyance de pouvoir faire deux choix contraires trace son contenu propre de manière *sensible*, et dans lequel elle est *sûre* à l'égard de son implication logique (l'indéterminisme). Comme nous l'avons rappelé en II.2, c'est tout ce qui est requis pour considérer qu'une croyance constitue une connaissance.

II.5. Conclusion : la possibilité d'une *connaissance* introspective de l'indéterminisme

La conclusion des deux sections précédentes est donc la suivante : quel que soit le modèle métaphysique que l'on donne pour le libre arbitre libertarien, il est également possible d'en construire un « modèle d'intégration », c'est-à-dire un modèle dans lequel le fait métaphysique du libre arbitre est connu de manière fiable sur la base de notre expérience phénoménale du libre arbitre. Dans cette dernière section, j'aborderai trois points. Premièrement, je poserai la question des mérites comparés des deux modèles du libre arbitre, relativement à la question d'intégration. Deuxièmement, je montrerai que l'approche introspectionniste, adoptée dès le début de cette étude, révèle particulièrement son intérêt pour résoudre ce dernier problème qu'est le problème d'intégration épistémique. Enfin, j'aborderai la réaction « d'incrédulité » que peuvent susciter les modèles d'intégration proposés, à cause de leur caractère hautement spéculatif ; je répondrai à cette difficulté en remplaçant ces modèles dans la dialectique générale de cette étude.

Le problème d'intégration donne-t-il un avantage au libertarianisme agentif ou au libertarianisme événementiel ? En conclusion du chapitre précédent, nous avons dit que, du point de vue des objections de premier ordre, les deux modèles étaient tous les deux aussi défendables. Mais il se pourrait que l'un des deux prenne l'avantage parce qu'il disposerait d'une réponse plus convaincante au problème d'intégration. Mon sentiment est que le problème d'intégration ne change pas sensiblement la donne. Si l'on compare les réponses agentive et événementialiste au problème d'intégration, on peut constater une grande hétérogénéité, mais cette hétérogénéité reste de même nature que celle observée au chapitre 5 : le modèle d'intégration événementialiste est plus détaillé d'un point de vue scientifique (il suppose ou prédit l'existence de certains circuits neuronaux de détection de l'état cérébral),

tandis que le modèle d'intégration agentif garde un vocabulaire et des conditions de vérité plus « métaphysique » ou moins empirique. Mais nous avons déjà observé cette asymétrie, et commenté que l'on pouvait voir des avantages et des inconvénients dans chacune des deux approches. En particulier, le caractère plus « empirique » du modèle événementialiste présente l'inconvénient d'être beaucoup plus spéculatif : qui pourrait prétendre à partir du seul sens commun qu'il existe réellement un système neuronal de détection des instabilités chaotiques dans le cerveau ? Nous envisagerons ce problème un peu plus loin, pour montrer qu'il n'a rien de dirimant. Mais il est certain que la solution agentive prête beaucoup moins le flanc à ce type de réaction, qu'elle soit ultimement légitime ou non.

Inversement, le caractère moins empirique de la solution agentive peut donner à certains le sentiment qu'il s'agit d'une théorie non soumise à la vérification et à la falsification, et donc purement gratuite. D'un point de vue poppérien (ou lakatosien), les « prédictions » empiriques du modèle d'intégration événementiel ouvrent assurément des possibilités d'investigation expérimentale pour les décennies à venir. Ces prédictions invitent en effet à déterminer empiriquement les corrélats neuronaux de notre expérience du libre arbitre, et à voir si ces corrélats neuronaux « tracent » ou non l'occurrence d'une certaine configuration neuronale chaotique. Après tout, dira l'événementialiste, il est certain que nos expériences de choix ouvert entre deux options sont réalisées cérébralement par un certain état neuronal, et il est à peu près certain que cet état neuronal n'apparaît pas de manière arbitraire et hasardeuse dans le cerveau mais qu'il est la conséquence systématique de certaines causes déterminées, c'est-à-dire qu'il doit bien « tracer quelque chose » ; même si nous ne savons pas encore ce qu'il trace, n'est-il pas raisonnable de s'attendre à ce qu'il trace un certain état du cerveau dans lequel les deux options ouvertes (mentalement) sont réalisées par deux réseaux neuronaux distincts ? Et n'est-il pas raisonnable de s'attendre à ce qu'il trace une forme « d'instabilité » neuronale générée par ces deux réseaux neuronaux ? Les hypothèses du modèle événementialiste semblent donc constituer des hypothèses de travail raisonnables pour une expérimentation empirique à venir. Bien sûr, cette expérimentation à venir pourrait fort bien réfuter le modèle proposé – c'est le lot de toute théorie scientifique. Mais il est important de noter ici qu'une réfutation empirique du modèle proposé ne serait pas nécessairement une réfutation de la connaissabilité du libre arbitre, dans la mesure où nous n'avons proposé qu'*un* modèle possible : il se pourrait tout à fait que l'investigation empirique révèle que notre expérience du libre arbitre trace en fait un autre fait cérébral F (auquel nous n'avons pas pensé ici) et que ce fait F soit corrélé de manière fiable ou sûre à l'indéterminisme macroscopique du choix. En somme, le modèle événementiel ouvre la voie à un grand projet de recherche expérimental :

Projet de recherche neuroscientifique du fait F

Quel est le fait cérébral « tracé » (de manière fiable) par l'occurrence du corrélat neuronal de nos expériences de libre arbitre ?

Si ce projet est particulièrement important pour le modèle événementiel, il me semble que le défenseur du modèle agentif devrait également en reconnaître la pertinence : une meilleure connaissance de notre expérience du libre arbitre et de ses conditions d'occurrence ne peut en effet qu'apporter des éléments intéressants pour rendre plus précise et plus réaliste toute épistémologie introspectionniste.

Au chapitre 4 (p.314), j'ai annoncé qu'une de mes raisons principales pour adopter une épistémologie introspectionniste du libre arbitre – une épistémologie fondée sur notre expérience agentive en première personne et non pas, par exemple, sur notre conviction de la responsabilité morale – était sa capacité à répondre au problème d'intégration. C'est pour

cette raison que j'ai choisi de ne pas approfondir les autres modèles épistémologiques. Nous sommes à présent en mesure de mieux comprendre ce choix.

Le grand avantage épistémologique de l'approche phénoménale ou introspectionniste vient du caractère « *occurrentiel* » de l'expérience agentive : l'expérience phénoménale du libre arbitre est une expérience qui ne se produit pas n'importe quand mais au moment même où l'on agit. Il est donc naturel, dans ces conditions, de soutenir qu'elle se produit (normalement) au moment même où a lieu le fait métaphysique qui constitue un indéterminisme libertarien. Autrement dit, le caractère *occurrentiel* de l'expérience phénoménale ouvre très naturellement la possibilité d'un *traçage* du fait métaphysique (de ses occurrences dans le temps) par l'expérience phénoménale. Or un tel type de traçage est un atout considérable pour construire une théorie convaincante de la fiabilité externaliste. De fait, dans l'un et l'autre modèle d'intégration que nous avons proposé, le traçage *occurrentiel* joue un rôle important. Inversement, si nous avons adopté une épistémologie de la responsabilité morale comme celle de Peter van Inwagen, un tel traçage aurait été impossible : la conviction d'être moralement responsable est une conviction permanente chez l'homme, et donc elle ne trace l'occurrence d'aucun fait particulier.

Pour finir, il faut envisager une réaction que les modèles d'intégration ne manqueront pas de susciter, à savoir la réaction d'incrédulité devant le caractère hautement spéculatif de ces modèles – peut-être tout spécialement du modèle d'intégration événementiel. Cette réaction peut s'exprimer ainsi : êtes-vous réellement en train de prétendre qu'il existe un système neuronal de détection des instabilités chaotiques dans le cerveau ? Qui pourrait prétendre savoir une telle chose aujourd'hui, à partir du seul sens commun ? Cette réaction n'est pas facile à éliminer et risque fort de rester un obstacle majeur à l'acceptation de nos conclusions ; cependant, si elle est difficile à éliminer, ce n'est pas à mon sens parce qu'il est impossible d'y répondre, mais parce que la réponse suppose d'avoir bien présente à l'esprit la situation dialectique complexe où nous nous trouvons.

La première réponse consiste donc à rappeler cette situation dialectique : dans ce chapitre, je n'ai pas cherché à décrire positivement comment le libre arbitre est connu, mais seulement à proposer des *modèles* de connaissance du libre arbitre, c'est-à-dire à défendre qu'il est *possible* de le connaître, pour répondre à une objection qui consistait à affirmer que c'est *impossible*. Si je n'ai pas cherché à m'engager d'avantage en faveur de la vérité de tel ou tel modèle, c'est que cela me semblerait en effet hautement spéculatif (en l'état actuel de nos connaissances) et que cela n'est pas nécessaire dialectiquement pour avoir une croyance justifiée au libre arbitre libertarien : ce qui est requis, à cause du principe de méta-cohérence, c'est de soutenir (rationnellement) *que* nous connaissons le libre arbitre (qu'il ne s'agit pas d'une croyance purement hasardeuse), mais non pas de prétendre savoir *comment* nous le connaissons. Dans de nombreux cas, lorsque nous sommes justifiés à dire qu'une chose est le cas (par exemple, qu'il pleut), nous sommes aussi en mesure de donner des détails sur la *manière* dont la chose est le cas (il pleut ici ou là, il pleut doucement ou à verse), parce que notre justification pour affirmer le fait consiste en une observation du fait. Mais dans d'autres cas, nous pouvons être justifiés à croire qu'il pleut sans être en mesure de dire s'il pleut ici ou là, doucement ou à verse : c'est le cas notamment lorsque notre justification est d'ordre testimonial (quelqu'un de fiable m'a dit qu'il pleuvait, sans me donner davantage de détails). Nous sommes ici dans une situation comparable : notre justification positive à affirmer que nous connaissons le libre arbitre ne repose pas sur une observation du fait que notre croyance trace le fait du libre arbitre ; elle repose sur des considérations épistémologiques fondamentales présentées au chapitre 3, à savoir que notre intuition du libre arbitre fait partie de celles auxquelles nous devons accorder, *prima facie*, le statut de connaissance. Bien sûr, il serait plus confortable épistémiquement d'avoir une confirmation observationnelle du fait que

nous connaissons, une confirmation qui nous renseignerait sur la manière dont nous connaissons. Mais cette confirmation n'est pas justificationnellement indispensable : il suffit que notre justification *prima facie* à croire que nous connaissons ne soit pas réfutée. Or les modèles d'intégration possibles permettent de montrer que notre justification *prima facie* n'est pas réfutée par un argument impossibiliste. Pour résumer, à la question « êtes-vous en train de prétendre qu'il y a vraiment un système de détection de réseaux neuronaux chaotiques ? », la réponse est : « non, évidemment. Je prétends que notre croyance au libre arbitre dépend *d'une manière ou d'une autre* de l'indéterminisme libertarien ; et si quelqu'un m'objecte que c'est *impossible*, je peux lui présenter *tel modèle* qui montre que cela n'est pas impossible. »

Cette réponse ne satisfera peut-être pas : après tout, si l'on affirme positivement que nous connaissons le libre arbitre, on ne peut pas s'en tenir à dire que les modèles de connaissance sont « simplement possibles » ; il *faut bien* que l'un ou l'autre d'entre eux soit *actuel*. Imaginons par exemple que les modèles M1 et M2 soient les *seuls* modèles possibles de connaissance du libre arbitre ; dans ce cas, pour affirmer rationnellement que nous connaissons le libre arbitre, il faut bien être prêt à affirmer que soit M1 soit M2 est bel et bien le cas – et non pas simplement *possible*. Et si chacun de ces modèles est absolument incroyable, alors leur disjonction sera, elle aussi, absolument incroyable, c'est-à-dire qu'il sera absolument incroyable que nous connaissions le libre arbitre. Toutes ces remarques me semblent tout à fait légitimes : en particulier, lorsque je dis que nous avons une justification *prima facie* à croire que nous connaissons d'une manière ou d'une autre, cela veut dire, évidemment, qu'un modèle ou un autre doit être effectivement vrai et non pas simplement possible. Mais je ne prétends absolument pas que les modèles présentés dans ce chapitre soient les *seuls* modèles possibles. Il se pourrait par conséquent que nous connaissions le libre arbitre d'une manière que je n'ai pas imaginée et modélisée ici. Par ailleurs, à supposer que les modèles présentés ici soient les seuls, je ne crois pas non plus qu'on puisse dire qu'ils sont absolument incroyables. Plus précisément, je ne pense pas qu'ils soient *intrinsèquement* absurdes ou improbables ; si on peut les trouver « incroyables », c'est seulement dans le sens faible où ils sont « sans aucun support observationnel », et donc dans le sens où il nous est impossible, pour l'instant, de les croire de manière justifiée. Mais s'il s'avérait que l'un d'eux soit finalement le seul modèle possible de connaissance du libre arbitre, alors notre justification métaépistémique à croire que nous connaissons le libre arbitre deviendrait une justification *suffisante* pour affirmer la vérité de ce modèle, même en l'absence de support observationnel. En l'état actuel de nos connaissances, la situation n'est pas du tout celle-là : en l'état actuel, il existe plusieurs modèles (certains présentés ici, d'autres sans doute que je n'ai pas imaginés), tous complexes et spéculatifs mais tous intrinsèquement possibles, et si nous sommes justifiés à dire que *l'un ou l'autre* de ces modèles doit être bel et bien actuel, nous n'avons pas à nous engager en faveur de l'un plutôt que de l'autre.

La comparaison avec les théories antiques de la vision permettra sans doute de mieux comprendre la place dialectique des modèles présentés dans ce chapitre. Comme nous l'avons vu au chapitre 3 (p.203), il existait dans l'antiquité diverses théories sur la manière dont nos impressions (et nos croyances) visuelles dépendent du fait extérieur observé : certains modèles (théories de l'émission d'Euclide ou Ptolémée) postulaient l'émission de certaines entités de l'œil vers l'objet regardé, d'autres (théories de l'intromission des atomistes et d'Aristote) postulaient le déplacement de certaines entités de l'objet regardé vers l'œil, d'autres enfin (notamment Platon) adoptaient un modèle mixte entre émission et intromission. Ce qui est clair, c'est que tous ces modèles étaient hautement spéculatifs compte tenu des connaissances scientifiques dont disposaient les anciens. Pourtant, il semble incroyable de dire que les anciens n'avaient pas de métajustification pour leurs croyances perceptives, c'est-à-dire qu'ils n'étaient pas justifiés à les tenir pour des connaissances. Les anciens étaient

certainement justifiés à tenir les croyances visuelles pour des connaissances et non pas pour des croyances purement hasardeuses ; ils étaient justifiés à penser qu'il existait un lien de dépendance entre l'impression visuelle et le fait extérieur, mais cette justification était indépendante de toute observation directe de la *manière* dont les impressions visuelles dépendaient du fait observé. Nous disposons *aujourd'hui* d'une telle confirmation observationnelle pour la vision, mais les anciens n'en disposaient pas. Aussi les anciens étaient-ils réduits à faire diverses hypothèses spéculatives, divers modèles. Ces modèles étaient-ils sans intérêt et sans justification parce que purement spéculatifs ? Assurément non : les anciens étaient justifiés à présumer que *l'un ou l'autre* de ces modèles était conforme à la réalité, et à espérer que cette construction de modèle pourrait servir à une éventuelle confirmation empirique future. Et c'est bien ce qui s'est passé. De la même façon, nous avons aujourd'hui une justification à croire que notre expérience du libre arbitre dépend *d'une manière ou d'une autre* du fait métaphysique du libre arbitre, et bien que nous n'en ayons pas (encore) de confirmation empirique, nous sommes justifiés à construire des hypothèses spéculatives, des modèles dont *l'un ou l'autre* devrait bien être conforme à la réalité (pour autant que nous soyons justifiés à le présumer), et à espérer que ces modèles théoriques pourront guider et alimenter une éventuelle confirmation empirique future.

Dans ce chapitre, j'ai essayé de répondre à la question suivante : si nous avons une justification *prima facie* à croire que notre expérience du libre arbitre nous procure une véritable *connaissance* (une croyance *fiable* du point de vue externaliste), cette justification *prima facie* est-elle réfutée, soit par la faiblesse du support épistémique, soit par la « lourdeur » de l'information qu'il est censé nous apporter ? Et j'ai tâché de montrer que cette justification *prima facie* (cette *méta*justification *prima facie*) n'était pas réfutée. Du côté du support épistémique, nous n'avons pas de raison de penser que notre expérience agentive serait systématiquement non fiable, même dans ses contenus « légers » : les données des neurosciences et de la psychologie montrent, au pire, que notre expérience est faillible, mais une épistémologie qui exigerait l'infailibilité des expériences mène au scepticisme. Du côté du contenu, il existe un argument plus grave à mon sens, l'argument de réfutation fondationnelle : il se pourrait que la découverte de l'incompatibilisme de notre expérience agentive (chapitre 4) constitue une réfutation de deuxième ordre pour notre croyance au libre arbitre, s'il s'avérait qu'un contenu aussi lourd ne peut pas être fiablement indiqué par une simple expérience phénoménale. Mais j'ai montré que notre expérience phénoménale pouvait en réalité dépendre fiablement de l'indéterminisme libertarien, aussi bien lorsque celui-ci est décrit de manière événementielle que lorsqu'il est décrit selon le modèle de la causalité agentive. Par conséquent, le fait que notre expérience du libre arbitre ait un contenu relativement « lourd » (qui engage à l'indéterminisme causal) n'empêche pas que nous soyons justifiés à continuer de la tenir pour une source fiable de connaissance. Autrement dit, notre croyance introspective au libre arbitre n'est pas réfutée fondationnellement ; nous n'avons pas de raison de penser que son fondement introspectif est incapable de la supporter épistémiquement.

Ainsi s'achève notre tour d'horizon des réfutations possibles de la justification *prima facie* par l'expérience du libre arbitre. Après avoir montré l'absence de réfutation directe (chapitre 5), nous venons de voir que les réfutations fondationnelles n'étaient pas convaincantes non plus. L'information indéterministe délivrée par notre expérience du libre arbitre est donc non seulement *recevable*, mais en outre nous n'avons aucune raison de contester la fiabilité de la source qui délivre cette information. En l'absence de réfutation, la justification *prima facie* devient donc justification *ultima facie*, non seulement pour « le libre arbitre » au sens indexical, non analysé, mais aussi pour le contenu analysé au chapitre 4, c'est-à-dire qu'en définitive, nous pouvons dire que nous sommes *ultima facie* justifiés à

croire à l'indéterminisme libertarien sur la base de notre expérience commune du libre arbitre. Dire qu'il n'y a pas de réfutation valable aujourd'hui ne préjuge pas, évidemment, d'éventuelles réfutations futures : rien n'exclut que des données de neurophysiologie viennent un jour démontrer une forme de court-circuitage de la conscience lors du choix ; rien n'exclut non plus que la discussion philosophique vienne renverser de manière définitive la possibilité des modèles du libre arbitre ou des modèles d'intégration proposés dans cette étude. Mais en l'état actuel de nos connaissances et de la discussion philosophique, l'attitude la mieux justifiée est de conserver notre confiance dans l'expérience commune du libre arbitre et d'accepter ses conséquences indéterministes.

Conclusion : le libertarianisme, à l'égal des croyances perceptives

Puisque les diverses genèses que l'on a tenté de faire de l'idée de liberté ne sont pas valables, nous admettons comme juste la preuve directe de notre liberté tirée de l'idée que nous en avons. (Durkheim 2002, cours XXXV)

L'étude qui s'achève peut être lue de deux manières différentes. On peut la lire tout d'abord comme un seul argument général dont le but est de défendre épistémologiquement une position particulière dans le débat sur le libre arbitre, à savoir la position libertarienne. La conclusion de cet argument général est que nous sommes pleinement justifiés à croire au libertarianisme, et pleinement justifiés à soutenir qu'il s'agit d'une *connaissance*. Et la démarche de l'argument consiste à montrer que la croyance introspective au libre arbitre peut satisfaire les mêmes réquisits épistémologiques qui sont satisfaits par les croyances perceptives. Je reviendrai un peu plus loin sur cette stratégie de comparaison avec la perception, pour en résumer les éléments essentiels et voir si elle est ultimement convaincante.

Mais il y a une autre manière de lire cette étude : non pas tant comme la défense d'une position particulière dans le débat sur le libre arbitre qu'un effort pour structurer l'ensemble du débat et pour « situer » à leur juste place dialectique les différents désaccords qui donnent lieu aux différentes positions. C'est cet effort qui a donné lieu à un plan inhabituel, et pour ainsi dire « à l'envers », où la « question d'existence » précède la « question de compatibilité », et où la « question épistémologique » constitue le cœur du problème. Je me suis efforcé, en effet, de construire ma démarche comme une ascension dans laquelle les premières étapes sont (ou devraient être) largement consensuelles, et les étapes suivantes de plus en plus controversées, de telle sorte que tous les participants de la discussion puissent me suivre jusqu'à l'étape qui constitue le nœud de leur désaccord, et où ils adopteront alors la plus proche solution de repli. C'est dans cet esprit que je vais résumer à présent les grandes étapes de mon raisonnement, en commentant plus particulièrement la « solidité » de chaque étape (telle que je l'évalue) et les solutions de repli que j'envisage.

I. Camp de base et positions de repli

Dans la première partie de cette étude, je me suis efforcé d'établir le « point de départ épistémologique » du problème du libre arbitre – c'est-à-dire la thèse selon laquelle, avant toute investigation spécialisée (philosophique ou empirique), nous sommes justifiés à nous fier à notre expérience du libre arbitre comme à une véritable source de connaissance. C'est seulement à partir de ce point de départ que l'on peut, à mon sens, comprendre et ordonner les grandes thèses possibles sur le libre arbitre. Et cette conclusion générale de la première partie me semble extrêmement forte dialectiquement – notamment parce que je me suis efforcé de lui donner une portée aussi prudente et modeste que possible.⁴³³ Mais il sera utile

⁴³³ Elle est modeste pour au moins trois raisons : 1) la justification défendue à ce stade est seulement une justification *prima facie* pouvant faire l'objet, en deuxième partie, de toutes sortes d'objections, 2) elle justifie un contenu « indexical » ou non analysé et n'engage donc à aucune théorie particulière sur le libre arbitre, enfin 3)

cependant de rappeler les arguments en faveur de cette conclusion et d'envisager leur échec éventuel. La conclusion de la première partie repose essentiellement sur cinq étapes :

- 1) le sens de « libre arbitre » est fixé par une expérience doxastique commune à propos de nous-mêmes et de nos actions
- 2) nous avons l'expérience E d'avoir le libre arbitre
- 3) cette expérience E nous justifie *prima facie* à croire que nous sommes libres
- 4) il existe des expériences que nous sommes justifiés *prima facie* à tenir pour des sources de connaissances
- 5) l'expérience E d'avoir le libre arbitre fait partie de ces expériences que nous sommes justifiés *prima facie* à tenir pour des sources de connaissances.

Le chapitre était consacré aux thèses 1 et 2, c'est-à-dire au « problème de la définition » et à la « question de l'expérience ». ⁴³⁴ L'originalité de ce chapitre est de traiter ces deux problèmes conjointement, la solution du premier donnant immédiatement la solution du second : (1) le concept de « libre arbitre » doit être défini par référence au conflit intuitif entre le concept de déterminisme et notre conception spontanée de nous-mêmes et de nos actions ; et de là il s'ensuit logiquement que (2) nous avons une tendance spontanée à nous représenter comme libres (ou une « expérience doxastique » d'être libres), celle-là même qui entre en conflit avec l'idée de déterminisme. Ma solution au problème de la définition n'est pas classique, et pourrait donc susciter des désaccords. Encore faut-il être au clair sur l'étape du raisonnement qui pourrait être refusée. La première étape consiste à rejeter les solutions « confiante » et « stipulative » au problème de la définition, et à définir le libre arbitre de manière « descriptive » par référence à la discussion traditionnelle. Cette étape me semble extrêmement difficile à refuser une fois que le problème est posé explicitement : de fait, il me semble que même les auteurs qui commencent leur discussion du libre arbitre par une définition stipulative – telle que le « pouvoir d'agir autrement » – ont généralement l'ambition plus ou moins explicite de saisir par là le sens qui est en jeu dans la discussion traditionnelle, celui qui a donné naissance au « problème du libre arbitre ». La deuxième étape du raisonnement consiste à dire que la *bonne* solution descriptive définit le libre arbitre par un « conflit intuitif originaire » entre le concept de déterminisme et un certain concept que nous avons de nous-mêmes et de nos actions. Cette deuxième étape pourrait être contestée de deux manières : soit en niant l'existence de tout conflit originaire intuitif entre le déterminisme et notre conception de nous-mêmes, soit en niant que ce conflit originaire (existant) soit la bonne solution descriptive pour définir le concept de libre arbitre. La première réaction me semble extrêmement difficile à adopter aujourd'hui, au vu des résultats de (Nichols and Knobe 2007) ⁴³⁵ : il y a bel et bien un conflit intuitif naturel avec l'idée de déterminisme. On pourrait alors nier que ce conflit intuitif soit à l'origine du problème traditionnel du libre arbitre – le concept de libre arbitre ne serait pas celui qui donne lieu à ce conflit, mais plutôt (par exemple) le concept signifié linguistiquement par l'expression « libre arbitre » (solution descriptive sémantique). Là encore, cette réaction me semble très peu plausible : puisqu'il y a un conflit naturel entre le déterminisme et un certain concept de nous-mêmes, il n'y a pas lieu de chercher ailleurs que dans ce concept lui-même l'origine du problème philosophique millénaire. Il se peut que ce concept ait quelque chose à voir avec l'expression linguistique

elle repose sur l'expérience « doxastique » du libre arbitre, sans se prononcer sur l'origine et la nature de cette expérience (*de se* ou *de altero*, phénoménale agentive ou morale).

⁴³⁴ Du moins une partie de la « question de l'expérience », celle qui consiste à savoir si nous avons oui ou non une expérience du libre arbitre. L'autre partie de la question concerne évidemment le contenu de l'expérience, et ce point est abordé dans le chapitre 4.

⁴³⁵ En particulier le fait que nous avons une tendance naturelle à juger que notre monde ressemble davantage à un monde indéterministe qu'à un monde déterministe (question d'actualité, voir p.66).

« libre arbitre » (il se peut qu'il soit *désigné* par cette expression), mais il semble bien qu'il soit *indépendant* de toute formulation proprement linguistique. Quoi qu'il en soit, quand bien même le « problème traditionnel » du libre arbitre aurait plus à voir avec la sémantique de telle ou telle expression linguistique qu'avec le concept que nous avons spontanément de nous-mêmes, cela n'affecterait pas le reste de mon étude. En effet, je choiserais alors de délaisser ce problème traditionnel pour m'intéresser spécifiquement au concept que nous avons de nous-mêmes, pour savoir si ce concept est illusoire ou véridique ; ce problème (savoir si nous vivons ou non dans une illusion généralisée) me paraît être, en lui-même, le plus intéressant et le plus important. Et pour pouvoir aborder ce problème, je choiserais de stipuler que l'expression « libre arbitre » désigne ce concept que nous avons de nous-mêmes et qui entre en conflit avec l'idée de déterminisme. Si un tel conflit n'existait pas, il n'y aurait pas, à mon sens, de problème pressant du libre arbitre. Mais puisque'il existe (comme on vient de le rappeler), c'est lui qu'il convient d'étudier, et il permet d'être sûr que (2) nous avons une tendance naturelle à concevoir (une expérience doxastique) que nous sommes libres.

Dans le chapitre 2, je me suis efforcé de montrer que (3) cette expérience doxastique apportait une justification *prima facie* en faveur de son contenu. C'est-à-dire que j'ai défendu un modèle « introspectionniste » de notre justification à croire au libre arbitre. J'ai commencé par montrer qu'il n'y avait aucune bonne raison de refuser à l'expérience du libre arbitre le caractère justifiant, en rejetant trois objections « anti-introspection ». On pourrait être tenté de reprocher à notre expérience du libre arbitre (i) de n'être qu'une impression, (ii) d'être faillible, ou (iii) d'avoir une conséquence « lourde » (le fait que nous ne sommes pas manipulés par des dispositifs martiens). Chacune de ces trois objections est extrêmement faible dialectiquement, car elle s'applique aussi bien aux expériences perceptives qu'à l'expérience du libre arbitre. Si donc l'expérience perceptive peut apporter une justification, alors l'expérience du libre arbitre le pourra aussi. Dans ce chapitre, je ne prétends pas avoir répondu de manière définitive au scepticisme à propos de la perception. Certains lecteurs seront donc peut-être tentés de suivre cette voie, en rejetant du même coup le caractère justifiant de l'expérience du libre arbitre. Mais cela ne porterait pas préjudice à ma démarche d'ensemble qui consiste à montrer, comparativement, que l'expérience du libre arbitre n'est *pas moins crédible* que la perception sensible (voir section suivante de la conclusion). D'autres auteurs pourraient être tentés d'adopter des solutions anti-sceptiques différentes de la théorie de la justification *prima facie* fondationnaliste ou « par expériences », par exemple une solution cohérentiste, ou infinitiste. Cela non plus ne porterait pas préjudice à ma démarche, pour autant que je puisse en juger : j'ai choisi la terminologie fondationnaliste dans cette étude principalement parce qu'elle me semble la plus adaptée pour formuler l'ensemble des objections épistémologiques qu'on peut faire au libre arbitre (notamment celles des chapitres 3 et 6) ; si l'on adopte une conception infinitiste ou cohérentiste, ces objections soit disparaîtraient totalement, soit pourraient simplement être traduites dans des termes différents et rejetées en traduisant dans ces nouveaux termes les solutions que j'apporte en termes fondationnalistes. Je ne vois pas d'objection à l'introspectionnisme du libre arbitre qui serait proprement infinitiste ou proprement cohérentiste. Si l'on met de côté ces cas du scepticisme, du cohérentisme et de l'infinitisme, il me semble extrêmement difficile de rejeter l'idée selon laquelle notre expérience naturelle du libre arbitre nous donne une justification *prima facie* en faveur de son contenu « indexical » (c'est-à-dire non analysé). Notez que cette conclusion est extrêmement modeste : une justification qui est simplement *prima facie* est susceptible d'être réfutée par toutes sortes de facteurs, notamment par une simple analyse du contenu de notre expérience. Supposons néanmoins que quelqu'un refuse cette conclusion modeste et fortement argumentée ; supposons que notre expérience du libre arbitre n'apporte aucune justification *prima facie* en faveur de son contenu. Quelle serait la conclusion à en tirer ? La première conclusion, c'est que, pour affirmer l'existence du libre arbitre, il nous faudrait

trouver une autre source d'information. C'est-à-dire qu'il nous faudrait d'abord établir que le contenu de notre expérience est le fait F, puis montrer que le fait F est indiqué par une source d'information différente de notre expérience de sens commun, probablement par des données empiriques. Malheureusement, il me semble que quelle que soit l'analyse qu'on donne du contenu de notre expérience E, même une analyse compatibiliste, nous n'avons pas des données empiriques qui permettent de montrer de manière directe que nous satisfaisons ces conditions. Autrement dit, si l'on rejetait l'introspectionnisme a-compatibiliste, il me semble qu'on serait conduit naturellement au scepticisme (a-compatibiliste) à propos du libre arbitre (« rien ne permet de dire si nous satisfaisons ou non les conditions du libre arbitre »). Mais encore une fois, cette réaction me semble fort peu plausible : à peu près tout lecteur anti-sceptique devrait accepter que notre expérience du libre arbitre donne au moins une justification *prima facie* en faveur de son contenu avant toute analyse – quoi qu'il advienne ensuite de cette justification *prima facie* au gré des objections.

Dans le chapitre 3, j'ai tâché de montrer que notre expérience du libre arbitre nous donnait en outre une « méta-justification » *prima facie*, c'est-à-dire que nous étions justifiés *prima facie* à accepter la proposition de second ordre « nous savons que nous sommes libres ». Cette conclusion nous a amenés à approfondir nos remarques en épistémologie fondamentale, en étudiant notamment la question de savoir si et comment nous pouvons considérer nos expériences perceptives comme des sources de *connaissance*, afin de satisfaire un principe de méta-cohérence. Le principe de méta-cohérence exige en effet que nous soyons en mesure de considérer comme connaissances (comme non hasardeuses) les croyances que nous affirmons, à défaut de quoi nos croyances perdent leur justification. La meilleure solution à ce problème consiste à dire que (4) il existe des expériences que nous sommes justifiés basiquement et *prima facie* à tenir pour des sources de connaissances (sans avoir d'indication directe sur la manière dont elles apportent une connaissance, ou sur le type de lien de dépendance entre fait et croyance). C'est la solution « dogmatiste » qui a pris de plus en plus d'importance dans la discussion épistémologique contemporaine. Cette position est-elle indiscutable ? Non, bien sûr, mais on peut faire ici le même genre de remarques que dans le chapitre précédent, à savoir que les réponses alternatives à ce problème (scepticisme radical, méta-cohérentisme, méta-infinitisme) ne constitueraient pas un problème spécifique contre l'expérience du libre arbitre, et ne porteraient donc pas préjudice à mon argument. Ce qui pourrait poser problème, ce serait d'accepter que certaines expériences sont dotées d'une méta-justification basique *prima facie*, mais que l'expérience du libre arbitre ne fait pas partie de ces expériences privilégiées. Mais là encore, tant qu'on n'a pas analysé le contenu de notre expérience du libre arbitre, il me semble qu'il n'y a pas de bonne raison de lui refuser ce statut, parce que cette expérience a toutes les caractéristiques (force intuitive, naturalité, expérience mooréenne) qui ont été avancées pour expliquer la méta-justification basique de la perception. Cette conclusion me semble très forte, mais c'est probablement la plus contestable de la première partie : on pourrait tenter de défendre que les expériences perceptives ont d'autres caractéristiques qui permettent de rendre compte de leur méta-justification, des caractéristiques qui font défaut à l'expérience du libre arbitre. J'aborderai cette question dans la section suivante. Pour l'instant, je ferai simplement deux remarques. Premièrement, je rappellerai que cette thèse d'une méta-justification basique *prima facie* du libre arbitre reste une thèse extrêmement modeste puisqu'elle est compatible avec toutes sortes de réfutations ; si donc on soupçonne que le libre arbitre ne pourra pas être méta-justifié *ultima facie*, on ne doit pas craindre d'accepter tout de même cette thèse modeste, car les objections qu'on aurait à la méta-justification seront très probablement exprimables (en deuxième partie) comme réfuteurs de la méta-justification *prima facie*. Deuxièmement, s'il s'avérait malgré tout que l'expérience du libre arbitre n'a *même pas* de méta-justification basique *prima facie*, la thèse que je défendrais dans ce cas serait le scepticisme. On pourrait être tenté de conserver la

croyance au libre arbitre en négligeant l'absence de méta-justification, par exemple en disant ceci : « tant que nous n'avons pas de preuve que notre expérience *ne peut pas* constituer une connaissance (chapitre 6), je choisirai de lui faire confiance, même si je n'ai pas de raison positive de la considérer comme une connaissance ». Cette réaction reviendrait à abandonner le principe de méta-cohérence positive pour se reporter sur un principe de méta-cohérence négative. Mais cette solution de repli n'est pas pour moi une option acceptable : pour défendre une thèse de manière justifiée, il faut être prêt à défendre *positivement* qu'il s'agit d'une connaissance. Aussi, si nous n'avons pas de méta-justification pour l'expérience du libre arbitre (et la seule méta-justification que nous puissions avoir est une méta-justification basique issue de l'épistémologie fondamentale, pas une méta-justification empirique), alors il nous faudrait abandonner notre croyance introspective au libre arbitre ; et, comme je l'ai dit dans le paragraphe précédent, si nous abandonnons l'introspection comme source d'information, il est douteux que nous ayons d'autres sources d'information sur la même question, donc ultimement il faudrait douter de l'existence du libre arbitre.

Pour résumer, le « point de départ épistémologique » selon lequel notre expérience du libre arbitre est à la fois justifiée et méta-justifiée *prima facie* me semble extrêmement solide et ne devrait être refusé que par peu de philosophes, principalement des philosophes sceptiques radicaux. Mais si l'on refusait tout de même ces thèses de la première partie, la seule option, à mon sens, serait de devenir sceptique à propos du libre arbitre.

Les positions les plus sérieuses à propos du libre arbitre sont, à mon sens, celles qui acceptent le point de départ épistémologique, mais qui divergeront à propos des questions abordées dans la deuxième partie.

Dans le chapitre 4, j'ai étudié la question du compatibilisme ou de l'incompatibilisme de notre expérience du libre arbitre et je suis arrivé à une conclusion incompatibiliste. Pour évaluer la solidité de cette conclusion, il est important de se rappeler que, là encore, la conclusion visée était de portée modeste : je n'ai pas cherché à rejeter le compatibilisme *en général*, mais seulement le « compatibilisme indépendamment motivé » par opposition au « compatibilisme d'ajustement ». C'est-à-dire que j'ai essayé de montrer que l'incompatibilisme est la meilleure hypothèse *si* on considère cette question de compatibilité ou d'incompatibilité *indépendamment* de la question d'existence (pouvons-nous réellement avoir ce type de liberté) et de la question d'intégration (pouvons-nous connaître ce type de liberté). A mon sens, le compatibilisme ne peut être sérieusement défendu que par « ajustement », pour sauver l'idée de libre arbitre face à une réfutation directe ou fondationnelle du libertarianisme. Le compatibilisme en général est à mes yeux une option très sérieuse dans le débat sur le libre arbitre. L'originalité de ma démarche est de montrer que cette position n'est très sérieuse que *pour un certain type de raisons* et que ces raisons ne résident pas, en fait, dans la question de compatibilité elle-même, mais dans une théorie générale du libre arbitre (qui inclut les traitements de la question d'existence et de la question d'intégration). Ce rejet du « compatibilisme indépendamment motivé » est-il irréfutable ? Sans doute pas ; je vais revenir à l'instant sur les arguments présentés dans le chapitre 4. Mais avant tout, je voudrais souligner que personne, à ma connaissance, n'a cherché à défendre *spécifiquement* le compatibilisme indépendamment motivé ; au contraire, plusieurs grands compatibilistes (en particulier John M. Fischer ou Terry Horgan comme on l'a vu) font voir explicitement que leur motivation en faveur du compatibilisme repose sur leur intention de préserver l'existence ou la connaissabilité du libre arbitre.⁴³⁶ En creux, cela signifie que leur motivation ne repose *pas* sur la considération isolée de la question de compatibilité.⁴³⁷ Par

⁴³⁶ Pour les motivations d'intégration de John Fischer voir p.10 sqq ; pour celles de Horgan voir p.340 sqq.

⁴³⁷ Au contraire, on peut rappeler ici la déclaration de Horgan déjà citée au chapitre 4 : « Je reconnais que lorsque nous examinons introspectivement notre phénoménologie de l'action libre, avec son aspect

conséquent, il me semble que bien peu d'auteurs voudraient vraiment contester la conclusion de ce chapitre 4, à savoir que la meilleure théorie de notre expérience du libre arbitre (si on considère cette expérience pour elle-même et sans se soucier de sa véridicité ou de son statut épistémologique) est qu'elle a un contenu incompatibiliste. Peut-être y aura-t-il certains « compatibilistes indépendamment motivés » qui quitteront la discussion à ce stade, mais je soupçonne qu'il y en aura en fait très peu si l'ensemble de la situation dialectique est bien comprise.

Voyons cependant les arguments sur lesquels repose ce statut privilégié *prima facie* de l'incompatibilisme. Deux arguments assez différents ont été proposés : un argument abductif (l'incompatibilité conceptuelle entre libre arbitre et déterminisme est la meilleure explication du conflit intuitif originaire), et un argument direct (l'examen phénoménologique de notre expérience du libre arbitre révèle un contenu incompatibiliste). De ces deux arguments, le premier est le moins « confortable » intellectuellement, parce qu'il demeure neutre et sans engagement sur le contenu exact de notre expérience du libre arbitre. Mais cette absence d'engagement est aussi ce qui fait sa force dialectique : le second argument, pour pouvoir étudier le contenu de l'expérience du libre arbitre, est obligé d'adopter une certaine théorie sur la nature de cette expérience, une théorie qui pourrait être contestée. Dialectiquement, je crois que l'argument abductif est tout à fait suffisant pour montrer le statut privilégié *prima facie* de l'incompatibilisme. En effet, pour rejeter l'argument abductif, il faudrait avancer une explication du conflit intuitif originaire qui soit plus simple que l'explication par l'incompatibilité conceptuelle, et qui fasse reposer le conflit intuitif sur une illusion ou sur une inférence fallacieuse. Or je crois avoir montré qu'aucune des théories de l'illusion n'était bien motivée. Plus précisément, je me suis efforcé de montrer que les inférences prétendument fallacieuses auxquelles ces théories ont recours ont au moins autant de chances d'être des inférences tout à fait valides appuyées sur des principes intuitivement très forts.⁴³⁸ C'est ici, dans ma dialectique, qu'apparaissent les grands arguments incompatibilistes de la discussion contemporaine. Il me semble que ces arguments sont beaucoup plus forts dialectiquement lorsqu'on les prend comme réfutations de théories de l'illusion, que lorsqu'on les prend comme preuves directes de l'incompatibilisme. Par exemple, pour l'argument de la conséquence, il existe encore des débats très complexes pour savoir si la règle (β) est un principe valide, si l'inévitabilité d'un antécédent (le passé lointain) se « transfère » au conséquent (mon action future) via une implication elle-même inévitable (les lois de la nature). Ces débats semblent aujourd'hui difficiles à trancher. Mais le simple fait que (β) soit *possiblement* vraie suffit à montrer, dialectiquement, que la théorie de la confusion NCNC est mal motivée : pour que cette théorie soit motivée, il faudrait avoir établi que l'inférence qu'elle nous attribue est fallacieuse ; or cette inférence pourrait fort bien être parfaitement valide si (β) est une règle valide. Il me semble donc que l'argument abductif, malgré sa complexité dialectique, est extrêmement fort : si l'on met de côté la question d'existence et la question d'intégration, nous n'avons aucune raison motivée de tenir le conflit intuitif pour un conflit illusoire.

présentationnel du soi comme source qui inclut l'aspect de la liberté [des alternatives] comme composant essentiel, et lorsque nous nous demandons en même temps si la véridicalité de cette phénoménologie requiert [...] le libertarianisme, nous sentons une *certaine* tendance à juger que la réponse à cette question est Oui. » (Horgan 2011a, 94)

⁴³⁸ A la théorie de l'erreur de catégorie de la volonté libre, j'ai montré qu'on pouvait faire correspondre le réquisit tout à fait défendable du contrôle des volitions (argument de Lehrer). Pour la théorie de la confusion entre nécessité de la conséquence et nécessité du conséquent, j'ai montré que l'inférence réalisée pouvait fort bien reposer en réalité sur un principe de transfert de l'inévitabilité (argument de van Inwagen). Enfin, pour la théorie de la confusion entre déterminisme et diverses sortes de manipulation, j'ai montré que l'assimilation entre déterminisme et manipulation pouvait fort bien reposer sur une intuition très forte d'absence de différence pertinente, et non pas sur une simple confusion (argument de Pereboom).

L'argument phénoménologique permet de mieux comprendre d'où vient exactement le conflit, s'il n'est pas illusoire. Cet argument comprend en fait deux étapes : tout d'abord l'idée selon laquelle notre expérience (doxastique) du libre arbitre vient, fondamentalement, de notre expérience *agentive* en première personne (et non pas de notre schème moral ou de notre perception des autres agents) ; deuxièmement, l'analyse phénoménologique de notre expérience agentive comme ayant un contenu incompatibiliste. La deuxième étape a fait l'objet de diverses contestations ; pourtant elle me semble aujourd'hui très difficilement réfutable, et c'est surtout la première étape qui, personnellement, me semble attaquable. Commençons donc par la seconde étape : j'estime que les résultats de (Deery, Bedke, and Nichols 2013) ont montré de manière vraiment convaincante que notre expérience agentive contenait une composante d'expérience des alternatives qui est naturellement et spontanément interprétée comme étant incompatible avec le déterminisme (bien compris). Cette donnée de « l'introspection commune » pourrait être remise en cause par des considérations d'existence ou d'intégration du libre arbitre mais si l'on s'en tient simplement à la question de compatibilité pour elle-même, c'est certainement la meilleure donnée que nous ayons sur la question. Si nous avons une expérience agentive incompatibiliste, cela veut-il dire que « le libre arbitre » est incompatibiliste ? Pas tout à fait puisque le concept de libre arbitre pourrait avoir une autre origine que cette expérience agentive, et notamment notre schème conceptuel moral. J'ai présenté trois arguments pour défendre que notre concept de libre arbitre vient de l'expérience agentive. *L'argument de la force* permet sans doute de montrer que notre expérience du libre arbitre n'est pas une expérience « scotiste » (par perception des autres agents), mais il laisse ouverte la possibilité de l'origine morale du concept. *L'argument de la disponibilité* permet de montrer que notre concept de libre arbitre (celui qui entre en conflit avec le déterminisme) est certainement constitué *en partie* de notre concept expérientiel agentif, puisqu'on vient de voir que ce concept agentif avait toutes les chances de produire un conflit avec le déterminisme... mais cela n'exclut pas une *double* origine : supposons, par exemple, que notre expérience agentive nous donne l'expérience des alternatives (qui est incompatibiliste) et que notre schème moral nous donne la conviction d'être la source ultime de nos actions (qui est également incompatibiliste). Il me semble qu'à ce stade de la discussion, un modèle double est en fait tout à fait défendable. La *seule* raison pour laquelle je le rejette est une raison d'intégration (c'est là mon troisième et dernier argument) : il me semble que si l'on partait avec cette hypothèse d'un double fondement de notre croyance au libre arbitre indéterministe, le fondement « moral » serait ultimement réfuté fondationnellement (au chapitre 6) parce que l'expérience doxastique morale n'est pas capable de répondre au défi d'Intégration.⁴³⁹ On se trouverait alors avec une demi-réfutation

⁴³⁹ Je n'ai pas développé, dans cette étude, mes raisons de penser que l'épistémologie « morale » du libre arbitre est réfutée fondationnellement. Ces raisons tiennent essentiellement au fait que l'expérience doxastique morale (l'impression que nous sommes libres fondée sur notre schème conceptuel moral) n'est pas une expérience « occurrentielle » dans le sens défini au chapitre 4, p.310. Sans cette caractéristique d'occurentialité, il est beaucoup plus difficile de satisfaire la contrainte épistémologique de traçage ou de sensibilité. La seule solution d'intégration que je vois pour l'épistémologie morale du libre arbitre serait d'avoir recours à une forme de garantie divine, telle que celle employée par Descartes (pour toutes les croyances) ou par Plantinga (pour la croyance en Dieu). On pourrait défendre en effet la dépendance contrefactuelle : « si nous n'étions pas indéterminés nous n'aurions pas un schème conceptuel moral nous portant à croire que nous sommes indéterminés ». Ce conditionnel serait vérifié de la manière suivante : dans le monde le plus proche où nous ne sommes pas indéterminés, nous ne sommes pas moralement responsables, et donc c'est un monde où Dieu n'a pas voulu nous rendre moralement responsables ; mais probablement, dans ce monde, Dieu n'a pas voulu non plus que nous soyons systématiquement trompés sur notre responsabilité morale, donc il ne nous a pas non plus équipés d'un schème conceptuel moral nous donnant l'illusion d'être indéterminés. Notre schème moral dépendrait ainsi de l'indéterminisme naturel *via* Dieu, qui est l'auteur et l'origine de l'un et de l'autre, mais *d'aucune autre manière plus directe*. Si cette solution théiste est la *seule* solution d'intégration de la croyance morale au libre arbitre (et si notre croyance au libre arbitre était véritablement d'origine morale), cela voudrait

fondationnelle : une partie de notre concept de libre arbitre (la partie phénoménale agentive) serait confirmée et étayée, tandis que l'autre partie (morale) serait réfutée fondationnellement. Cette demi-réfutation n'aurait pas de conséquence directe sur la question de l'indéterminisme (puisque la partie agentive *suffit* à établir cette conséquence), mais elle reviendrait tout de même à dire qu'une partie de notre conception de sens commun de nous-mêmes et de nos actions est fondamentalement illusoire. C'est pour cette raison d'intégration (et *non pas* pour des raisons indépendantes) que je suis tenté, personnellement, de considérer que notre croyance au libre arbitre ne repose absolument pas sur notre schème conceptuel moral. Mais je reconnais que cette conclusion est sujette à caution, et si l'on me persuadait que le schème moral est à l'origine, au moins partiellement, de notre croyance au libre arbitre, ma solution de repli serait de dire que notre expérience agentive est certainement *aussi* à l'origine d'une partie de notre concept de libre arbitre, et que cette expérience agentive *suffit* à nous justifier à croire au libertarianisme.

C'est à partir du chapitre 5 qu'apparaissent, à mon sens, les désaccords les plus fondés et les plus importants dans le débat sur le libre arbitre. C'est-à-dire que presque tous les participants au débat reconnaissent ou devraient reconnaître à la fois l'intuition épistémologique qui fait la conclusion de la première partie :

(IM) notre expérience du libre arbitre nous donne (présomptivement) une *connaissance* du libre arbitre

ainsi que l'intuition initialement (ou indépendamment) incompatibiliste qui fait la conclusion du chapitre 4 :

(IE) notre expérience du libre arbitre (si on la considère en elle-même) a toutes les chances d'avoir des conditions de vérité incompatibilistes

Et il me semble, par conséquent, que les vrais désaccords concernant le libre arbitre sont des désaccords concernant l'éventuelle *réfutation* de ces deux intuitions initiales par d'autres sortes de considérations – des considérations de premier ordre (réfutation directe du contenu indéterministe – chapitre 5) ou bien des considérations de second ordre (le problème d'intégration proprement dit – chapitre 6).

Dans le chapitre 5, j'ai examiné la question de savoir si le contenu incompatibiliste de notre expérience était irrecevable parce qu'impossible ou incompatible avec nos connaissances scientifiques actuelles. En m'appuyant largement sur les réponses contemporaines à la « question d'intelligibilité », j'ai soutenu que le libre arbitre incompatibiliste n'était ni inconcevable ni incompatible avec nos connaissances scientifiques, c'est-à-dire qu'on peut construire des modèles intelligibles et scientifiquement informés qui laissent une place à une forme indéterministe de libre arbitre. Quelle est la force de ces conclusions ? On doit constater que, dans le débat contemporain, les réponses libertariennes à la question d'intelligibilité n'ont pas convaincu tout le monde. En introduction, j'ai souligné que ces réponses étaient très largement considérées comme *sérieuses* ou dialectiquement défendables (de telle sorte que le discrédit du libertarianisme vient plutôt, à mon sens, du problème épistémologique), mais cela ne veut pas dire qu'elles soient aussi largement considérées comme *vraies*, ou pleinement satisfaisantes. Je reconnais volontiers que la discussion de la question d'intelligibilité est ouverte à des désaccords raisonnables, et je ne prétends pas avoir apporté dans cette étude des nouveautés argumentatives susceptibles de clore le débat. Plus précisément, si l'on reprend les deux modèles du libertarianisme et les

dire que *seuls les théistes* peuvent continuer de croire au libre arbitre, les athées quant à eux étant réfutés fondationnellement.

quatre grandes objections (argument de la chance, argument basique, argument déterministe et argument d'intelligibilité empirique), je donnerais l'évaluation suivante. Il me semble que l'argument basique et l'argument déterministe peuvent être rejetés de manière totalement satisfaisante. Pour l'argument de la chance, j'estime qu'il s'agit d'une difficulté très sérieuse pour le libertarianisme, en particulier dans sa version du rembobinage. La réponse originale que j'apporte à cette version du problème (par analogie avec le problème épistémologique des façades de granges) ne satisfera peut-être pas tous les lecteurs. En revanche, je ne crois pas (contrairement à un certain nombre de philosophes) que l'argument de la chance donne un avantage au modèle agentif sur le modèle événementiel (en réfutant le second, mais pas le premier). La version du problème qui est censée favoriser ainsi le modèle agentif est l'argument du contrôle diminué ; mais je suis convaincu que le libertarianisme événementiel peut répondre à cette version de l'argument aussi bien que la causalité agentive (voir p.362), et cette version de l'argument ne représente donc pas, à mon sens, le cœur du problème de la chance. Si l'argument de la chance marche, c'est probablement dans sa version du rembobinage, et dans ce cas il réfute toute forme de libertarianisme. Pour ce qui est de l'argument d'intelligibilité empirique, enfin, il me semble qu'un certain nombre de personnes conserveront l'intuition que les modèles à la Kane ou même les modèles de survénance de causalité agentive sont des spéculations théoriques très peu plausibles. Cependant il est très important de distinguer dans cette réaction ce qu'elle a de proprement *métaphysique* et ce qu'elle a d'*épistémologique*. C'est-à-dire qu'il faut distinguer l'intuition « ce modèle est probablement faux parce qu'il est intrinsèquement incroyable » de l'intuition « ce modèle est probablement faux parce que nous n'avons pas la moindre indication en sa faveur ». Si l'on fait clairement cette distinction, on trouvera, je crois, qu'une majorité de personnes qui trouvent les modèles libertariens scientifiquement « incroyables » le font pour des raisons épistémologiques et non pour des raisons d'impossibilité ou du non plausibilité *intrinsèques*.⁴⁴⁰ Par conséquent, il me semble qu'il y aura assez peu de philosophes convaincus par l'argument d'intelligibilité empirique en tant que tel. En somme, si le libertarianisme est jugé impossible ou inconcevable par certains philosophes, ce devrait être essentiellement à cause de l'argument de la chance dans la version du rembobinage. Si une telle preuve est apportée, alors il existe clairement deux réactions possibles. La première est de considérer que notre justification *prima facie* en faveur du libre arbitre serait alors réfutée. C'est la solution incompatibiliste dure. La seconde consisterait à réviser l'analyse incompatibiliste du chapitre 4 en disant ceci : « si un libre arbitre indéterministe est inconcevable, alors probablement notre expérience commune n'a pas des conditions de vérité incompatibilistes ». Ce serait une première forme de « compatibilisme d'ajustement ».⁴⁴¹ A titre personnel, si l'on parvenait à me convaincre que l'argument de la chance est concluant, ma solution de repli serait cette première forme de compatibilisme d'ajustement. Mais pour l'instant, j'estime que les modèles libertariens sont concevables et scientifiquement possibles, et que nous n'avons donc pas de raison de réviser notre analyse incompatibiliste de l'expérience agentive, ni de penser que notre expérience agentive est réfutée de manière directe.

⁴⁴⁰ C'est le cas, notamment, de Bertrand Russell lorsqu'il écrit : « Perhaps in the brain the unstable equilibrium is so delicate that the difference between two possible occurrences in one atom suffices to produce macroscopic differences in the movements of muscles. And since, according to quantum physics, there are no physical laws to determine which of several possible transitions a given atom will undergo, we may imagine that, in a brain, the choice between possible transitions is determined by a psychological cause called "volition". All this is possible, but no more than possible; there is not the faintest positive reason for supposing that anything of the sort actually takes place. » (Russell 2009, 42).

⁴⁴¹ Mais j'estime que la majorité des « compatibilistes d'ajustement » appartiendront en fait à la deuxième forme, que nous allons voir dans le résumé du chapitre 6.

Dans le chapitre 6, j'envisage la question de la réfutation fondationnelle de notre expérience agentive. J'ai distingué deux types de réfutations fondationnelles : réfutation par la source faible ou réfutation par le contenu alourdi. Les réfutations par la source faible qui existent dans la littérature sont les prétendues preuves (neuroscientifiques ou psychologiques) que notre expérience agentive est purement illusoire ou systématiquement trompeuse. J'estime que les données expérimentales de Libet et de Wegner n'apportent, en réalité, aucune preuve de ce type, si on les décrit de manière philosophiquement précise et adéquate. Il existe, bien sûr, un certain nombre de personnes qui ont été convaincues par Libet (ou Wegner) que le libre arbitre est une illusion, mais il me semble qu'il y a très peu de *philosophes* (ou de personnes philosophiquement informées) qui acceptent ce genre de conclusions. Par conséquent, on peut rejeter l'idée selon laquelle notre expérience agentive serait empiriquement discréditée. S'il y a un problème fondationnel pour notre croyance au libertarianisme, ce n'est pas parce que l'expérience agentive sur laquelle elle repose serait en elle-même discréditée, mais parce qu'elle ne serait pas le bon type de source épistémique pour donner une connaissance de l'indéterminisme.

C'est là le problème d'Intégration proprement dit. Plus précisément, le problème repose sur le fait que la connaissance implique une condition « anti-hasard » : comment notre simple expérience agentive pourrait-elle dépendre de manière non hasardeuse de l'indéterminisme que postule le libertarien ? Ce problème de dépendance adéquate entre « fait libertarien » et « expérience libertarienne » prend des formes différentes selon le modèle que l'on choisit pour le fait libertarien. Pour le modèle événementiel, c'est un problème « d'hétérogénéité des niveaux ontologiques » – l'expérience libertarienne étant au niveau macroscopique du cerveau tandis que le fait libertarien événementiel est constitué (en partie) d'un indéterminisme au niveau microscopique des particules élémentaires. J'ai répondu à ce problème en utilisant la distinction, aujourd'hui fondamentale en épistémologie, entre condition de sensibilité et condition de sûreté : une expérience doit être sensible à son contenu propre, mais il suffit qu'elle soit *sûre* relativement à ses conséquences logiques. Or s'il est difficile pour l'expérience libertarienne d'être sensible à quelque chose qui ne se situe pas à son niveau ontologique (l'indéterminisme quantique), elle peut très bien être sensible à une certaine structure neuronale chaotique qui se situe à son niveau ontologique, et qui à son tour est corrélée de manière *sûre* (et non sensible) à l'indéterminisme quantique. Pour ce qui est du modèle de la causalité agentive, le problème prend une forme très différente : le problème est ici celui de la perception du non actuel – comment percevoir un pouvoir s'il est simplement de l'ordre du possible ? De nouveau, la solution utilise la distinction entre sensibilité et sûreté. Si les potentialités du pouvoir agentif sont corrélées à une certaine base catégorique actuelle, alors il est possible de percevoir (de manière sensible) cette base catégorique, qui à son tour sera corrélée de manière sûre à une certaine potentialité. Quelle est la force dialectique de ces réponses ? Il me semble qu'elles laissent une certaine place au désaccord raisonnable. Certaines personnes estimeront peut-être que ces modèles d'intégration ne satisfont pas réellement les contraintes épistémiques qui pèsent sur la perception sensible. Je reviendrai sur cette question dans la section suivante. Mais ces réponses me semblent néanmoins très fortes pour deux raisons. Premièrement, ces deux réponses sont indépendantes l'une de l'autre, de telle sorte que si l'on considère que l'une d'elles n'est pas satisfaisante (par exemple l'intégration du libertarianisme événementiel), on n'a pas par là réfuté fondationnellement le libertarianisme, mais seulement l'un de ses modèles, et la réaction à adopter sera alors de se reporter sur l'autre modèle. Pour réfuter le libertarianisme, il faut considérer que *tous* les modèles libertariens rencontrent, indépendamment, une objection dirimante.⁴⁴² Deuxièmement, pour bien évaluer la force dialectique de ces modèles d'intégration, il est

⁴⁴² On pourrait défendre que l'un rencontre une objection de premier ordre et l'autre une objection d'intégration.

fondamental de bien comprendre que l'objection du « caractère spéculatif », l'objection qui sera probablement la principale résistance à ces modèles, n'est pas pertinente. L'objection du caractère spéculatif consiste à dire que nous n'avons pas le début d'une indication en faveur de la vérité de ces modèles. Mais dans la situation dialectique où nous nous trouvons (la justification *prima facie* à tenir notre expérience pour une source de connaissance), il est suffisant que ces modèles soient *possibles* pour que notre présomption de connaissance ne soit pas réfutée ; nous n'avons pas besoin d'avoir la moindre confirmation de leur réalité effective. C'est donc uniquement sur cette question, la question de la *possibilité* de ces modèles d'intégration, que devra porter un désaccord bien compris. Encore une fois, je pense qu'un tel désaccord peut être raisonnable, mais je crois avoir montré, en tout cas, que la thèse d'une *possible* intégration du libertarianisme est au moins aussi raisonnable.

Quelle serait la réaction à adopter si l'on parvenait à montrer l'impossibilité de ces modèles d'intégration ? Comme je l'ai dit, si l'un des deux modèles d'intégration échouait, la solution la plus justifiée serait de se reporter sur l'autre modèle. Mais si les deux modèles d'intégration échouaient, alors nous serions véritablement en face d'un trilemme d'intégration. On aurait alors trois solutions de repli :

- soit réviser totalement notre épistémologie : adopter des standards plus souples *dans le but* de pouvoir continuer à défendre spécifiquement la connaissance du libre arbitre – c'est ce que j'ai appelé la solution d'intégration forcée
- soit réviser notre métaphysique du libre arbitre : ce serait la deuxième forme de compatibilisme d'ajustement
- soit enfin abandonner l'idée que notre expérience agentive nous permet de connaître l'existence du libre arbitre : la position serait alors celle du « libertarianisme empirique », c'est-à-dire une position qui rejette l'information donnée par notre expérience mais préserve les résultats des chapitres 4 et 5, c'est-à-dire que le libre arbitre requiert l'indéterminisme, et qu'un libre arbitre indéterministe est quelque chose de *possible* en l'état actuel de nos connaissances.⁴⁴³

En ce qui me concerne, ma solution de repli serait très certainement le compatibilisme d'ajustement. En effet, l'intuition mooréenne selon laquelle nous connaissons le libre arbitre (quel que soit le contenu de cette expérience du libre arbitre) est pour moi très difficile à rejeter. Pour ce qui est de la solution d'intégration forcée, je n'y suis pas opposé par principe, mais mon degré de tolérance dépend ici de la nature du « forçage » : si l'intégration du libertarianisme supposait de réviser drastiquement notre épistémologie, par exemple en abandonnant totalement la contrainte anti-hasard (« puisque notre expérience de l'indéterminisme est hasardeuse, il faut bien admettre que des expériences hasardeuses peuvent être sources de *connaissance* »), alors je ne pourrais pas l'accepter ; en revanche, si l'intégration du libertarianisme supposait d'abandonner des contraintes épistémologiques plus contraignants, dont on a des raisons de penser qu'ils ne valent que pour la perception (par exemple le principe de traçage occurrentiel), je ne serais pas opposé à adapter l'épistémologie aux besoins de la connaissance libertarienne. La démarche qui serait alors adoptée est une démarche bien connue en épistémologie, celle du *particularisme* : le propre d'une démarche particulariste est de partir d'un stock de croyances que l'on considère comme archétypes de connaissance, puis de définir les standards de la connaissance à partir de ces connaissances et des conditions qu'elles peuvent satisfaire. La solution d'intégration forcée consisterait donc simplement à faire entrer la croyance au libre arbitre dans le « stock » de croyances archétypes. Une telle démarche ne me semblerait pas du tout injustifiée. Selon la nature et la radicalité de ce « particularisme libertarien », ce pourrait être une solution que j'adopterais de

⁴⁴³ Il est important de noter que le problème d'intégration, en tant que tel, ne *devrait pas* mener au déterminisme dur : seul une objection de premier ordre peut amener à une telle conclusion.

préférence au « compatibilisme d'ajustement ». Ce type de démarche, à ma connaissance, n'a jamais été adopté dans l'histoire de la philosophie, mais il me semble qu'elle aurait constitué la position de repli la plus naturelle d'un certain nombre d'introspectionnistes libertariens s'ils n'avaient été convaincus de la possibilité d'une intégration directe.

L'ensemble de cette dialectique, avec ses diverses positions de repli, est résumé dans le schéma placé en annexe 4.

Si mes arguments du chapitre 6 sont satisfaisants, alors il ne sera pas nécessaire de prendre la croyance au libre arbitre dans le stock des croyances archétypes d'une solution particulariste. Je crois au contraire qu'en prenant simplement les croyances perceptives comme croyances archétypes, on arrive à des contraintes épistémologiques que la croyance libertarienne peut satisfaire. Cette conviction, selon laquelle la croyance libertarienne et les croyances perceptives sont épistémiquement « à égalité », traverse l'ensemble des chapitres 1, 2, 3 et 6. Presque toutes mes remarques épistémologiques sont motivées par l'analogie avec le cas de la perception. Il est donc important de faire le bilan de cette démarche analogique pour voir si elle est, toutes choses considérées, véritablement satisfaisante.

II. N'y a-t-il vraiment aucune différence (pertinente) entre perception et libertarianisme ?

Pourquoi l'analogie avec la perception a-t-elle été un fil conducteur fondamental ? Principalement à cause de la « méthodologie anti-sceptique » présentée à la fin du chapitre 3 : un certain nombre de principes épistémologiques qui pourraient sembler, à première vue, intuitifs, ont la particularité de conduire au scepticisme pour les croyances perceptives. Si donc on veut éviter le scepticisme radical, on est bien obligé d'adopter une épistémologie modeste, dont j'ai essayé de prouver qu'elle avait pour conséquence de permettre à la croyance libertarienne de constituer une connaissance.

Cette méthodologie anti-sceptique peut être interprétée de deux manières différentes : une manière « permissive » et une manière « modeste ». L'interprétation permissive consiste à dire ceci : le problème sceptique est un problème extrêmement profond, qui fait apparaître la très grande fragilité de nos croyances perceptives ; de telle sorte que si l'on veut vraiment éviter le scepticisme pour la perception, on est obligé d'adopter des standards épistémologiques si permissifs que à peu près n'importe quelle croyance peut compter comme connaissance. Si l'on adoptait cette lecture, alors la conclusion générale de cette étude (que le libertarianisme peut constituer une connaissance) serait une conséquence triviale (et régionale) d'une épistémologie généralement permissive.⁴⁴⁴ Personnellement, je ne crois pas que le problème sceptique nous oblige à abandonner tout standard épistémologique digne de ce nom, de telle sorte que « anything goes ». Au cours de cette étude, j'ai eu l'occasion de défendre un certain nombre de contraintes épistémologiques substantielles, qui permettent de rejeter du domaine de la connaissance un grand nombre de croyances : principe évidentialiste (ch.2, II.1.c), principe fiabiliste (ch.2, II.2.c), principe de clôture disjonctive (ch.2, II.3.c), principe de méta-cohérence disjonctive (ch.3, II.3), principe de sensibilité des expériences et de sûreté

⁴⁴⁴ Ce genre de critique a pu être formulé contre (Plantinga 2000), qui utilise également une épistémologie largement anti-sceptique pour défendre la croyance religieuse comme connaissance. On a pu lui reprocher de tirer parti du scepticisme pour « faire passer » des croyances qui ne satisfont quasiment aucun standard épistémologique. Mais je crois que Plantinga pourrait légitimement faire la réponse que je vais faire à l'instant pour ma défense de la croyance libertarienne, à savoir que, bien qu'il ait rabaisé certaines prétentions épistémologiques démesurées, il a néanmoins conservé et mis en évidence de véritables standards épistémologiques, non triviaux et non permissifs, et montré que la croyance religieuse pouvait satisfaire ces standards non triviaux.

des conséquences (ch.3, III.3). Montrer que la croyance libertarienne peut satisfaire ces mêmes contraintes qui s'appliquent à la perception est un résultat non trivial, qui montre plutôt la force épistémologique de la croyance libertarienne que la faiblesse des croyances perceptives.

Reste néanmoins à s'assurer que la croyance libertarienne satisfait bel et bien les mêmes contraintes que les croyances perceptives. Pour être précis, mon propos n'est pas de dire qu'il n'y a *aucune* différence entre les croyances perceptives et la croyance libertarienne, mais plutôt que les différences qui existent ne sont pas épistémologiquement pertinentes. Pour faire échouer ma démarche analogique, il ne suffit pas en effet de faire apparaître une caractéristique qu'auraient les croyances perceptives et pas la croyance libertarienne, il faut encore que cette caractéristique puisse plausiblement constituer un réquisit épistémologique (soit un réquisit de justification soit un réquisit de connaissance). Or, à mon sens, lorsqu'on s'intéresse aux forces et aux faiblesses épistémologiques de ces expériences, on ne trouvera que trois types de caractéristiques :

- des points forts épistémologiques des croyances perceptives... qui sont en fait partagés par la croyance libertarienne,
- des faiblesses épistémologiques de la croyance libertarienne... qui sont en fait partagées par les croyances perceptives,
- des différences entre croyances perceptives et croyance au libre arbitre... qui ne peuvent pas plausiblement être considérées comme épistémologiquement pertinentes.

Mais voyons cela de plus près : quelle pourrait être la faiblesse propre de la croyance libertarienne ?

Dans le chapitre 2, nous avons envisagé trois « faiblesses » que l'on pourrait reprocher à la croyance introspective au libre arbitre.

Premièrement, elle repose sur « une simple impression », c'est-à-dire qu'elle ne repose pas sur un argument. Mais il est clair que cette « faiblesse » ne distingue pas la croyance libertarienne des croyances perceptives : la croyance qu'il y a une table en face de moi repose également sur l'impression visuelle d'une table en face de moi, et non pas sur un argument ou une inférence. Telle est en tout cas la conception foundationaliste de la perception sensible. Et si l'on choisit une autre conception de la perception, s'il y a en fait une inférence cachée derrière nos croyances perceptives, alors rien n'empêche de postuler également une inférence cachée dans le cas de la croyance au libre arbitre.

Deuxièmement, on pourrait reprocher à l'expérience du libre arbitre d'être une expérience faillible. La faillibilité est indéniablement une faiblesse du point de vue épistémologique. Mais il est évident que c'est une faiblesse qu'elle partage avec les expériences perceptives. Mon expérience visuelle d'une table en face de moi *pourrait* être une hallucination ou le produit dans mon cerveau des manipulations d'un neuroscientifique malintentionné.

Enfin, on pourrait reprocher à notre expérience du libre arbitre d'avoir des implications « lourdes ». On peut penser ici à deux implications : d'une part, notre expérience d'être libre implique que nous ne sommes pas manipulés par des dispositifs martiens, et d'autre part (d'après le chapitre 4), elle implique l'indéterminisme causal. Pour le premier type d'implications (qu'on peut appeler les « implications anti-sceptiques »), il est clair qu'elles ne font aucune différence avec le cas de la perception sensible⁴⁴⁵ : le contenu de mon expérience visuelle d'avoir deux mains implique que je ne suis pas manipulé par un malin génie. Le fait que nos expériences communes aient des implications anti-sceptiques est sans doute une faiblesse ou en tout cas un problème épistémologique – parce qu'il n'est pas aisé de

⁴⁴⁵ C'est ce que ne semble pas avoir aperçu van Inwagen dans son argument anti-introspection. Voir en Annexe 5 cet argument et la réponse que j'y apporte dans (Guillon 2014).

comprendre comment nous pourrions connaître la vérité de ces implications anti-sceptiques – mais ce n'est clairement pas un problème qui serait propre à l'expérience du libre arbitre.

L'implication indéterministe a plus de chance de faire apparaître une première différence entre l'expérience du libre arbitre et l'expérience perceptive. Encore faut-il montrer que l'implication lourde de l'indéterminisme n'est pas « du même ordre » que les implications lourdes anti-sceptiques. Par exemple, il est exclu de reprocher à l'expérience du libre arbitre d'avoir une implication *qui est une thèse générale sur la nature du monde extérieur* parce que les thèses anti-sceptiques (par exemple le fait que le monde n'est pas constitué d'une âme unique manipulée par un malin génie) sont *aussi des thèses générales sur la nature du monde extérieur*. Quelle est donc la particularité de la thèse indéterministe qui la distinguerait des thèses anti-sceptiques ? Probablement le fait que les hypothèses sceptiques sont de pures hypothèses théoriques, forgées de toute pièce, et sans le moindre support épistémique positif. On aurait donc là le principe d'une distinction pertinente entre expérience du libre arbitre et expérience perceptive :

(P1) Les seules implications lourdes de nos expériences perceptives sont les négations de pures hypothèses théoriques dont personne n'a jamais cherché à soutenir positivement la vérité.

(L1) L'expérience du libre arbitre a une implication lourde qui est la négation d'une hypothèse *sérieuse* (le déterminisme), une hypothèse qui dispose d'un support épistémique positif.

Pour voir si nous avons là une différence pertinente épistémologiquement, il faut encore préciser cette notion de support épistémique positif.

Prise en un sens fort, l'objection de l'implication déterministe voudrait dire que la thèse déterministe dispose d'une *preuve* positive ; l'expérience du libre arbitre aurait alors (à la différence des expériences perceptives typiques) la faiblesse épistémique d'impliquer une thèse dont la fausseté est prouvée. Mais nous avons vu au chapitre 5 que cela n'était pas vrai : aujourd'hui, pour autant que nous puissions en juger, le déterminisme universel semble plutôt scientifiquement rejeté que scientifiquement prouvé, à cause de la présence probable d'indéterminismes au niveau quantique, et nous n'avons pas non plus de preuve que le niveau macroscopique du cerveau se comporte de manière quasi déterministe. Ce n'est donc pas une « preuve » qui peut faire de l'hypothèse déterministe une « hypothèse sérieuse », et constituer ainsi une asymétrie entre expérience du libre arbitre et expérience perceptive.

Une deuxième possibilité pour comprendre en quel sens l'hypothèse déterministe est sérieuse, et donc problématique, est la suivante : on pourrait être tenté de dire que l'hypothèse du déterminisme macroscopique, même si elle n'est pas prouvée, est tout de même une « option ouverte », tandis que les hypothèses sceptiques ne seraient pas des options ouvertes (peut-être pas des options du tout). Mais que veut dire une « option ouverte » ? Cela semble vouloir dire que rien ne permet de l'exclure, que rien ne nous justifie à croire le contraire. Si c'est cela que veut dire l'objection, alors il est clair que l'objecteur fait une pétition de principe contre l'expérience du libre arbitre : si l'expérience du libre arbitre est une expérience de l'indéterminisme de l'action, alors elle apporte *elle-même* une justification au moins *prima facie* à rejeter le déterminisme de l'action. Pour réfuter cette justification *prima facie*, on ne peut pas objecter que « rien ne justifie à rejeter le déterminisme ». Si l'on considère la thèse déterministe comme une « option ouverte » en ce sens fort, c'est qu'on a *déjà* refusé tout crédit à l'expérience du libre arbitre. On pourrait l'avoir

Troisièmement, on pourrait considérer que le déterminisme est une « option ouverte » dans un sens plus faible, celui dans lequel William James parle « d'option vivante » (*live option*) : une option est ouverte, en ce sens plus faible, s'il semble envisageable que l'on soit

amené à la croire. Par exemple, bien que la communauté scientifique accepte largement le modèle standard des particules en physique élémentaire, d'autres modèles restent des « options ouvertes ». Cela ne veut pas dire que nous ne sommes pas, pour l'instant, pleinement justifiés à rejeter ces options ; cela veut simplement dire qu'il n'est pas exclu que cette justification soit un jour révisée par de nouvelles informations. On pourrait donc défendre que l'expérience du libre arbitre, à la différence des expériences perceptives, est dans cette situation : elle nous justifierait, actuellement, à rejeter l'hypothèse déterministe, mais cette hypothèse resterait « ouverte » dans le sens où des informations futures pourraient nous amener à réviser cette justification. Les expériences perceptives, quant à elles, n'auraient pas de telles implications. Que penser de cette asymétrie entre expérience du libre arbitre et expérience perceptive ? On pourrait remarquer tout d'abord que l'asymétrie est très douteuse : le contenu d'une expérience perceptive est presque toujours *révisable* par de plus amples informations. Mais on pourrait insister que, dans le cas de la perception, les hypothèses révisables sont des hypothèses « légères » (spatio-temporellement limitées) tandis que dans le cas du libre arbitre, l'expérience aurait une implication à la fois ouverte aux révisions futures *et* lourde (portant sur la totalité de l'univers physique et de ses lois). Admettons donc qu'il y ait là une certaine asymétrie.⁴⁴⁶ Mais quand bien même il y aurait là une asymétrie, il est clair dans la manière même dont je viens de la décrire qu'elle ne pose pas un problème pour l'introspectionnisme, puisqu'il reste admis que *en l'état actuel* nous sommes pleinement justifiés à rejeter le déterminisme sur la base de notre expérience du libre arbitre ; l'introspectionniste ne cherche pas à contester qu'une révision éventuelle de cette justification est envisageable, en fonction de découvertes futures. On serait donc ici dans la situation d'une asymétrie non problématique épistémologiquement.

Concluons cette première série de difficultés qui concerne la justification apportée par l'expérience : le fait que l'introspectionnisme fasse reposer la croyance au libre arbitre sur une expérience (faillible) ne constitue pas une asymétrie avec le cas de la perception, au contraire. Et si l'on admet que l'expérience (faillible) du libre arbitre apporte une justification (révisable) pour son contenu, alors la seule asymétrie avec la perception est que le contenu admis de manière révisable (l'indéterminisme) est peut-être plus « lourd » (plus général) que le contenu typique des perceptions sensibles. Mais du moment que ce contenu lourd est *justifié et non réfuté*, le fait qu'il soit révisable n'a aucune conséquence épistémologique problématique. La

⁴⁴⁶ A vrai dire, je ne suis pas convaincu du tout qu'il y ait là réellement une asymétrie : après tout, le contenu de notre expérience visuelle n'exclut pas seulement les scénarios détaillés que sont les scénarios anti-sceptiques (comme l'expérience du libre arbitre exclut le scénario du Malin Génie). Elle exclut également le scénario *général* de l'Idéalisme, comme l'expérience du libre arbitre exclut le scénario général du Déterminisme. En effet, si l'on admet, contre Berkeley, que notre expérience perceptive a pour contenu l'existence d'objets matériels, alors ce contenu s'oppose à l'Idéalisme. Or comme le Déterminisme, l'Idéalisme est une hypothèse (i) qui a été défendu par des arguments positifs par le passé, (ii) qui concerne l'ensemble de la réalité et sa constitution objective, (iii) dont on pourrait envisager qu'il revienne en force sous l'effet de nouveaux arguments. On peut estimer qu'une future « preuve idéaliste » est moins probable qu'une future « preuve déterministe », mais c'est là essentiellement une question de plus et de moins. L'expérience perceptive semble donc rejeter une thèse tout aussi « lourde » et tout aussi « ouverte » que le Déterminisme qui est rejeté par l'expérience du libre arbitre. Si l'on voulait insister sur cette analogie entre justification de l'anti-Idéalisme (par la perception) et justification de l'indéterminisme (par l'expérience du libre arbitre), il faudrait sans doute souligner également que chacune de ces deux implications semble faire partie de notre « métaphysique spontanée », c'est-à-dire la manière dont nous nous représentons naturellement le monde dès que nous envisageons la question. (Pour ce qui est de la représentation indéterministe spontanée, voir la « question d'actualité », ch.1 p.66.) La conclusion serait donc la suivante : nos expériences d'une manière générale (perceptive ou non) ont le pouvoir de justifier non seulement le rejet des hypothèses sceptiques particulières, mais aussi l'admission de certaines croyances à propos de la structure générale et fondamentale de la réalité (l'existence de la matière, la non-détermination de tous les événements), des croyances qui ne sont pas justifiées autrement et qui composent notre métaphysique fondamentale de sens commun. Je ne développe pas ici cette analogie parce qu'elle ne me semble pas indispensable à l'argument, et qu'elle n'est donc pas apparue dans le cours de l'argumentation de la thèse.

seule manière de faire apparaître une asymétrie problématique serait donc de défendre que l'implication indéterministe ne peut même pas être justifiée sur la base d'une expérience. Pour cela, il faudrait alors réfléchir au niveau « méta-épistémique ».

Y a-t-il, au niveau méta-épistémique, une asymétrie problématique entre expérience du libre arbitre et expérience perceptive ? Autrement dit : sommes-nous plus ou mieux justifiés à considérer nos expériences perceptives comme sources de connaissance que notre expérience du libre arbitre ? Ce type d'objection pourrait prendre deux formes différentes. On pourrait prétendre que, dans le cas du libre arbitre, nous avons une preuve que cette expérience ne peut pas constituer une connaissance – une preuve que nous n'aurions pas dans le cas de la perception. Ou bien, on pourrait prétendre que nous n'avons pas de preuve contre la valeur épistémique de l'expérience du libre arbitre, mais on n'aurait pas non plus de preuve *positive* de sa valeur épistémique, alors que nous aurions une telle preuve dans le cas de la perception. Dans le premier cas, on reprocherait à l'expérience du libre arbitre de violer le principe de méta-cohérence négative ($J \sim Kp \rightarrow \sim Jp$) ; dans le second cas, on lui reprocherait de violer le principe de méta-cohérence positive ($\sim JKp \rightarrow \sim Jp$).

Commençons par le problème de méta-cohérence négative. Avons-nous (pour le libre arbitre, et pas pour la perception) une preuve de non fiabilité ? On pourrait prétendre tout d'abord que nous avons une preuve de non fiabilité de l'expérience du libre arbitre indépendamment du contenu lourd indéterministe. C'est ainsi que j'ai interprété les arguments de Libet et de Wegner au chapitre 6. Mais il me semble que la discussion de ces résultats scientifiques permet de se convaincre qu'il n'y a là, en fait, aucune réfutation de l'expérience du libre arbitre. Si l'on fait abstraction du contenu indéterministe, nous n'avons aucune raison de nous défier de notre expérience du libre arbitre plus que de nos expériences perceptives.

Le contenu indéterministe est à l'origine d'un problème méta-épistémique plus sérieux, que j'ai appelé le Défi d'Intégration du libertarianisme. On pourrait prétendre – et beaucoup d'auteurs ont exprimé l'intuition – que *à cause* de son implication déterministe, une expérience libertarienne ne *pourrait pas* constituer une source de connaissance. Le contenu indéterministe serait une implication « lourde » dans le sens, ici, où elle serait « inconnaissable sur la base de cette expérience ». La vraie asymétrie entre expérience perceptive et expérience du libre arbitre serait alors la suivante :

(L2) L'expérience du libre arbitre a une implication (l'indéterminisme) que nous ne pouvons pas connaître sur la base de cette expérience (et dont nous ne connaissons pas indépendamment la vérité).

(P2) Les expériences perceptives n'ont pas d'implications que nous ne puissions pas connaître sur la base de ces expériences.

Cette objection pose cependant un problème initial : à cause des implications anti-sceptiques des expériences perceptives, il n'est pas immédiatement clair que (P2) soit vraie. En effet, il n'est pas sûr qu'il soit possible de connaître, sur la base de mon expérience visuelle d'avoir deux mains, que « je ne suis pas un cerveau dans une cuve ». En tout cas, ce qui est clair c'est qu'il est impossible de satisfaire une certaine forme de « détection » de ce contenu anti-sceptique par l'expérience perceptive, à savoir la détection par sensibilité ou traçage : les expériences perceptives *ont* des implications dont elles ne peuvent pas « tracer » la vérité. Pour faire apparaître une asymétrie, il faudrait alors se reporter sur une autre notion de dépendance épistémique, une notion que la perception satisfait y compris pour ces conséquences anti-sceptiques, à savoir la notion de sûreté. Les expériences perceptives n'ont pas, semble-t-il, d'implications dont la vérité ne soit corrélée de manière sûre à l'occurrence de ces expériences. Mais y a-t-il alors asymétrie avec le cas du libre arbitre ? Non, et c'est un

des aspects fondamentaux de notre solution d'Intégration du libertarianisme : l'expérience du libre arbitre peut fort bien être corrélée de manière sûre à la vérité de son implication indéterministe, tout en traçant de manière sensible son contenu propre. La discussion du problème d'Intégration mène à mon sens à la conclusion suivante : il y a là une asymétrie apparente qui a été envisagée par de nombreux auteurs, une asymétrie qui *serait* problématique si elle était réelle, mais cette asymétrie disparaît à l'analyse. Mon intuition est que cette asymétrie apparente est en fait la raison *principale* pour laquelle de nombreux auteurs traitent différemment la perception et l'expérience du libre arbitre, et qu'en supprimant cette asymétrie, on devrait donc rétablir, pour beaucoup de philosophes, la situation initiale (épicurienne) de traitement analogue. Mais voyons tout de même les autres asymétries envisageables entre libre arbitre et perception.

Le deuxième type d'asymétrie méta-épistémique est fondé, comme on vient de le voir, sur le principe de méta-cohérence *positive* : l'idée serait que, même si nous n'avons pas de preuve *contre* la valeur épistémique de l'expérience du libre arbitre, nous n'avons pas non plus de preuve *positive* en sa faveur, alors que nous avons une telle preuve dans le cas de la perception. Et la toute première intuition qu'on pourrait avoir en ce sens, c'est que nous avons une théorie détaillée sur la manière dont nos expériences visuelles (par exemple) dépendent causalement du fait observé (au moyen de la réflexion des rayons lumineux, du fonctionnement de la rétine, puis de la circulation d'information dans les neurones), alors que nous n'avons aucun élément scientifique nous indiquant la manière dont l'expérience du libre arbitre pourrait dépendre du fait du libre arbitre. Cette asymétrie est réelle mais elle n'est pas épistémologiquement pertinente : pour qu'elle soit épistémologiquement pertinente, il faudrait soutenir que notre justification à tenir la perception pour fiable *repose* sur ce genre de théories scientifiques, et que notre expérience du libre arbitre, n'ayant pas ce genre de théories pour support, n'est donc pas bien « étayée ». Mais on ne peut pas dire que notre justification à tenir la perception pour fiable (notre méta-justification) repose sur ce genre de théories, et ce pour deux raisons. Tout d'abord, parce que ce genre de théories scientifiques repose clairement sur des données sensibles et ne peut donc être fondée que si l'on a déjà accepté, sur un autre fondement, la fiabilité de la perception sensible. Deuxièmement, parce que les philosophes de l'antiquité étaient certainement justifiés à tenir la perception pour fiable, or ils ne disposaient pas de données scientifiques de ce genre (ils n'avaient pas d'indication de la manière dont fonctionne causalement la vision). Le fait que la perception ait une méta-justification scientifique, et que l'expérience du libre arbitre n'en ait pas, est donc une asymétrie réelle mais sans pertinence épistémologique. Pour comparer de manière pertinente la méta-justification perceptive et la méta-justification du libre arbitre, il sera plus simple de prendre désormais pour repère la situation épistémologique des philosophes antiques.

Deuxièmement, on pourrait tenter de soutenir que les philosophes antiques avaient une méta-justification pour la perception dont nous ne disposons pas pour l'expérience du libre arbitre. Mais quelle peut être, plausiblement, la méta-justification perceptive dont disposaient les anciens ? Il est important de noter tout d'abord que cette méta-justification était forcément de l'ordre d'une présomption de connaissance, sans aucune information sur le « comment » de la connaissance en question : on vient de voir que les philosophes antiques n'avaient pas d'informations *scientifiques* sur le « comment » de la fiabilité visuelle, mais ils n'avaient pas non plus d'autres sources d'information sur cette question. La vision était présumée fiable non pas sans raisons ni justification, mais sans des raisons qui donnent la moindre indication sur la modalité de la dépendance fiable entre fait et expérience. D'où pouvait donc venir cette présomption de connaissance ? Dans le chapitre 3, j'ai envisagé deux caractéristiques des expériences perceptives que l'on pourrait plausiblement considérer comme fondements de méta-justification : on peut être justifié à présumer la fiabilité de nos expériences perceptives parce que (i) elles font partie de nos sources *naturelles* de croyance, ou bien parce que (ii)

elles sont accompagnées d'une « expérience mooréenne », c'est-à-dire de l'intuition de deuxième ordre qu'elles constituent des sources de connaissance. Ces deux caractéristiques positives des expériences perceptives sont les meilleurs candidats pour rendre compte de notre méta-justification *prima facie* en leur faveur, pour des raisons que je rappellerai un peu plus loin. Or aucune de ces deux caractéristiques ne permet d'établir une distinction avec l'expérience du libre arbitre. L'expérience du libre arbitre a toutes les chances d'être une expérience qui fait partie intégrante de notre constitution naturelle comme agents confrontés à des choix (difficiles) ; et plusieurs auteurs, compatibilistes ou non, ont donné raison à l'intuition de Johnson pour qui il est directement intuitif d'accepter la proposition de second ordre selon laquelle « nous savons que nous sommes libres » (sur la base de notre expérience). Pour ces deux raisons, il y a donc toutes les raisons de supposer que notre expérience du libre arbitre dispose, comme la perception, d'une méta-justification positive *prima facie*.

Cependant, on pourrait être tenté de dire que la perception a d'autres caractéristiques positives qui, même à l'époque antique, pouvaient constituer le fondement de notre présomption de fiabilité. Ces deux autres caractéristiques positives sont (i) la force intuitive de ces expériences perceptives, et (ii) le fait que les différents sens peuvent s'étayer les uns les autres de manière « intermodale ». A propos de la force intuitive, ce genre de théorie pourrait être défendu à partir du cas des croyances mathématiques. Pourquoi présumons-nous que nos intuitions mathématiques sont fiables, alors qu'il y a un problème d'Intégration (il ne semble pas que nous ayons un accès causal à des objets abstraits) ? Pourquoi la « solution sceptique » est-elle quasiment exclue pour ce problème d'Intégration particulier.⁴⁴⁷ Une réponse possible est que les intuitions mathématiques (certaines d'entre elles en tout cas) sont *beaucoup trop évidentes*, beaucoup trop fortes intuitivement, pour qu'on puisse adopter une attitude sceptique à leur égard. Nous savons que $1 + 1 = 2$, un point c'est tout. On pourrait défendre également que, lorsque je vois une table en face de moi dans les meilleures conditions, en plein lumière, sans avoir bu, etc. mon impression visuelle est si évidente que je ne peux pas faire autrement que de présumer qu'elle est une connaissance. Que penser de cette méta-justification par la force intuitive ? Je ne suis pas certain que cette solution soit réellement distincte de la solution par « expérience mooréenne ». Mais quoi qu'il en soit, si c'était la *bonne* théorie de la méta-justification perceptive, il est clair qu'elle ne ferait aucune différence avec l'expérience du libre arbitre, puisque notre expérience du libre arbitre est *aussi* une expérience extrêmement vive.

Il reste donc une seule caractéristique positive de la perception qui pourrait faire une différence avec le libre arbitre, c'est la caractéristique de « renfort intermodal » : lorsque je vois un coin de table en face de moi, je peux confirmer la fiabilité de cette expérience visuelle par le toucher. Inversement, si je sens un objet pointu sous la plante de mon pied, je peux vérifier la fiabilité de cette expérience tactile par la vue. Les différents sens s'apportent les uns aux autres un renfort épistémique, des raisons de supposer qu'ils sont fiables, et ce avant toute théorie scientifique sur la manière exacte dont la vision ou le toucher sont fiables. Or nous ne disposons pas d'un tel « renfort intermodal » pour l'expérience du libre arbitre : lorsque je me sens libre de faire tel ou tel choix, je ne peux pas « regarder » ou « toucher » pour vérifier que les deux choix sont possibles. L'expérience du libre arbitre est une expérience qui est peut-être naturelle, accompagnée d'une expérience mooréenne, et très vive, mais elle semble une expérience *isolée*. Nous avons là une troisième asymétrie entre perception et libre arbitre qui me semble mériter discussion :

⁴⁴⁷ Personne, à ma connaissance, n'a répondu au trilemme d'Intégration de Benacerraf en refusant l'intuition mooréenne, et en disant par exemple : « puisque les objets mathématiques sont des objets abstraits, et puisque nous ne pouvons pas connaître des objets abstraits, c'est donc que nous ne savons pas si $1 + 1$ est égale 2 – peut-être est-ce en fait égale à 3, ou à 4, nous ne pouvons pas savoir ».

(P3) Les expériences perceptives d'un sens particulier peuvent être étayées par les expériences perceptives d'autres sens.

(L3) L'expérience du libre arbitre ne reçoit aucun renfort d'autres sens et ne peut donc pas être étayée (ou méta-justifiée) par renfort intermodal.

Je ne conteste pas qu'il y ait là une différence entre la perception et l'expérience du libre arbitre. Mais je conteste que cette différence soit épistémologiquement pertinente. Pourquoi ? Pour que cette différence soit épistémologiquement pertinente, il faudrait défendre que le renfort intermodal dont bénéficie la perception est le fondement même de notre méta-justification (préscientifique) : si notre raison de nous fier à nos sens *vient du fait* qu'ils s'étayent les uns les autres, alors on ne peut pas soutenir que l'expérience du libre est méta-justifiée de la même façon que la perception – et on conclura probablement qu'elle n'est pas méta-justifiée du tout au niveau préscientifique, c'est-à-dire qu'Epicure ou Aristote étaient infondés à présumer que l'expérience du libre arbitre devait constituer (d'une manière ou d'une autre) une connaissance. (Notez qu'un compatibiliste contemporain serait alors tout aussi infondé à présumer cela : il ne pourrait donc affirmer l'existence du libre arbitre que comme un *résultat*, sur la base de son analyse compatibiliste *plus* d'une indication empirique que nous possédons effectivement tel type de libre arbitre.) Personnellement, cette idée selon laquelle Epicure ou Aristote (ou aussi bien tout auteur introspectionniste compatibiliste) avaient *tort* de présumer la fiabilité de notre conviction naturelle du libre arbitre me semble en elle-même peu plausible. Mais puisque c'est justement le point qui est en jeu, je ne m'appuierai pas ici sur cette intuition.

On peut montrer, je crois, que la caractéristique de « renfort intermodal » n'est pas un critère épistémologiquement pertinent qui permettrait de distinguer les expériences méta-justifiées des expériences non méta-justifiées (à la différence du critère de « naturalité » ou du critère d'« expérience mooréenne »). Un critère pertinent pour fonder la méta-justification doit être un critère qui permet de résoudre le *problème* de la méta-justification, c'est-à-dire le problème sceptique de la méta-régression tel que nous l'avons présenté au chapitre 3 (sec. II.2). Le problème fondamental de la méta-justification, c'est qu'il est impossible (à moins d'être « infinitiste ») de donner une théorie de la méta-justification qui ne soit pas ultimement circulaire. D'une manière ou d'une autre, la méta-justification de nos croyances de base doit provenir des sources mêmes qui sont ainsi méta-justifiées. Or la circularité épistémique est clairement problématique. Certains auteurs soutiennent même qu'elle est toujours et nécessairement vicieuse. Si c'était le cas, alors il serait tout bonnement impossible de méta-justifier nos sources épistémiques, et donc (à cause du principe de méta-cohérence) nous serions conduits au scepticisme. La seule manière d'éviter le scepticisme est de faire une distinction entre circularités vicieuses et circularités vertueuses. Par conséquent, un bon critère de méta-justification devra nécessairement être un critère qui permet d'expliquer pourquoi la circularité qu'il recèle est vertueuse et non pas vicieuse. C'est à ce titre que les critères de « naturalité » et d'« expérience mooréenne » présentent des avantages que ne présente pas le critère de « renfort intermodal ». En effet, on peut comprendre qu'il est légitime *prima facie* de se fier à nos yeux plutôt qu'à une jauge parce que nos yeux sont des indicateurs qui nous sont *naturels* – même si c'est sur la base de la sensation elle-même que nous savons qu'ils sont des organes naturels (circularité vertueuse).⁴⁴⁸ De même, on peut comprendre que le caractère non inférentiel d'une expérience mooréenne (« cette source S semble fiable ») permet d'éviter le problème particulièrement évident d'une *inférence* épistémiquement circulaire.⁴⁴⁹ En revanche, si la caractéristique de « renfort intermodal » est

⁴⁴⁸ Pour les détails de cette solution naturaliste, voir p. 184 sqq.

⁴⁴⁹ Pour les détails, voir p. 210 sqq. Pour rappel, le problème particulier d'une *inférence* épistémiquement circulaire, c'est que la première *étape* de l'inférence est irrémédiablement « fragile » : si je dis « (i) (la vision

une caractéristique qui semble, à première vue, un atout épistémique, son intérêt s'évanouit lorsqu'on fait observer le caractère circulaire de ce renfort intermodal : la circularité propre au renfort intermodal n'offre aucune ressource argumentative pour expliquer qu'il s'agirait d'une circularité épistémiquement vertueuse plutôt que vicieuse. Une situation de renfort intermodal est une situation de circularité dans laquelle, au lieu d'avoir une source S qui s'étaye elle-même (circularité à un maillon), nous avons une première source S1 qui étaye une autre source S2, et S2 qui en retour étaye S1 (circularité à deux maillons, ou plus). Autrement dit, le renfort intermodal ne fait qu'« agrandir le cercle » ; or il est bien connu, dans les problèmes de circularité vicieuse, que le fait d'agrandir le cercle a tendance à faire *oublier* ou perdre de vue le caractère vicieux de l'opération, mais pas à supprimer ce caractère vicieux. Le cercle reste illégitime ou vicieux ; c'est seulement notre sensibilité à son caractère vicieux qui change. Par conséquent, on voit bien pourquoi le renfort intermodal pourrait donner *l'impression* d'être une source de méta-justification non vicieuse (parce qu'elle a tendance à faire oublier son caractère vicieusement circulaire), mais on ne voit pas en revanche pourquoi ce type de circularité serait non vicieux.

La caractéristique de renfort intermodal de la perception ne peut donc pas être la source fondamentale de notre méta-justification préscientifique en faveur de la perception. C'est pour cette raison que ce critère n'est pas apparu au cours du chapitre 3, où nous ne cherchions pas de manière systématique toutes les différences possibles entre expérience du libre arbitre et expérience perceptive, mais seulement les différences qui auraient une chance d'être pertinentes du point de vue épistémologique. Trois remarques encore concernant le renfort intermodal. Premièrement, si on considérait que le renfort intermodal est *la* source de notre méta-justification perceptive, cela aurait aussi pour conséquence d'effectuer une sélection au sein des différentes informations sensibles : si la vision et le toucher peuvent établir la fiabilité l'un de l'autre concernant les « sensibles communs » (les formes, les mouvements, etc.), rien ne permettra de montrer la fiabilité de la vision ou du toucher *quant à leurs sensibles propres* (la couleur, le chaud et le froid, etc.). Et rien ne permettra de montrer la fiabilité de l'odorat, qui ne semble donner d'informations *que* sur des sensibles propres. Finalement, il serait très abusif de dire que le critère de renfort intermodal permettrait une méta-justification « de la perception » ; il permettrait au mieux une méta-justification d'une certaine partie restreinte des informations de certains sens. La conclusion d'une telle théorie serait alors qu'il faut être sceptique concernant tous les sensibles propres, tant que nous n'avons pas une théorie scientifique de la fiabilité des sens pour leurs sensibles propres. Cela me semble une conclusion suffisamment « sceptique » pour rechercher une autre source de méta-justification de la perception. Deuxièmement, on peut soutenir que le renfort intermodal n'est pas *la* source de méta-justification de la perception ne sans pour autant nier que ce renfort intermodal soit une caractéristique épistémique positive : on peut considérer, par exemple, que le renfort modal de deux sources qui sont *déjà* méta-justifiées tend à augmenter le degré de *cohérence interne* du système épistémique qui contient ces deux sources. Et on peut défendre qu'entre deux systèmes épistémiques tous les deux cohérents et méta-justifiés, le système le plus cohérent a une certaine forme de supériorité épistémique (il est épistémiquement « plus confortable »). J'accepte volontiers ce diagnostic. Mais il n'a aucune conséquence pour notre problème. Tout ce que cela permet de conclure, c'est qu'il y a une certaine différence de *degré* entre le statut épistémique de certains aspects de la perception, et l'expérience du libre arbitre ; mais cela ne remet pas en cause que l'expérience du libre arbitre soit, absolument parlant, méta-justifiée, et ce de la même façon que la perception.

indique que) les rayons lumineux fonctionnent de telle manière, donc (ii) la vision est fiable », alors bien que la vision soit méta-justifiée à l'étape (ii), elle ne l'est pas *encore* à l'étape (i), au moment où pourtant nous devons bien nous y fier pour accepter ses données. Une méta-justification *basique* par expérience mooréenne ne laisse pas un tel délai, une telle étape fragile.

Troisièmement, le fait que l'expérience du libre arbitre ne soit pas « renforcée » par d'autres sens n'est pas nécessairement une situation définitive : si nous découvrons un jour empiriquement des indéterminismes bien placés dans le cerveau humain, nous pourrions dire alors que l'expérience du libre arbitre trouve un renfort dans la perception sensible. C'est à peu près ce qui s'est passé pour l'odorat : dans l'antiquité, la fiabilité des odeurs n'avait aucun renfort intermodal, mais on peut supposer qu'elle était néanmoins méta-justifiée soit par sa naturalité, soit par expérience mooréenne ; lorsque les sciences ont montré le fonctionnement des cellules odorantes et des cils olfactifs, l'odorat a trouvé là un renfort épistémique qui a sans doute rendu « plus cohérent » notre système cognitif, mais qui n'était pas nécessaire pour la méta-justification de l'odorat. De la même façon, l'expérience du libre arbitre, qui est aujourd'hui dans la situation épistémique de l'odorat il y a quelques siècles, pourrait être un jour renforcée par des découvertes scientifiques, mais elle n'a pas besoin de ces découvertes scientifiques pour être dès aujourd'hui absolument méta-justifiée.

On voit ici que l'absence de renfort intermodal pour l'expérience du libre arbitre est une « faiblesse » qui est du même ordre que l'absence de méta-justification scientifique : c'est une différence réelle avec (certains aspects de) la perception, et une différence qui permet sans doute de dire que l'expérience du libre arbitre est épistémiquement « moins forte » que (certains aspects de) l'expérience sensible. Mais si on décrit bien cette différence, elle ne permet pas de fonder une différence *de nature* avec la perception, celle-ci étant méta-justifiée, et l'expérience du libre arbitre ne l'étant pas. Elle ne peut fonder qu'une différence de degré de cohérence, les deux types d'expériences étant, absolument parlant, tous les deux méta-justifiés.

Pour conclure cet inventaire des différences entre expérience du libre arbitre et expérience sensible, je dirais qu'il existe à mon sens trois différences significatives auxquelles on pourrait penser si l'on veut dire que la perception est justifiée tandis que l'expérience du libre arbitre ne l'est pas :

- 1) le renfort intermodal de (certains aspects de) la perception, dont ne dispose pas (aujourd'hui) l'expérience du libre arbitre.
- 2) le fait que l'expérience du libre arbitre a une implication (l'indéterminisme) qui n'est pas la négation d'une pure hypothèse sceptique et qui semble plus « substantielle » que les implications (non anti-sceptiques) des expériences perceptives.
- 3) le fait que l'expérience du libre arbitre a une implication (l'indéterminisme) qui n'est pas la négation d'une pure hypothèse sceptique et qui semble inconnaissable (ou inintégréable) tandis que les expériences perceptives n'ont que des implications intégrables (à part leurs implications anti-sceptiques).

En ce qui concerne les points 1 et 2, je n'ai pas cherché à contester qu'il y ait là une différence réelle entre expérience du libre arbitre et expériences perceptives. Seulement, j'ai défendu que ces différences ne sont pas épistémologiquement pertinentes dans le sens où elles ne permettent pas de mettre en doute le caractère fondamentalement justifié de la croyance introspective au libre arbitre.

Il est vrai que l'expérience du libre arbitre n'a pas de renfort modal, alors que la vision et le toucher en ont à propos des sensibles communs. Si la méta-justification reposait sur le renfort intermodal, alors l'introspectionisme du libre arbitre serait réfuté indépendamment de toute implication indéterministe. Mais en réalité, la méta-justification de la perception ne peut pas reposer sur le renfort intermodal ; celui-ci ne peut apporter qu'un renforcement de degré de cohérence interne, un renforcement dont l'expérience du libre arbitre bénéficiera peut-être un jour, au même titre que l'odorat en a tardivement bénéficié.

Il est peut-être vrai que l'expérience du libre arbitre, à cause de son implication indéterministe, nous amène à adopter une croyance plus générale et plus « substantielle » que les expériences perceptives (qui n'auraient, outre leurs implications anti-sceptiques, que des implications « légères »).⁴⁵⁰ Cela entraîne sans doute que l'information apportée par cette expérience pourra plus facilement être réfutée par des informations contraires. Mais tant que cette différence (de degré, encore une fois) n'empêche pas à l'expérience du libre arbitre de fournir une justification *prima facie* pour son contenu relativement « lourd », elle n'aura pas d'incidence épistémologique. Or le fait qu'une expérience apporte ou non une justification *prima facie* pour son contenu ne peut pas dépendre du degré plus ou moins général, plus ou moins « lourd » de ses implications : une expérience peut justifier des implications de diverses « lourdeurs » du moment qu'elle peut constituer une source de connaissance pour une information de cette lourdeur. Bien sûr, si l'on établissait que l'implication de l'indéterminisme est trop lourde *pour être connue* par l'expérience du libre arbitre, alors cette expérience ne pourrait plus justifier une implication si lourde. Mais c'est là le troisième problème, que nous allons envisager dans un instant. Le second problème, dont nous parlons pour l'instant, consiste à reprocher à l'implication indéterministe le simple fait d'avoir un contenu relativement plus lourd (plus général, plus substantiel) que les contenus perceptifs, indépendamment de la question de savoir si l'on peut *connaître* ce contenu. Or il me semble clair que cette objection n'est pas pertinente : si une implication n'est pas trop lourde *pour être connue*, elle ne sera pas non plus trop lourde *pour être justifiée*. Encore une fois, une expérience peut justifier des croyances de n'importe quelle « lourdeur » à condition de pouvoir être une source de connaissance pour ces croyances. Le *seul* vrai problème est donc de savoir si l'implication indéterministe pourrait être *connue* par l'expérience du libre arbitre.

C'est là la troisième asymétrie envisageable entre expérience du libre arbitre et expérience perceptive : on *pourrait* défendre que l'implication indéterministe ne peut pas être connue sur la base de notre expérience du libre arbitre, à cause d'un problème de hasard épistémique. C'est sur cette asymétrie que j'ai concentré dans cette étude tous mes efforts, parce que j'estime (pour les raisons qu'on vient de voir) qu'elle est en fait la seule capable de poser un véritable problème épistémologique pour la justification de la croyance introspective au libre arbitre. C'est-à-dire que si il y avait une telle asymétrie, alors il faudrait reconnaître que libre arbitre et perception ne sont pas à égalité épistémologiquement, et que la croyance introspective au libre arbitre n'est pas, ultimement, justifiée. A mon sens, c'est aussi cette asymétrie que de nombreux adversaires de l'introspectionnisme ont plus ou moins explicitement à l'esprit. Pour répondre à cette difficulté, je me suis efforcé de montrer que cette asymétrie ne résistait pas à une analyse approfondie : en réalité, l'implication indéterministe peut être *connue* sur la base de notre expérience du libre arbitre, si l'on applique les critères épistémologiques définis pour la perception sensible (à savoir le critère de « sensibilité pour le contenu propre et sûreté pour les conséquences »).

La comparaison entre expérience du libre arbitre et expériences perceptives permet donc de tirer la conclusion suivante : en termes quantitatifs, la croyance introspective au libre arbitre n'est peut-être pas « aussi forte » que les croyances perceptives (parce qu'elle n'a pas, pour l'instant, de renfort intermodal, ni d'étayage scientifique, et que son contenu substantiel pourrait peut-être être plus facilement réfuté par des découvertes éventuelles). Mais ce ne sont là que des questions de *degrés*. Absolument parlant, la croyance introspective au libre arbitre dispose exactement des mêmes structures fondamentales de justification et de méta-justification que les croyances perceptives : elle est justifiée *prima facie* par expérience, méta-justifiée *prima facie* par sa naturalité ou par expérience mooréenne, non réfutée directement

⁴⁵⁰ Voir cependant la note 446 p.480.

par les données actuelles de la philosophie et des sciences, et non réfutée fondationnellement parce qu'elle peut constituer une connaissance.

Par conséquent, il n'existe que deux options : soit on considère que ces structures épistémologiques sont insuffisantes pour fonder une croyance rationnelle, et dans ce cas on admet un scepticisme très radical pour la perception (non seulement nos croyances perceptives ne sont pas des connaissances, mais elles sont tout bonnement *irrationnelles*). Soit on refuse le scepticisme radical de la perception, et dans ce cas, on doit reconnaître que la croyance introspective au libre arbitre, avec son implication indéterministe, est une croyance tout à fait rationnelle. Plus précisément, j'ai même défendu que la position introspectionniste libertarienne était la position *la plus rationnelle* à adopter dans le débat sur le libre arbitre, si du moins on n'accepte pas le scepticisme radical. Dans cette étude, je n'ai pas entrepris la lourde tâche de réfuter la position sceptique radicale. Je me suis adressé principalement aux personnes qui, comme moi, rejettent le scepticisme radical, mais qui pourtant croient trouver une faiblesse épistémologique spécifique dans l'introspectionnisme libertarien.

En introduction, j'ai présenté un petit dialogue qu'il m'est arrivé d'avoir avec plusieurs personnes, philosophes ou non :

- Etes-vous libre ?
- Oui bien sûr !
- Comment le savez-vous ?

Et j'ai mentionné la réaction d'incompréhension et d'embarras d'une certaine partie de mes interlocuteurs. Mais une autre partie de mes interlocuteurs (presque tous non philosophes) ne semblait pas embarrassée et semblait trouver la réponse évidente. Cette étude n'a consisté finalement qu'à défendre, en beaucoup de mots, la réponse évidente de ces interlocuteurs :

- Etes-vous libre ?
- Oui bien sûr !
- Comment le savez-vous ?
- Je le sens bien, tout simplement !

ANNEXES

Annexe 1 : Pourquoi la notion internaliste de « Justification » ne préjuge pas de l'Internalisme ou de l'Externalisme

En introduction, j'ai défini la justification de la manière suivante :

S croit que p de manière justifiée =_{df}

(i) S croit que p

(ii) S ne viole aucune des normes strictement internalistes proprement épistémiques en croyant que p .

La notion de justification ainsi définie est une notion qu'on peut appeler « internaliste », mais qui, en tant que stipulation terminologique, n'entend pas prendre parti dans le débat de fond entre Internalisme et Externalisme. Pourtant, il semble difficile d'exprimer une thèse externaliste à partir de cette définition, et donc il semble que cette définition implique malgré tout, *volens nolens*, l'Internalisme.

Pour répondre à ce problème, le plus simple est de repartir de ce que la définition présuppose effectivement : cette définition présuppose l'existence de normes strictement internalistes et proprement épistémiques (c'est-à-dire non déontologiques).⁴⁵¹ Les exemples apportés pour défendre ce présupposé sont notamment la non-contradiction ou la non-circularité épistémique. La non-contradiction est clairement une norme épistémique importante, qu'on ne peut pas réduire à une obligation morale et dont la satisfaction peut être déterminée de manière purement internaliste (au moins dans certains cas⁴⁵²). De même, si un juge naïf adopte le raisonnement suivant « le témoin T vient de m'assurer qu'il est fiable, donc sur sa foi, je crois désormais que T est fiable », on peut déterminer de manière purement internaliste qu'il viole une norme épistémique (quand bien même il serait parfaitement excusable moralement). Il semble donc difficile de refuser l'idée qu'il y a de telles normes internalistes non déontologiques.

Mais la première chose à noter, c'est que l'externaliste ne *nie pas* l'existence de telles normes. Il ne nie pas qu'une croyance soit épistémologiquement problématique lorsqu'elle est contradictoire, ou lorsqu'elle est aussi trivialement circulaire que celle du juge naïf. Il ne nie pas non plus que ce problème (cette violation) soit déterminable (au moins dans certains cas) de manière purement internaliste. Et par conséquent, l'externaliste peut concevoir et manipuler le concept d'impeccabilité doxastique interne que j'appelle « justification », c'est-à-dire le statut d'une croyance qui n'a violé aucune de ces normes épistémiques internalistes. Seulement, il appellera généralement ce statut épistémique d'un autre terme, car il réserve celui de « justification » pour la satisfaction de normes épistémiques externalistes. Pour parler des normes internalistes et de leur satisfaction, un externaliste utilisera le vocabulaire des « defeaters » de justification – c'est-à-dire des conditions d'annulation d'une justification (Plantinga 2000, 359–366; Bergmann 2006, 154–177). Autrement dit, dans le langage externaliste, le terme de « justification » sanctionne principalement la satisfaction (ou non satisfaction) des normes externalistes, mais la satisfaction (ou non satisfaction) des normes

⁴⁵¹ Cette présupposition est générée par l'article défini « des » dans « aucune des normes etc. ».

⁴⁵² Evidemment, certaines contradictions seront non déterminables internalistiquement – par exemple, la croyance que « l'étoile du matin est une planète tandis que l'étoile du soir n'en est pas » qsd – mais d'autres le seront – par exemple la croyance que « l'étoile du matin est une planète et n'est pas une planète ». Quelqu'un qui adopterait la seconde croyance violerait clairement une norme épistémique *internaliste* (quand bien même il serait parfaitement excusable moralement).

internalistes est néanmoins prise en compte à travers la distinction « defeated / undefeated » (annulée / non annulée).

Ces considérations pourraient facilement donner l'impression qu'il n'y a finalement aucun désaccord substantiel entre Internalistes et Externalistes. Que ce que les premiers appellent « être justifié », les seconds l'appellent simplement « ne pas être annulée ». Si l'internalisme n'affirme rien de plus que l'existence (et l'importance) de normes épistémiques internes, alors l'externaliste peut être d'accord. Or, certaines défenses de l'internalisme semblent aller dans ce sens. Ça semble être le cas de Jim Pryor dans son argument des trois cerveaux en cuve, qui a pour but de séparer les critères déontologiques des critères proprement épistémiques (dans un contexte sceptique où il est évident que les seuls critères satisfaisables sont *internalistes*). Pryor imagine ainsi trois cerveaux en cuve : le premier est clairement épistémiquement blâmable ; le second applique consciencieusement de très mauvais standards de raisonnement, mais n'est pas moralement responsable des standards qu'il a (on les lui a inculqués de force) ; le troisième enfin applique consciencieusement de bons standards de raisonnement et n'échoue donc que sur la fiabilité externaliste (puisqu'il reste un pauvre cerveau en cuve). Pryor tire la conclusion suivante :

« Ceci indique qu'il y a un type de statut épistémique positif qui va au-delà de la simple inblâmabilité épistémique, mais que l'on peut posséder même si nos croyances sont systématiquement fausses, comme le sont les croyances des cerveaux. De telles considérations me font pencher vers une conception internaliste non déontologique de la justification. » (Pryor 2001, 117)

Mais comme on vient de le voir, ces considérations ne font pas une différence substantielle avec l'externalisme bien considéré. Elles peuvent au mieux rendre utile une *terminologie* internaliste (comme celle utilisée dans cette thèse) ; elle ne permettent pas de défendre une thèse particulière, une « thèse internaliste » en face d'une « thèse externaliste ».

D'autres caractérisations de l'internalisme, pourtant, permettent de faire apparaître le véritable désaccord de fond. Bergmann définit le cœur de l'intuition internaliste sous la forme d'un « *Réquisit de Conscience* » :

« *Le Réquisit de Conscience* : la croyance B de S est justifiée seulement si :
(i) il y a quelque chose, X, qui contribue à la justification de B [...] et
(ii) S a conscience (ou potentiellement conscient) de X. » (M. A. Bergmann 2006, 9)

Ce qui est important pour nous, ce n'est pas la condition de conscience en elle-même, le (ii), c'est plutôt le (i). Le (i) nous dit qu'il doit *exister quelque chose*, un contributeur de justification (*justification-contributor*) pour que la croyance soit justifiée. Il est important de remarquer que cette condition constitue en elle-même une norme épistémique, et en l'occurrence, évidemment, une norme internaliste et non déontologique. Par rapport aux *autres* normes internalistes non déontologiques que nous avons déjà vues (la non-contradiction ou la non-circularité), cette nouvelle norme est d'un type totalement différent : les autres étaient des normes *négatives*, celle-ci est une norme *positive*. En d'autres termes, les autres normes exigeaient seulement *l'absence* d'un élément interne et qui mettrait en péril la justification ; le réquisit de conscience exige en outre la *présence* d'un élément interne qui contribue positivement à la justification. On peut donc donner la caractérisation suivante de la *thèse* internaliste :

Internalisme :

Parmi les normes internalistes non déontologiques, certaines sont des normes *positives* au sens suivant : elles n'exigent pas seulement *l'absence* d'éléments internes problématiques (l'absence de contradiction, l'absence de circularité), mais elles exigent la *présence* d'un élément interne qui contribue positivement à la justification (un élément qu'on peut appeler un *justifieur* internaliste).

Et la thèse adverse s'exprimera alors de la manière suivante :

Externalisme :

Les *seules* normes internalistes non déontologiques qui existent sont des contraintes négatives (*non-contradiction*, *non-circularité*, etc.). Une croyance n'a pas besoin, pour être correcte, d'un *justifieur* internaliste positif, mais seulement d'être internalistiquement « non défaite ».⁴⁵³

On a bien là deux thèses incompatibles, et qui sont indépendantes de la terminologie « internaliste » définie plus haut. La différence entre Internalisme et Externalisme n'est pas que l'un accepterait un statut d'impeccabilité internaliste que l'autre ne peut comprendre. L'externaliste comprend fort bien l'impeccabilité internaliste, et reconnaît qu'une croyance *doit* être internalistiquement impeccable (ne violer aucune norme internaliste non-déontologique). Seulement, pour l'externaliste, on peut être internalistiquement impeccable de manière « vide » ou triviale, puisque les seules normes internalistes non déontologiques sont des contraintes négatives.

Imaginons par exemple un esprit éternel E, dont toute la vie mentale, de toute éternité, serait constituée d'une unique croyance occurrente que *p*. Il est clair que la croyance de E n'est en contradiction avec aucune autre croyance de E (il n'a *aucune* autre croyance), ni en conflit ou incohérence avec aucune expérience vécue de E (il n'a *aucun* autre vécu que cette croyance) ; elle n'est pas non plus fondée sur un élément de sa vie mentale qui reposerait circulairement sur *p* (il n'y a *aucun* autre élément dans sa vie mentale). D'une manière générale, il est clair que E ne peut violer aucune norme internaliste négative, car il faudrait pour cela qu'*existe* dans son esprit quelque chose qui entre en conflit épistémique avec la croyance que *p*, or il n'y a que la croyance que *p* en son esprit, et que cette croyance que *p* ne saurait entrer en conflit avec elle-même.

Donc si l'externaliste a raison, c'est-à-dire si toutes les normes épistémiques internalistes, sans exception, sont des normes négatives alors il s'ensuit que E ne peut violer aucune norme épistémique internaliste, et donc E est nécessairement internalistiquement impeccable. Inversement, si l'internaliste a raison, alors il s'ensuit que E ne sera (très probablement) *pas* internalistiquement impeccable, car pour être impeccable, il lui faudrait trouver dans sa vie mentale un élément *x* tel que *x* sert de fondement à sa croyance que *p*, mais comme il n'y a rien d'autre que cette croyance dans sa vie mentale, et qu'une croyance fondée directement sur elle-même serait certainement une circularité problématique, il semble que E ne *puisse* pas satisfaire les normes d'impeccabilité qu'un internaliste reconnaît.

Pour l'instant, j'ai soigneusement évité le terme de « justification » lui-même dans la description du scénario de l'esprit éternel E. Je l'ai fait pour rendre plus clair le désaccord substantiel entre Internalistes et Externalistes, grâce à des termes neutres. Mais il est clair que ce désaccord peut se retraduire dans n'importe quel vocabulaire, et notamment dans mon vocabulaire « internaliste » où « justification » signifie « impeccabilité internaliste ». Dans ce vocabulaire, l'internaliste dira ceci : « l'esprit E ne peut pas être justifié dans sa croyance que

⁴⁵³ En revanche, une croyance a besoin, pour être correcte, d'un justifieur *externaliste*, c'est-à-dire qu'il y a bien des réquisits épistémiques *positives* mais que les seuls réquisits épistémiques positifs sont externalistes.

p, car pour être justifié, il faut disposer d'un élément justifieur ». L'externaliste, quant à lui, dira ceci : « l'esprit *E* est nécessairement justifié, bien qu'il n'ait pas de justifieur ; il est justifié de manière vide... ce qui montre à quel point il est facile et peu significatif d'être « justifié » dans ce sens qui est un sens strictement négatif. »

On peut donc, pour résumer, formuler ainsi les deux thèses :

Internalisme :

Pour être justifié à croire que *p*, *S* doit avoir justifieur (interne) en faveur de *p*.

Externalisme :

S peut être justifié à croire que *p* sans avoir de justifieur (interne) en faveur de *p*.

Bien que le désaccord entre Internalistes et Externalistes soit substantiel, je crois qu'on peut aussi, assez facilement, expliquer pourquoi il aboutit aussi fréquemment à une divergence terminologique. Au-delà du désaccord substantiel, il y a paradoxalement un accord entre Internalistes et Externalistes sur l'idée suivante : pour qu'une notion de « justification » soit vraiment pertinente, il faut qu'elle repose sur un « justifieur ». L'Internaliste en fait une condition sur la justification internaliste : comme il y a une notion pertinente de justification internaliste, ça doit être que d'une manière ou d'une autre, il y a un réquisit de justifieur internaliste. L'externaliste raisonne dans l'autre sens : comme on ne peut pas imposer aux croyances un réquisit de justifieur internaliste, toute notion de justification internaliste sera strictement négative, ce sera une notion de « ne pas être défait », mais pas une notion de « être justifié par *x* », et donc il n'y a pas de notion de justification internaliste vraiment intéressante. Il vaut donc mieux garder le nom de « justification » pour la seule contrainte de « justifieur » qui soit raisonnable, à savoir la contrainte de justifieur externaliste.

Donc l'internaliste (classique) et l'externaliste (classique) semblent partager une même pratique terminologique qui consiste à n'appeler « justification » un ensemble de normes épistémiques que si il assure un justifieur positif. C'est cette pratique (et non pas une thèse quelconque) que ma terminologie remet en cause : telle que je la définis, la « justification » n'implique pas (elle n'exclut pas non plus) qu'il y ait besoin d'un « justifieur ». J'entends le terme de « justifié » *d'abord* comme un terme négatif signifiant « ne prêtant pas le flanc à la critique ». C'est ainsi qu'il s'applique, notamment, au domaine moral ou juridique : « ne prêtant pas le flanc au blâme moral, ou à une inculpation juridique ». Je ne crois pas que la notion de justification juridique implique nécessairement d'avoir un élément qui rend juste ; on peut être « non inculpable » sans avoir un élément positif qui « exculpe » ou excuse, simplement parce qu'il n'y a pas non plus d'élément qui inculpe.⁴⁵⁴ De façon analogue, je ne vois pas pourquoi il devrait faire partie du *concept* de « justification épistémique internaliste » que pour être justifiée une croyance doive avoir un justifieur positif.

L'origine de cette habitude terminologique réside, à mon sens, non pas dans une analyse conceptuelle ou dans une thèse quelconque, mais dans l'ambiguïté bien connue des substantifs verbaux : quand on parle de la « justification » (*justification*) d'une croyance, est-ce qu'on désigne un « justifieur » (*justifier*) ou « le fait qu'elle soit justifiée » (*justifiedness*) ? Pour peu qu'on ne remarque pas cette ambiguïté, il est très facile de glisser d'un réquisit de *justifiedness* à un réquisit de *justifier*, et d'une terminologie de la *justifiedness* interne à une thèse exigeant un *justifier* internaliste. Malheureusement, plusieurs expressions de la famille lexicale de la justification ont tendance à aggraver l'ambiguïté. Ainsi l'expression « avoir une

⁴⁵⁴ Certes, dans certains cas, il faudra sans doute un élément qui excuse. Et peut-être certains théoriciens du droit, mettant en cause la présomption d'innocence, défendront la *thèse* qu'il faut dans tous les cas un élément positif d'excuse. Mais cela restera une *thèse* et non une analyse du concept de justification juridique.

justification », qui peut très facilement être prise comme synonyme de « être justifié »... mais tout aussi facilement comme synonyme de « avoir un justifieur ». Même l'expression « être justifié » peut se prêter à une certaine ambiguïté entre une lecture passive et donc à complément d'agent implicite (« être justifié *par* », cf. « être étayé ») ou une lecture purement adjectivale (cf. « être impeccable »).

Annexe 2 : Métaphilosophie ou Epistémologie de la philosophie ?

Comme il a été dit en Introduction (sec. 3.1), cette étude aborde la question métaphysique du libre arbitre avec un angle particulier : je n'y défends pas, de manière directe, que telle ou telle thèse sur le libre arbitre est *vraie*, mais j'examine plutôt les critères et les méthodes de justification qui permettront de considérer telle ou telle thèse comme *justifiée*. Etudier (philosophiquement) les méthodes qui doivent être appliquées pour résoudre une question philosophique, c'est faire, en un sens, de la philosophie *à propos* de la philosophie, c'est-à-dire de la métaphilosophie. Ma démarche pourrait donc être comparée à celle de Richard Double dans *Metaphilosophy of Free Will* (Double 1996). Dans cet ouvrage, Double soutient que la bonne approche du problème du libre arbitre est l'approche métaphilosophique, dans le sens où c'est seulement la discussion des critères métaphilosophiques qui permettra de trancher le débat métaphysique dans un sens ou dans un autre. Cependant, la démarche de Double, et la notion même de « métaphilosophie », contiennent à mon sens des ambiguïtés et même des erreurs qui m'amènent à préférer la catégorie « d'épistémologie de la philosophie ». Je commencerai par exposer la démarche de Double, puis j'exposerai toutes les difficultés que j'y vois.

Pour Richard Double, une « métaphilosophie » est une certaine conception de la finalité de la philosophie ; c'est ce *pour quoi* un auteur choisit de faire de la philosophie. Il existe essentiellement quatre grandes métaphilosophies :

1. la *philosophie comme conversation*, qui « correspond à la conception de l'art comme un enrichissement non tourné vers la vérité, c'est-à-dire comme une activité qui contribue à nos vies intellectuelles sans se prononcer sur le caractère de la réalité ultime. » (Double 1996, 24)
2. la *philosophie comme praxis*, qui « correspond à la conception de l'art comme contribuant à la moralité en prétendant que la philosophie devrait être un instrument pour faire de nous de meilleures personnes. » (Double 1996, 25)
3. la *philosophie comme apportant un soubassement : au Sens Commun, à la Religion, à la Loi, aux Sciences Naturelles, ou aux Sciences Spéciales*, qui « conçoit le rôle propre de la philosophie comme un support d'un autre domaine, ou d'autres domaines, d'intérêt intellectuel. » (Double 1996, 26)
4. la *philosophie comme construction de Vision du Monde* (Worldview), qui « considère que le but de la philosophie est semblable à la conception du but de la science par le réaliste scientifique : une tentative pour caractériser la réalité aussi précisément que nous le pouvons. » (Double 1996, 29) C'est cette conception seulement que l'on peut considérer comme « orientée vers la vérité » (*truth-tracking*) (ibidem).

D'après Double, les grands désaccords sur la question du libre arbitre reposent fondamentalement sur des choix différents en matière de métaphilosophie. Par exemple, pour la question de savoir si nous avons une justification *prima facie* à croire que nous sommes libres (indépendamment de la question de compatibilité), Double (1996, 56–60) estime que la seule justification qu'on pourrait avoir est la justification de van Inwagen par la responsabilité morale (nous sommes moralement responsables, donc libres). Mais cela suppose que la croyance à la moralité soit à son tour justifiée : cela sera le cas si l'on a une conception de la philosophie comme *praxis* (croire à la moralité est important pour se comporter moralement), ou une conception de la philosophie comme soubassement du Sens Commun (la responsabilité morale fait partie des éléments essentiels du Sens Commun). Mais si l'on a une conception de la philosophie comme soubassement des *Sciences*, on n'aura probablement pas

cette justification *prima facie* en faveur de la moralité et du libre arbitre. A cause d'une différence de métaphilosophie, les *données de base* du problème seront différentes, et on arrivera donc certainement à une conclusion générale différente. L'étude de Double est très minutieuse et fait apparaître plusieurs sous-catégories, de telle sorte qu'on ne peut pas dire de manière simple et directe que (par exemple) la *philosophie comme conversation* amènerait au scepticisme, la *philosophie comme praxis* au compatibilisme, la *philosophie comme soubassement* du Sens Commun au libertarianisme, et la philosophie comme vision du monde amène au déterminisme dur. Mais l'idée centrale de Double est bien celle-là : les divergences sur la question du libre arbitre reposent fondamentalement sur des différences de métaphilosophie.

Que penser de cette approche métaphilosophique, et quel est son rapport avec l'approche que j'adopte dans la présente étude ? En un sens, on pourrait dire que l'approche de Double est tout simplement indépendante de la mienne, et repose sur une définition différente de ce qu'est la « métaphilosophie ». On pourrait en effet distinguer trois *sens* de « métaphilosophie » :

- a. dans un premier sens, la métaphilosophie est l'étude de ce qu'est la philosophie, de ce que veut dire le mot « philosophie ». Ce mot, apparu historiquement en Grèce environ cinq siècles avant Jésus-Christ, a pu prendre des significations différentes au cours de son évolution. La métaphilosophie consisterait alors à appliquer la méthode philosophique de « tentative de définition » à l'objet « philosophie ».
- b. dans un deuxième sens (celui de Double), une métaphilosophie est le choix d'une certaine attitude ou d'un certain objectif : en abordant le thème du libre arbitre, je peux choisir de « rechercher la vérité », ou je peux choisir de « parler pour le plaisir esthétique de la conversation », ou encore « d'exhorter à agir moralement ».⁴⁵⁵
- c. dans un troisième sens enfin (celui qui correspond à ma démarche), la métaphilosophie est la réflexion philosophique sur les bonnes méthodes à employer pour rechercher la vérité dans les questions philosophiques.

Ce troisième sens de « métaphilosophie » présuppose une conception de la philosophie comme « orientée vers la vérité », et constitue à partir de cette conception une forme de « philosophie de la philosophie », c'est-à-dire un ensemble de considérations orientées vers la vérité sur la question « comment aborder les questions philosophiques de manière à avoir le plus de chance d'être dans la vérité ? ». Cette analyse philosophique (orientée vers la vérité) des méthodes de la philosophie (orientée vers la vérité) est, par exemple, au centre de l'ouvrage de Williamson *The Philosophy of Philosophy* (Williamson 2007), de l'étude par Joel Pust de l'usage des intuitions à la base des arguments philosophiques (Pust 2000), ou encore de la collection *Metametaphysics* (Chalmers, Manley, and Wasserman 2009) qui s'intéresse plus spécifiquement à la question des méthodes à adopter *en métaphysique*. Lorsqu'on regarde ces études de plus près, il est clair que leur contenu est fortement épistémologique (quel est la valeur épistémologique des *intuitions* ? y a-t-il une connaissance *a priori* ? comment les inférences à la meilleure explication sont-elles justifiées ? etc.) Et il n'est donc pas étonnant que la métaphilosophie, dans ce troisième sens, soit souvent pratiquée par des épistémologues. De même, dans la présente étude, je m'interroge sur la question de

⁴⁵⁵ Notez que ce deuxième sens est indépendant du premier : on peut très bien penser que le mot « philosophie » n'a jamais voulu dire « esthétique de la conversation » et pourtant avoir pour soi-même cet objectif lorsqu'on a abordé les problèmes philosophiques. En parlant de différentes « conceptions de la philosophie », Double ne fait pas clairement la distinction entre les sens a. et b. : une conception semble être à la fois « ce que l'on conçoit que la philosophie est » (sens a.) et « la manière dont on conçoit sa propre activité lorsqu'on aborde des questions philosophiques » (sens b.). Mais c'est clairement le sens b. qui intéresse fondamentalement Double.

savoir « comment telle ou telle thèse dans le débat sur le libre arbitre pourra être justifiée, ou pourra constituer une connaissance ? » et les outils conceptuels auxquels j'ai recours sont des outils d'épistémologie. Compte tenu de l'ambiguïté du terme « métaphilosophie », il me semble donc plus clair de parler, pour cette troisième démarche spécifiquement, « d'épistémologie de la philosophie ». ⁴⁵⁶

En un sens, donc, le terme de « métaphilosophie » est simplement ambigu : il peut désigner aussi bien la démarche d'épistémologie de la philosophie (orientée vers la vérité) que la démarche de Double de « choix d'un projet », qui n'a tout simplement rien à voir.

Mais en fait, la présentation de Double me paraît problématique pour des raisons beaucoup plus graves qu'une simple ambiguïté portant à confusion. Il me semble en effet que sa présentation du problème a des implications qui sont absolument fausses. La présentation de Double a pour conséquence de donner une représentation tout à fait erronée des désaccords qui existent sur la question du libre arbitre, et une également une représentation erronée de la philosophie comme orientée vers la vérité. Je soulèverai ici cinq erreurs importantes.

Premièrement, si l'on suivait Double, on devrait conclure que les désaccords en philosophie du libre arbitre ne sont pas de véritables désaccords. En effet, si ces désaccords reposent sur des « attitudes » aussi fondamentalement différentes que « parler pour le plaisir esthétique de la conversation », « exhorter moralement » ou « chercher la vérité », alors il n'y a aucune contradiction possible entre deux auteurs qui ont recours à des conceptions différentes du libre arbitre. Si par exemple un auteur défend que « le compatibilisme est une théorie qui nourrit la conversation de manière élégante », un autre que « le libertarianisme est bon pour promouvoir des actions morales », et un troisième que « le déterminisme dur est probablement vrai », où est la contradiction entre eux ? Ces trois auteurs pourraient tout à fait avoir raison en même temps. Il n'y a aucune sorte « d'opposition » entre la représentation d'un corps humain à la manière cubiste par Picasso, et celle d'un traité de physiologie : ces deux représentations ont des finalités trop différentes pour pouvoir s'opposer l'un à l'autre. Cette conséquence de la théorie de Double me paraît absolument fautive : les différents participants dans le débat sur le libre arbitre *s'opposent* bien les uns aux autres ; ils défendent des théories *concurrentes* et non pas juxtaposées.

Deuxième point, qui est probablement l'explication du premier : Richard Double semble penser qu'une partie seulement des philosophes du libre arbitre a une visée orientée vers la vérité, et que les désaccords viennent largement de l'existence d'autres visées non aléthiques. Autrement dit, si tout le monde poursuivait l'objectif de donner une description vraie de la réalité, il n'y aurait probablement pas tant de désaccords sur le libre arbitre. Cela aussi me semble profondément erroné : une description honnête du débat (en tout cas du débat analytique) en métaphysique du libre arbitre doit reconnaître que l'immense majorité des philosophes cherche à donner une théorie « vraie » ou « conforme à la réalité », sans doute avec des méthodes et des présupposés différents, qui mènent à des conclusions différentes, mais toujours d'une manière ou d'une autre avec la finalité générale de bien décrire la réalité. Et c'est bien dans le cadre de cet objectif commun qu'une grande divergence de vues reste possible. ⁴⁵⁷ Ce deuxième point est relié au premier de la manière suivante : c'est bien parce

⁴⁵⁶ Bien que, dans un monde idéal, je préférerais que le terme « métaphilosophie » soit réservé à ce troisième sens.

⁴⁵⁷ Pour être précis, Double n'exclut pas totalement que la conception de la philosophie comme « construction d'une Vision du monde » (c'est-à-dire orientée vers la vérité) puisse mener à différentes conclusions. En effet, il considère qu'il existe deux grandes versions de cette métaphilosophie, à savoir la « philosophie comme en continuité avec la science » et la « philosophie comme pas en continuité avec la science » (Double 1996, 30). La philosophie comme en continuité avec la science a pour particularité de considérer que la bonne manière de chercher la vérité est d'appliquer les méthodes générales de la science. Ainsi, avec le même objectif de la vérité, on pourra arriver à différents résultats. Mais il est significatif de voir ce qui, d'après Double, pourrait amener les défenseurs de la deuxième conception (la philosophie comme pas en continuité avec la science) à une conclusion

qu'ils ont (presque) tous la visée du vrai que les différents philosophes du libre arbitre sont en véritable désaccord, ou en contradiction, les uns avec les autres.

Troisièmement, Double présente les désaccords sur le libre arbitre non seulement comme de faux désaccords, et comme résultants d'une variété d'objectifs différents de la visée du vrai, mais en outre il présente ces désaccords comme reposant sur un choix d'attitude qui est *arbitraire*. En effet, Double prend la peine de montrer qu'aucune métaphilosophie ne peut être « prouvée » (1996, sec. 2.5), mais c'est clairement une conséquence triviale de la conception des métaphilosophies comme différentes « attitudes » ou différentes « finalités » : si quelqu'un souhaite utiliser le langage d'une manière poétique ou esthétique, il n'y a là rien de meilleur ou de pire que d'utiliser le langage dans une visée aléthique ; c'est simplement un projet différent et non concurrent. Puisque le choix de tel ou tel objectif est arbitraire, et puisque la réponse au problème du libre arbitre dépend de ce choix, il est clair que la réponse au problème du libre arbitre sera elle-même ultimement arbitraire. Je suis en profond désaccord avec cette thèse de Double. Certes, si un poète souhaite écrire un sonnet sur le libre arbitre, il pourra bien, arbitrairement, « défendre » la conception qui sert le mieux ses intérêts esthétiques. Mais dans le débat *philosophique* sur le libre arbitre, la norme fondamentale est celle du vrai. Faire de la métaphilosophie, dans le sens que je défends (le sens de l'épistémologie de la philosophie), ce n'est certainement pas relativiser les principes fondamentaux sur lesquels repose telle ou telle thèse. Au contraire, l'épistémologie de la philosophie recherche des principes normatifs universels – par exemple, le principe évidentialiste (chapitre 2, sec. II.1.c), ou le principe de méta-cohérence (chapitre 3, sec. II.3) – qui permettront éventuellement de faire un arbitrage objectif entre les différentes thèses. D'une manière plus générale, la conception de Double a pour conséquence de donner une vision faussée de la philosophie comme visée du vrai. Les deux points suivants développent cette erreur particulière.

Quatrièmement, la philosophie comme visée du vrai est présentée par Double comme un projet parmi d'autres, qui n'intéresserait pas, en particulier, les défenseurs de la « philosophie comme soubassement du Sens Commun ou des Sciences ». Mais cela revient à donner une vision caricaturale de ces attitudes philosophiques : un auteur qui considère que la philosophie ne pourra jamais apporter de réfutation du Sens Commun et doit se contenter d'en expliciter les fondements partage entièrement la vision de la philosophie comme orientée vers la vérité. De même qu'un auteur qui considère que la philosophie ne pourra jamais apporter de résultat propre et doit se contenter d'expliciter les fondements et les résultats des Sciences. Ces deux types d'auteurs ne diffèrent pas sur la visée plus ou moins aléthique de l'entreprise philosophique, mais seulement sur la meilleure manière de parvenir à la vérité. Une nouvelle fois, on peut faire remarquer que leur désaccord est proprement *épistémologique*. Les uns estimeront que le Sens Commun se trouve en position de fondement de toute la structure de notre connaissance, de sorte qu'il est impossible de réviser ces fondements sur la base d'éléments ou de raisonnements qui reposent ultimement sur eux. Les autres estimeront que les fondements de la connaissance sont de nature strictement empirique ou scientifique et qu'il est donc illusoire de chercher d'autres données que celles des sciences.

Cinquièmement, Double soutient non seulement que la philosophie comme visée du vrai est un objectif parmi d'autres, mais aussi que cet objectif *présuppose* certaines thèses

différente des premiers : « Philosophy as Non-Continuous with Science thinkers may assign weight to various non-truthtracking considerations that are important to the other metaphilosophies. » (Double 1996, 31). Il semble que, pour Double, adopter une autre conception de la « visée du vrai » que la conception « en continuité avec la science » revienne toujours à mâtinier les considérations orientées vers la vérité par des considérations non orientées vers la vérité, et que les désaccords possibles viennent finalement de cette réinsertion de considérations non orientées vers la vérité. En somme, même au sein des philosophies orientées vers la vérité, le désaccord ne pourrait venir que d'une plus ou moins grande orientation vers la vérité.

substantielles sur les capacités de l'esprit humain : « Cette conception présuppose que la philosophie peut faire des approximations plus ou moins bonnes de la vérité réelle et que faire cela est son meilleur objectif. » (Double 1996, 29) Mais la visée du vrai n'a, en réalité, aucun présupposé particulier. On peut très bien avoir une attitude orientée vers la vérité et estimer qu'il nous est impossible d'atteindre la moindre approximation de la vérité : dans ce cas, on adoptera une position *sceptique*. Le scepticisme n'est pas moins orienté vers la vérité que le plus optimiste des fondationalismes philosophiques. De même, on peut très bien avoir une attitude orientée vers la vérité et estimer que seules les sciences parviennent à découvrir des vérités, à l'exclusion de toute autre méthode qui serait proprement « philosophique » : dans ce cas, on adoptera une conception de la philosophie comme simple soubassement des sciences ; et comme on vient de le voir, cette conception n'est pas moins orientée vers la vérité que la démarche philosophique la plus a-prioriste.

Par conséquent, la démarche « métaphilosophique » de Richard Double me semble à la fois trompeuse et tout bonnement erronée sur plusieurs points. Si elle consiste simplement à distinguer des « attitudes » que l'on pourrait arbitrairement choisir dans la considération du problème du libre arbitre, elle est indépendante de la démarche « métaphilosophique » que j'adopte, puisque toute mon enquête repose sur l'unique attitude de la visée du vrai, le niveau « méta » étant simplement une interrogation sur les bonnes méthodes épistémologiques à mettre en œuvre pour atteindre le vrai. Mais Double ne se limite pas à envisager d'autres finalités possibles, à côté de la visée du vrai : sa démarche a pour conséquence de relativiser considérablement cette visée du vrai. Contrairement à Double, j'estime que la visée du vrai est en fait l'objectif commun d'à peu près tous les philosophes du libre arbitre, qu'elle est en particulier l'objectif des partisans du Sens Commun ou de ceux qui voient la philosophie comme simple soubassement de la science, et que les désaccords nombreux sur le libre arbitre sont des désaccords véritables entre personnes qui partagent cette même visée du vrai.

Annexe 3 : Table récapitulative des objections épistémologiques à l'introspectionnisme libertarien

Objections	de premier ordre (contre <i>Jp</i>)	méta-épistémique (contre <i>JKp</i>)
préscientifiques	A. pure impression (ch.2, II.1) impression faillible (ch.2, II.2) impression lourde (ch.2, II.3)	B. impression non étayée (ch.3, II) conséquences lourdes non étayables (ch.3, III)
spécialisées (avec analyse conceptuelle et/ou données empiriques)	C. argument de la chance (ch.5, I, 1) argument basique (ch.5, I, 2) preuves directes du déterminisme (ch.5, II, 1) argument d'intelligibilité empirique (ch.5, II, 2)	D. arguments de la source faible (Libet, Wegner) (ch.6, I) problème d'Intégration de l'expérience indéterministe (ch.6, II)

principe général de l'objection		solution au scepticisme, qui résout du même coup le problème pour l'introspectionnisme.
suite de l'objection pour l'introspectionnisme libertarien	argument analogue sceptique	

A. Objections préscientifiques de premier ordre (chapitre 2)

pure impression (ch.2, II.1)

objection	analogie sceptique	solution favorisée
(Infer.) Une croyance ne peut être justifiée que par inférence à partir d'autres croyances elles-mêmes justifiées.		(Fond.) Une croyance peut être justifiée : - soit par inférence à partir d'autres croyances justifiées - soit basiquement sur la base d'une impression
(PI-2') Une croyance introspective au libre arbitre ne repose que sur une pure impression. (C) Donc une croyance introspective au libre arbitre n'est pas justifiée.	(PI-3) Une croyance perceptive basique ne reposerait que sur une pure impression (C) Donc une croyance perceptive basique ne serait pas justifiée.	

impression faillible (ch.2, II.2)

(Infal.) S'il est métaphysiquement possible qu'une expérience doxastique soit non véridique, alors elle ne confère pas de justification épistémique à son contenu.		(faillib. fort) Ma croyance que p peut constituer une connaissance même si elle est vulnérable à des possibilités d'erreur de type courant, i.e. un type d'erreur que j'ai déjà eu l'occasion de détecter pour moi-même.
(2) Il est métaphysiquement possible que l'expérience E d'être libre soit non véridique. (C) L'expérience E d'avoir le libre-arbitre n'apporte aucune justification à croire au LA.	(P) Il est métaphysiquement possible que nos expériences perceptives soient non véridiques. (C) Nos expériences perceptives n'apportent aucune justification à croire à leur contenu.	

impression lourde (ch.2, II.3)

(J-Clot.*) S est justifié à croire que p seulement si S est justifié à croire à toutes les conséquences logiques qu'il reconnaît à p de manière justifiée.		(transmission de J_{pf}) Lorsque je suis justifié <i>prima facie</i> à croire une proposition anodine A et que je comprends que A implique logiquement une conséquence anti-sceptique $\sim HS$, je suis justifié <i>prima facie</i> à croire $\sim HS$ par déduction compétente à partir de A .
(1) Nous n'avons aucune justification internaliste à rejeter le scénario de manipulation martienne. (2*) Nous sommes justifiés à croire que le libre arbitre est incompatible avec le scénario de manipulation martienne. (C) Nous ne sommes pas justifiés (par notre expérience interne) à croire au libre arbitre.	(1) Nous n'avons aucune justification internaliste à rejeter le scénario sceptique des cerveaux dans des cuves. (2*) Nous sommes justifiés à croire que les croyances perceptives anodines sont incompatibles avec le scénario sceptique. (C) Nous ne sommes pas justifiés à croire nos expériences perceptives anodines.	

B. Objections préscientifiques de deuxième ordre (chapitre 3)

impression non étayée

(MC+) Si <i>S</i> a une justification <i>prima facie</i> à croire que <i>p</i> sur la base d'un fondement épistémique <i>F</i> , mais n'a pas de justification non circulaire à croire que ce fondement peut constituer un fondement de <i>connaissance</i> , alors sa justification à croire que <i>p</i> est fondationnellement réfutée.		(Méta-circularité) Lorsque l'on soulève la question de la méta-justification d'une croyance (de la fiabilité de sa source), il est légitime d'apporter une méta-justification circulaire si elle repose sur : - une preuve de naturalité de la source ou - une intuition mooréenne étayant la source et s'étayant elle-même.
(NE) Nous n'avons pas de justification non circulaire à croire que notre expérience du libre arbitre peut fournir une connaissance. (C) La justification <i>prima facie</i> apportée par notre expérience du libre arbitre est réfutée fondationnellement.	(NEP) Nous n'avons pas de justification non circulaire à croire que nos expériences perceptives peuvent fournir une connaissance. (C) La justification <i>prima facie</i> apportée par nos expériences perceptives est réfutée fondationnellement.	

conséquences lourdes non étayables

(Principe d'indéfectibilité) Si mes expériences ne me permettent pas de détecter si <i>p</i> est vraie, alors je ne peux pas <i>savoir</i> que <i>p</i> est vraie.		(Sûreté) Pour connaître les conséquences logiques d'une expérience donnée, il n'est pas nécessaire que cette expérience soit <i>sensible</i> à la conséquence en question ; il suffit que cette conséquence soit <i>sûre</i> .
(Indéfect.) Nous ne pouvons pas détecter que $\sim MM$ est vraie. (C1) Donc nous ne pouvons pas savoir $\sim MM$ (C2) Donc (par engagement de méta-cohérence et de clôture) notre justification <i>prima facie</i> à croire au libre arbitre est réfutée.	(Indéfect.P) Nous ne pouvons pas détecter que $\sim HS$ est vraie. (C1) Donc nous ne pouvons pas savoir $\sim HS$ (C2) Donc (par engagement de méta-cohérence et de clôture) notre justification <i>prima facie</i> à croire nos perceptions est réfutée.	

C. Objections spécialisées de premier ordre (chapitre 5)

Objections dont le cœur n'est pas épistémologique ; sans analogue sceptique avec le cas de la perception.

D. Objections spécialisées de deuxième ordre (chapitre 6)

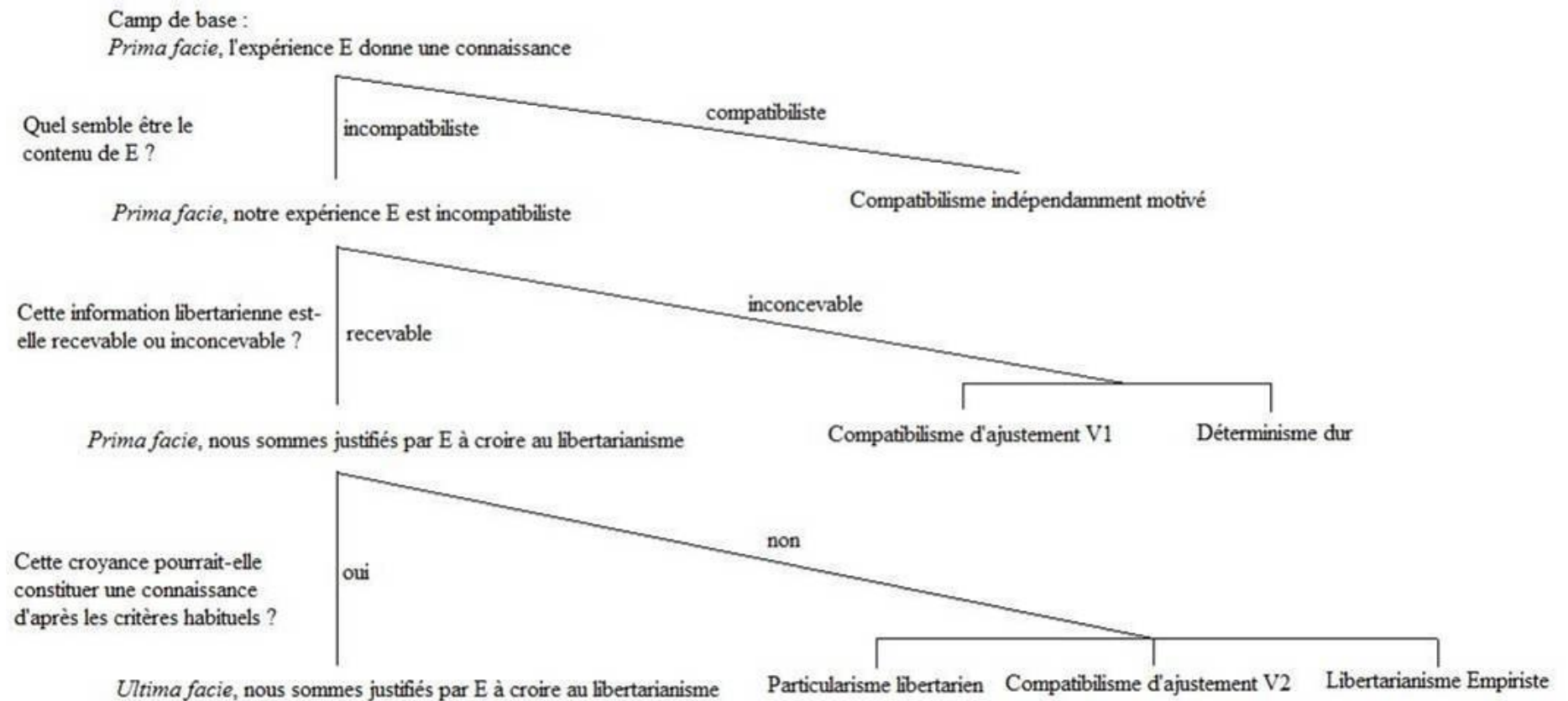
L'argument d'erreur de Wegner

(Infal.) Si une expérience doxastique fait l'objet d'erreurs courantes, elle est généralement discréditée.		(faillib. fort) Ma croyance que p peut constituer une connaissance même si elle est vulnérable à des possibilités d'erreur de type courant, i.e. un type d'erreur que j'ai déjà eu l'occasion de détecter pour moi-même.
(Illus.) Les neurosciences et la psychologie ont accumulé un grand nombre d'erreurs de l'expérience d'agir. (C) Donc l'expérience d'agir en général est une illusion.	(Illus.P) Les neurosciences et la psychologie ont accumulé un grand nombre d'erreurs de l'expérience perceptive. (C) Donc la perception en général est une illusion.	

Le Problème d'Intégration

(Etayage) Si nous n'avons pas d'indication de la manière dont une expérience dépend fiablement de son objet, cette expérience ne peut être considérée comme connaissance et ne peut donc apporter de justification (<i>ultima facie</i>).		(Modèles d'Intégration) Pour qu'une expérience apporte une justification, il suffit que l'on puisse donner un <i>modèle</i> dans lequel elle dépend fiablement de son contenu (et dans lequel ses conséquences logiques sont sûres), si l'on a par ailleurs une méta-justification pour cette expérience.
(Dép.) Pour autant que nous pouvons le savoir, notre expérience du libre arbitre ne dépend pas fiablement de son contenu. (C) Donc notre expérience ne peut nous apporter une justification.	(Dép.P) Pour autant que nous pouvons le savoir au Vème siècle av. J.C., nos expériences visuelles ne dépendaient pas fiablement des faits visuels perçus. (C) Donc nos expériences visuelles ne pouvaient pas nous apporter de justification au Vème siècle av. J.C.	

Annexe 4 : Les grandes positions possibles dans le débat sur le libre arbitre



Annexe 5 : Van Inwagen on Introspected Freedom⁴⁵⁸

Any philosopher who defends Free Will should have an answer to the epistemological question: “how do we know that we have such a capacity?” A traditional answer to this question is that we have some form of introspective access to our own Free Will. In recent times though, many philosophers have considered any such introspectionist theory as so obviously wrong that it hardly needs discussion, especially when Free Will is understood in libertarian terms. One of the rare objections to appear as an explicit argument was proposed by van Inwagen in his *Essay on Free Will*. In this paper, I address van Inwagen’s anti-introspection argument; I argue that it is both inconsistent with his overall treatment of the Existence Question (namely, with his defence of the existence of Free Will from reflections about morality), and inconclusive in itself (at least for anyone not ready to endorse general scepticism about perception). In passing, I give a clarification of the notion of Introspection, in the case of Freedom, that also sets a more favourable stage for the evaluation of further objections.

A Libertarian, in the traditional sense, is someone who accepts not only that free will is incompatible with determinism, but also that free will exists. There is a large agreement among libertarians as to the proper method to ascertain the first half of this view: incompatibilism is to be argued for on conceptual grounds, by philosophical reasoning. But there is much less agreement on how the second half – the “Free Will Thesis”, to use van Inwagen’s terminology – could and should be justified. How could a libertarian claim to *know* that he is free?

An answer that seems to naturally come to mind is that we just *feel* that we’re free. In more philosophical terms, a libertarian could claim that we have some kind of introspective access to, or direct experience of, our own freedom. And some prominent libertarians, like Reid⁴⁵⁹, Campbell⁴⁶⁰ or O’Connor⁴⁶¹, have indeed defended varieties of such a view⁴⁶².

⁴⁵⁸ Article publié dans les *Philosophical Studies* en 2014 (Guillon 2014).

⁴⁵⁹ “In certain motions of my body and directions of my thought, I know, not only that there must be a cause that has power to produce these effects, but that I am that cause; and I am conscious of what I do in order to the production of them. From the consciousness of our own activity, seems to be derived, not only the clearest, but the only conception we can form of activity, or the exertion of active power.” (Reid 1788, 36). And of course, Reid’s notion of agent-causation (or “active power”) is an incompatibilist one, since according to him: “power to produce any effect implies power not to produce it” (Reid 1788, 35).

⁴⁶⁰ “Can we *trust* the evidence of inner experience? [...] The misguided, and as a rule quite uncritical, belittlement, of the evidence offered by inner experience has, I am convinced, been responsible for more bad argument by the opponents of Free Will than has any other single factor. [...] Are there objections to a freedom of this [libertarian] kind so cogent that we are bound to distrust the evidence of ‘inner experience’? [...] The arguments which seem to carry most weight with Determinists are, to say the least of it, very far from compulsive.” (Campbell 1957, 168–174)

⁴⁶¹ “It does not seem to me (at least ordinarily) that I am caused to act by the reasons which favor doing so; it seems to be the case, rather, that I produce my decision in view of those reasons, and could have, in an unconditional sense, decided differently. [...] then it is natural for the agency theorist to maintain that they involve the perception of the agent-causal relation” (O’Connor 2000, 30).

⁴⁶² Other recent defenders of some variety of introspectionism include Richard Swinburne and Keith Lehrer. For the former, see (Swinburne 2011, 82): “It is a basic principle of rationality that things are probably the way they seem to be (in the sense that we are inclined to believe that they are) in the absence of counter-evidence. [...] When we make a decision, it seems that we choose and are not caused to choose as we do.” For Keith Lehrer, see (Lehrer 1960, 156): “The introspective datum that we deliberate about future actions provides adequate evidence for believing that some of our actions are in our power, that is, that we can choose to perform the action and also that we can choose not to perform the action, and in this sense we have adequate evidence that we have

But there are other answers on the market: some libertarians have claimed to know about freedom from reflections about morality (van Inwagen 1983), from the crucial role that this belief plays in our system of beliefs (Donagan 1987), or from some kind of self-referential or transcendental argument (Jordan 1969). Other authors, who could be called libertarians in the weak sense that they endorse a libertarian concept of free will, even though they remain agnostic about whether we do have free will, consider that the answer could only come from future empirical research (Balaguer 2010; cf. Kane 1996). Most of these libertarian authors join with compatibilists and hard-incompatibilists in pronouncing a sweeping condemnation of any introspectionist strategy for libertarianism. Such pronouncements are typically not very long and not always argumentative in form, because of the wide consensus expected among readers. In this paper, I want to address one explicit anti-introspection argument, proposed by Peter van Inwagen in the *Essay on Free Will*.

I shall argue that van Inwagen's argument against Introspectionism suffers from two defects. First, it is inconsistent with van Inwagen's own strategy in answering the Existence Question, for *if* the anti-introspection argument works, so should a parallel argument devised against our knowing Free Will from reflections about morality. But van Inwagen wants to maintain that this parallel objection against knowledge from morality *doesn't* work. So the anti-introspection argument shouldn't work either (sections II to IV). Second, I shall argue that the argument in itself is unsuccessful, at least for anyone not prepared to swallow general scepticism about perception (section V).

I. Van Inwagen's anti-introspection argument

Van Inwagen's objection to Introspectionism goes as follows:

Some philosophers believe that it is possible to find out whether we have free will by *introspection*. But this seems just obviously wrong, since, if it were right, we could find out by introspection whether we were fitted out with Martian devices like those we imagined in our discussion of the Paradigm Case Argument in Chapter IV. And, of course, we cannot do this. (van Inwagen 1983, 204)

The form of the argument is obvious enough:

- (1) If we could know by introspection that we are free, then we could know by introspection (plus *a priori* reasoning) that we are not fitted out with Martian devices.⁴⁶³
- (2) We cannot know by introspection (plus *a priori* reasoning) that we are not fitted out with Martian devices.

From (1) and (2) follows by modus tollens:

- (C) We cannot know by introspection that we are free.

free will." It is a disputable matter whether Lehrer's conception is ultimately as introspectionist as he himself thinks it is, but this is not the place to discuss his particular view.

⁴⁶³ I assume that what van Inwagen has in mind with the phrase "find out by introspection" is the possibility to *deduce* that "we are not fitted out with Martian devices" from the (hypothetical) introspective knowledge of "we are free" and the (presumably *a priori*) knowledge that freedom excludes Martian manipulation. The resulting knowledge, therefore, wouldn't be strictly introspective, or "by introspection" full stop, but rather "by introspection plus *a priori* reasoning".

The acceptability of the premises depends on two crucial features of the Martian Devices here mentioned (hereafter MD). First, MD are used by Martians to manipulate every single one of our choices, and therefore they are logically inconsistent with our having free will (even in most compatibilist construals of free will)⁴⁶⁴. In other words:

(a) we are fitted out with MD => we are not free

or conversely:

(a') we are free => we are not fitted out with MD

This feature, together with a principle of epistemic closure, gives support to premise (1). There are of course various principles in the literature which attempt to capture the core of the closure intuition, and not all of them are very cogent. What we need here, though, is only the weak and very plausible version which hinges on the notion of Competent Deduction⁴⁶⁵:

(CD-Closure) If x knows that p , and x is in a position to acquire the belief that q by competently deducing q from p (due to *a priori* considerations), then x is in a position to thereby come to know that q .⁴⁶⁶

The second crucial feature of MD, is that they generate in us whatever “feeling” or “experience” of free will an introspectionist may purport to have.⁴⁶⁷ In other words:

(b) the experience of freedom we would have in the MD scenario is subjectively indiscernible from the experience of freedom we would have if we were indeed free.

And feature (b) is brought in support to premise (2).

II. Analogy between van Inwagen’s argument and the sceptical argument

There is an obvious analogy between this argument and the famous family of sceptical arguments based on the same principle of closure. For comparison, just consider:

⁴⁶⁴ See (van Inwagen 1983, 109): “whenever that person must make a decision, the device *causes* him to decide one way or the other according to the requirements of a table of instructions.”

⁴⁶⁵ It should also be noted that what we need is a principle of closure for *knowledge* rather than closure for *justification* alone, or some other epistemic notion.

⁴⁶⁶ Competent Deduction Closure was famously defended by Hawthorne (2004; 2005). I have chosen to use a version of CD-Closure slightly different from Hawthorne’s exact wording for two main reasons. First, for simplicity, I have omitted some cautious qualifications that don’t play any particular role in our context. Second, in order to fit the “modal” wording of van Inwagen’s argument (“if we *could* know... then we *could* find out...”), I have translated in dispositional terms (“ x is in a position to competently deduce...” what Hawthorne states in conditional terms (“if x competently deduces...”).

⁴⁶⁷ See (van Inwagen 1983, 109): “someone might object that [MD] is not in fact consistent with our observations, since we can normally “feel” our decisions “flowing” naturally from our desires and our beliefs; but if [we were fitted out with MD], we should “feel” ourselves being interfered with. But to meet this objection we need only suppose that the Martian device causes us to have desires and beliefs appropriate to the decisions it will cause us to make.”

- (1*) If we could know by perception that we have hands, then we could know by perception (plus *a priori* reasoning) that we are not brains in vats.
- (2*) We cannot know by perception (plus *a priori* reasoning) that we are not brains in vats.
- (C*) We cannot know by perception that we have hands.⁴⁶⁸

The premises are supported by the following two features of the brains-in-vats scenario:

- (a*) we have hands => we are not brains in vats
- (b*) the experience as of having hands we would have in the brains-in-vats scenario is subjectively indiscernible from the experience as of having hands we would have if we did have hands.

More precisely, premise (2*) is supported by feature (b*). And premise (1*) is supported by feature (a*) together with the Closure Principle for Competent Deduction: anyone who gives (a*) a thought is able to come to know it to be true *a priori*, and to competently produce the corresponding modus ponens.

This analogy quite straightforwardly suggests a way of defending the Introspectionist stance by the following analogical argument:

Despite its plausibility, the anti-introspection argument here mentioned cannot be right, because it is just analogous to the anti-perception sceptical argument. And of course, the latter must be wrong some way or other. So whatever defeater we choose to rebut scepticism will also defeat the anti-introspection argument.

⁴⁶⁸ The detailed formulation of the argument was chosen in order to elicit the analogy with van Inwagen's argument. One may wonder whether this formulation is really equivalent to more usual formulations, in particular with formulations that do not mention the *modes* of knowledge, and require only that we should have knowledge (of *whatever kind*) that we are not brains in vats (hereafter ~BIV) (see Pritchard 2005, 28; DeRose and Warfield 1999, 2). The main worry here is that our formulation may be a weak one if it ignores the possibility of our having some kind of independent or antecedent knowledge of ~BIV (knowledge in no way perceptually based, presumably *a priori*). I shall discuss this possibility in part V. For now, it need only be noticed that this possibility isn't *ignored* but *negated* by one of the premises – namely (1*) – so that it remains a possible way to refuse the argument. Here is how the dialectics should be understood. I take it that a developed account of the closure-based reasoning – like that of (Avnur 2012, 2) – has three premises rather than two, namely: (4) We don't have antecedent (*a priori*) knowledge of ~BIV; (5) we cannot legitimately come to know ~BIV by deduction from everyday perceptual premises (because this would constitute "Easy Knowledge"); (6) if we had everyday perceptual knowledge, Competent Deduction Closure would allow us to come to know any trivial consequence of such knowledge which we wouldn't antecedently know (in particular ~BIV, according to (4)); therefore, we don't have everyday perceptual knowledge (or at least can't retain it upon reflection). The most classical two premise versions of the closure-based argument conjoin premises (4) and (5) to assert that we cannot have *any kind* of knowledge of ~BIV (neither antecedent nor resultant). Our version, rather, conjoins premises (4) and (6) as premise (1*): if we know by perception that we have hands, then given our lack of antecedent knowledge of ~BIV, Competent Deduction would allow us to acquire easy knowledge of ~BIV. But, second premise ((premise (5) or (2*)), easy knowledge is illegitimate. This way of presenting the sceptical reasoning obviously lays the stress on Easy Knowledge – and for that reason, it comes closest to (Cohen 2002)'s discussions. But it should be noticed for our discussion that this presentation differs from more traditional ones *only in emphasis* or stress: any reply to the traditional formulations of the argument are available to our formulation as well. In particular, the "antecedent knowledge" reply, whatever its plausibility may be, is available by refusing the conditional premise (1*).

But analogy isn't identity. And though such a defence by analogy may have some intuitive pull, still there *might* be a defeater of the sceptical argument that wouldn't also be a defeater of the anti-introspection argument.

Nevertheless, I believe the comparison with the sceptical argument can be quite devastating for van Inwagen's discussion, for two reasons I first present in a nutshell.

First, because van Inwagen uses just such an argument by analogy in order to defend his own conception of "knowledge from morality" (against a parallel objection). But if some analogy has any argumentative value, so should a *closer* analogy. And the introspectionist case is, if anything, *more* analogous to the perception case than is "knowledge from morality". Therefore, *if* van Inwagen's analogical defence of "knowledge from morality" works, *a fortiori* should the analogical defence of introspective knowledge work as well. Or conversely, *if* van Inwagen's objection against Introspection worked, *so* would the parallel objection against his version of knowledge from morality (sections III and IV).

Second, I shall argue that, if we properly construe the case for introspective knowledge of free will, the anti-introspection argument isn't really an *analogon* of sceptical anti-perception arguments but rather an *instance* of them. Therefore, if one doesn't want to be committed to scepticism, one has to consider that van Inwagen's argument against Introspection fails (section V).

Let us examine those two points more carefully.

III. Problem of consistency for van Inwagen's general strategy

My first point is that van Inwagen's *acceptance* of the anti-introspection argument is inconsistent with his own *rebuttal* of a parallel argument against knowledge from morality. Here is the objection he addresses (and rebuts) against knowledge from morality. The objector speaks:

If all your arguments are correct, then our (alleged) knowledge of the existence of moral responsibility, coupled with certain arguments *a priori*, can constitute a good reason for believing that determinism is false. But these things are not the *sorts* of things that can be a good reason for believing in indeterminism. [...] Therefore, your arguments represent just one more attempt by a philosopher to settle by intellectual intuition and pure reason a question that should be left to empirical science. (van Inwagen 1983, 210)

Again, it is not difficult to get at the structure of the argument:

(1**) If we could know from moral responsibility that we are free, then we could know from moral responsibility (plus *a priori* reasoning) that indeterminism is true.

(2**) We cannot know from moral responsibility (plus *a priori* reasoning) that indeterminism is true.

(C**) Therefore, we cannot know from moral responsibility that we are free.⁴⁶⁹

(1**) and (2**), like their counterparts in the anti-introspection argument and the sceptical argument, are backed by two assumptions:

⁴⁶⁹ Here again, some adaptation (from "reasons to believe" to "knowledge") was necessary for the analogy with the other arguments to be clearer. My adaptation requires only to assume that "knowledge from reasons" entails "having good reasons to believe".

(a**) we are free => indeterminism is true⁴⁷⁰

(b**) the belief in moral responsibility isn't sensitive to indeterminism (only scientific investigations, if anything, could be sensitive to such a state of affairs)

Van Inwagen acknowledges that this argument has some force, but his “considered opinion” is that it fails. His reason for rejecting this argument is the *analogy* with the kind of sceptical argument we mentioned a few paragraphs above. It is important to emphasize here that van Inwagen doesn't consider the argument to be an *instance* of such sceptical arguments: he presents the two kinds of arguments as “very much alike”, and sufficiently similar to suggest “two types of response [...] that are exactly *parallel* to the two types of response available to the anti-sceptic” (van Inwagen 1983, 212, emphasis mine).

The first type of response is to refuse the closure principle on which (1*) and (1**) rely: maybe we can know that we have hands (by perception) and that “having hands” entails “not being a brain in a vat” (*a priori*), but that doesn't provide us with a way to come to know that we are not brains in vats. And similarly, maybe we can know that we are free (from morality), and that “being free” entails “being undetermined” (*a priori*), but that doesn't provide us with a way to come to know that we are undetermined.

The second type of response (the one favoured by van Inwagen) is, so to speak, to make the *modus ponens* instead of the sceptic's *modus tollens*: since we *can* know by perception that we have hands, maybe we can, after all, know that we are not brains in vats. And since we *can* know from our belief in morality that we are free, maybe we can, after all, come to know that we are undetermined from that same belief. (see van Inwagen 1983, 211–212)

Again, van Inwagen's point is *not* that the two arguments (the sceptical argument on the one hand, and the argument against knowledge from morality on the other), being of the same *kind*, necessarily stand or fall together. His point is rather that he fails to see how the disparities between the two cases could be relevant to justify a different treatment between the two.

I am not quite sure what to say about this defence by analogy. What seems to me quite clear, though, is that the anti-introspection argument presented by van Inwagen is, if anything, *more* analogous to the sceptical argument than is the argument against knowledge from morality. It is more analogous in two respects.

First, the alternative scenarios considered in the anti-introspection argument (Martian devices) and in the sceptical argument (brains in vats) are both philosophical inventions lacking any kind of positive support, scenarios just made up for the sake of philosophical argument. Whereas the alternative scenario in the argument against knowledge from morality is arguably a live option for which we might have some kind of inductive support from science. It should therefore be “easier” to justifiably rule out the first kind of alternative than it is to rule out the second kind.

Second, the kind of knowledge considered in the anti-introspection and the sceptical arguments (introspection on the one hand, perception on the other) are arguably *basic* (or non-inferential) kinds of knowledge, whereas knowing that we are free *from* (our belief in) morality is presumably inferential knowledge. And the defence of basic beliefs *as such* has traditionally played an important role in the refutation of scepticism.

⁴⁷⁰ This incompatibilist assumption may be either accepted wholeheartedly, or assumed for the sake of argument by van Inwagen's objector, according to whether she is a hard incompatibilist or a compatibilist. We could of course bypass the compatibility problem by setting the whole argument in terms of “libertarian freedom” or “being free in the libertarian sense”. But I share van Inwagen's doubts about such phrases (see van Inwagen 2008), and anyway, given the dialectical situation of this paper, it is clearer to keep with van Inwagen's setting of the problem.

In these two respects (which both seem potentially relevant to rebutting the sceptical argument), van Inwagen's anti-introspection argument is more analogous to the sceptical argument than is the argument against knowledge from morality.

Now, even though the logic of analogical reasoning is rather hard to ascertain, it seems reasonable to accept the following "*a fortiori* principle":

(AF) if it is legitimate to reason by analogy from A is F to B is F because of the analogy between A and B,
then if C is *more* analogous to A than B is, then it is legitimate to reason by analogy from A is F to C is F because of the (closer) analogy between A and C.⁴⁷¹

If we apply (AF) to our discussion, the result is that if it is legitimate to reason by analogy from "the sceptical argument must somehow be inconclusive" to "the anti-knowledge-from-morality argument must somehow be inconclusive" because of the analogy between these two arguments, then given that the anti-introspection argument is *more* analogous to the sceptical argument, then *a fortiori* it is legitimate to reason by analogy from "the sceptical argument must somehow be inconclusive" to "the anti-introspection argument must somehow be inconclusive". In short, *if* the analogy with the sceptical argument does save knowledge from morality *then* *a fortiori* does a closer analogy save introspective knowledge. And conversely: *if* the argument against introspective knowledge *does* work in spite of its analogy with the sceptical argument, then a more remote analogy cannot undermine the argument against knowledge from morality.

So if I am right up to that point, there are two conclusions one might draw from our discussion: one way to go would be to accept van Inwagen's argument against introspection and reject his defence of knowledge from morality. Another way to go would be to reject the anti-introspection argument and say that, for all this argument can show, it remains possible that we should know by introspection that we are free.

IV. Objection: the nature of Introspection

There is an important objection against the *a fortiori* line of reasoning presented in the previous section; this objection relies on the peculiar nature of introspection.

Isn't it a *grammatical error*, or a *logical falsehood*, to say that we could know *by introspection* about what Martians are doing? Since introspection is by definition knowledge *of our experiences* or internal states, it is also impossible *by definition* that we should know by introspection about something that is *not* an experience. Now, in the other two arguments, there may be some degree of initial *implausibility* to the fact that we should know that we are not brains in vats by perception, or know that we are undetermined from morality, but this is certainly not *impossible by definition*. So there seems to be an important reason why the analogical argument could save knowledge from morality while it couldn't possibly save introspective knowledge of free will.

The easiest way to understand this objection in formal terms is probably to view it as exploiting the inductive nature of analogical reasoning: it is well known that reasoning by

⁴⁷¹ This seems to be a natural specification of Bartha's very first principle of analogy: "(CS1) The more similarities (between the two domains), the stronger the analogy" (Bartha 2009, 19).

analogy isn't *valid* reasoning, but only reasoning giving *some inductive support* to the conclusion. So suppose we accept the conclusion that the analogy with scepticism gives the introspectionist stance *at least as much inductive defence* than it gives to the knowledge from morality stance. Still, whether this amount of inductive defence is enough to save the introspectionist stance depends on the rest of the evidence for and against that stance, in other words, it depends on its prior plausibility. So given the inductive nature of analogical reasoning it might be that the analogical defence of knowledge from morality is weak but successful (given a favourable prior plausibility), while the analogical defence of introspection is stronger but unsuccessful (given a much less favourable prior plausibility). In other words: the introspectionist stance might be, for other reasons, too implausible to be saved.

In our case, the other reason would be provided by a logical analysis of "introspective knowledge": this analysis is seen as giving a *null* prior plausibility to the notion of introspecting freedom (this notion being considered as a "logical falsehood"). But I shall argue that this logical analysis is misconceived. Or more precisely: this line of reply relies on an ambiguity in the notion of "knowledge by introspection", so that the analysis here given is appropriate for some uses of that phrase but not for the uses in the context of "knowledge of Free Will by Introspection". (The following discussion, being an analysis of "knowledge by introspection", will also have upshots much beyond the present section.)

"Knowledge by Introspection" may designate either of three kinds of knowledge in which experience plays a decisive role. Let us take, as an example, the experience as of hearing a roaring lion. From this experience, we may gain:

a) knowledge *of* experience: when I have the experience as of hearing a roaring lion, I can form the belief (and usually I thereby come to know) that *I have an experience as of a roaring lion*. The content of the belief (and the knowledge) in that case is not the *content* of the experience but, so to speak, the experience itself (or rather, its occurring).

b) knowledge *from* experience: when I have acquired a knowledge *of* experience, of the kind just mentioned, I can also draw some conclusions from this first belief. I might for example draw the conclusion that *some experiences have a rich content*. The content of the belief is now *inferred*... but it is not inferred from the *content* of the experience, it is inferred from the belief *that* this experience occurs.

c) knowledge *by* experience: of course, when I have an experience as of a roaring lion, I can also form the belief (and in favourable circumstances come to know) that *a lion is roaring*. That is: I can come to have a belief the content of which is the very same content as that of the experience.⁴⁷²

The third kind of knowledge (knowledge *by* experience) is what we typically call *perception*, while introspection in the narrow sense (the sense involved in the "logical falsehood reply") designates the first kind of knowledge (knowledge *of* experience). Now, if anyone were claiming that we can know about free will by introspection in this narrow sense, she would be asserting the rather weird view that "being free" requires no more than "feeling

⁴⁷² Of course, some epistemologists, most notably Descartes, have tried to reduce what I call "knowledge by experience" to a certain form of "knowledge from experience" – for instance, I could be said to acquire knowledge *of* my experience as of a roaring lion, and then *infer* from my having such an experience plus a belief in my creator's trustworthiness that (probably) there is a roaring lion. Whether or not such a reduction succeeds, my point here is only the logical point that knowledge by experience doesn't have the same kind of content as knowledge of experience. In the former case – and not in the latter – experience and belief share the same (object-oriented) content. The mere distinction between the two is all that is really needed here.

free” just because free will is that experience.⁴⁷³ Whatever the philosophical defensibility of such an extreme view, it surely cannot be a libertarian conception of Free Will.⁴⁷⁴

So, when libertarians like Reid or O’Connor assert that our experience of freedom can give us knowledge that we are free, they certainly don’t understand it as introspection in the narrow sense. It seems obvious, on the contrary, that what they have in mind is a knowledge the content of which is the very content of the experience that grounds it... in other words a kind of perception⁴⁷⁵. Since this kind of “perception” isn’t perception of states of affairs *external* to our body, but rather perception of some state *internal* to it, it was natural to give it a name with the prefix “intro-” (while the radical “spection” originally means no more than a certain kind of perception). Hence the broad sense of “introspection”: perception of some state internal to one’s body. Because of the narrow sense this word now has, it could be helpful to talk of “interoception”⁴⁷⁶ instead of introspection. I shall stick to that terminology for the rest of the paper.

Given the distinction we now have between introspection (knowledge of experience) and interoception (knowledge by experience), what should we think of the aforementioned disanalogy between the anti-introspection argument and the sceptical argument? Is it true that the anti-introspection argument rejects a knowledge claim which is a logical falsehood, while the sceptical argument rejects a knowledge claim which is only *prima facie* surprising? The answer is quite obvious: if the “anti-introspection” argument really is against “introspecting free will” in the narrow sense of introspection, then of course the argument is disanalogous with the sceptical argument, and is indeed conclusive. But its conclusion has no bearing whatsoever on the traditional “interoceptionist” stance.

If on the other hand the anti-introspection argument is read with the third meaning of introspection in mind (interoception), then the proposed disanalogy with the sceptical argument disappears: knowing by interoception that we are not fitted out with MD is by no means a *logical falsehood*, or a *grammatical mistake*. For sure, it might be a hard and *prima facie* implausible achievement, but that couldn’t count as a disanalogy with the situation of the sceptical argument, for it is also a hard and *prima facie* implausible achievement to know by perception that we are not brains in vats.

Therefore, the proper analysis of “introspection” in the case of Free Will is unable to render the anti-introspection argument less analogous to the sceptical argument than is the argument against knowledge from morality. To the contrary, it renders it, if anything, more analogous. So our previous conclusion still holds: if interoception cannot be saved by the analogy with scepticism, then *a fortiori* knowledge from morality cannot be saved by the

⁴⁷³ This kind of view was discussed by Galen Strawson, as an attempt to spell out our common phenomenology of freedom. Strawson interprets this phenomenology as an experience that presents itself as somehow logically sufficient to ascertain freedom: “Consider [...] *a*, an ordinary intelligent self-conscious purposive agent. [...] It is (it seems) quite clear to him that he is able to choose freely. But then surely he *is ipso facto* able to choose freely?” (Strawson 1986, 60) The crucial phrase here is “*ipso facto*”. Strawson develops this point with a theory of what he calls “experiential facts”: “What is experienced? Freedom. This really is experience of freedom. It is not just experience *as of* freedom. For freedom is what is experienced.” (Strawson 1986, 68). This is not the place to discuss such a view: I doubt that there is any wholehearted defender of it; and if there is, he doesn’t need any special argument to accept an introspectionist view on free will anyway.

⁴⁷⁴ Nor could it be any kind of traditional compatibilism. Indeed, all traditional conceptions of Free Will are, according to Strawson’s terminology (Strawson 1986, 12), “Objectivist theories” – that is: theories in which “*x* believes/experiences himself to be free” doesn’t play any constitutive role in defining “*x* is free”.

⁴⁷⁵ Indeed, O’Connor explicitly uses the word “perception” for the kind of knowledge he has in mind. See footnote 3.

⁴⁷⁶ The word was first introduced by (Sherrington 1906) in a technical physiological sense, and this specialized meaning has been refined a few times since Sherrington. My use of the word is of course quite independent of the physiological terminology.

more remote analogy with scepticism. Or in other words: if van Inwagen's anti-interoception strategy succeeds, his own defence of knowledge from morality fails. Therefore, van Inwagen's overall strategy in treating the Existence Question fails. This conclusion is not just an ad hominem argument since many philosophers of Free Will seem to have agreed with van Inwagen on the general idea that a libertarian epistemology based on morality was somehow better off than any epistemology having to do with our experience of free will.

V. The anti-interoception argument as an instance of the sceptical argument-type

The clarification of the notion of interoception has been brought to bear, so far, on the comparison between knowledge from morality and interoceptive knowledge of free will. And I have argued that this clarification shows the anti-interoception argument to be so close to the sceptical argument that it certainly is closer than the anti-knowledge-from-morality argument.

Now, I think we can push the point further, and bring it to bear on the more substantial question whether the anti-interoception argument, in and of itself, succeeds or fails. Indeed, the analysis of interoception as a particular kind of perception reveals much more than a close "analogy" between the anti-interoception argument and the closure-based sceptical argument—or rather should we say, the closure-based sceptical "argument-type"⁴⁷⁷, for closure-based arguments can be multiplied ad infinitum, changing only the particular sceptical scenario: demons, brains in vats, martians etc., or the particular source of knowledge: sight, touch, memory, etc. What our clarification shows is that the anti-interoception argument is really an *instance* of that type of sceptical arguments. And if this is true, then refuting (or refusing) scepticism (the sceptical type of argument) directly *implies* refuting (or refusing) the anti-interoception argument, without the weaknesses or uncertainties involved in analogical reasoning.

We shall see that the anti-interoception argument is just an instance of the closure-based sceptical argument-type if we focus on the common form of closure-based arguments, whether they involve demons, brains in vats, or indeed Martians, and whether they involve sight, touch, or memory. The essential apparatus of the closure-based sceptical argument is the following:

- (i) x has an experience E as of *p*, to the effect that x could prima facie be claimed to know that *p* on the ground of experience E.
- (ii) some sceptical hypothesis SH just made up for philosophical purposes (and that has no positive support) is such that (necessarily) if that scenario were the case, then
 - (a) *p* would be false (and x is able to apprehend *a priori* that $SH \Rightarrow \sim p$) and
 - (b) x would have just the same experience E.

With those two features, the sceptical argument proceeds as follows⁴⁷⁸:

⁴⁷⁷ Or, to follow (Pritchard 2005), the "template closure-based argument".

⁴⁷⁸ As discussed in footnote 10, the exact presentation of the sceptical argument type isn't the most classical one. This is due especially to premise (1°) which conjoins the Closure intuition on the one hand, and the "no antecedent knowledge of $\sim SH$ " intuition on the other hand. Correspondingly, there are two different ways of refusing (1°), by denying Closure, or by accepting antecedent knowledge of $\sim SH$. I discuss the possibility of antecedent knowledge – for both the sceptical argument and van Inwagen's argument – in the next paragraph. It may be important to remember here that this different grouping of the premises is only a matter of presentation, as long as the set of possible replies to the argument remains unchanged.

- (1°) if x could know by E that p , then x could know by E (plus *a priori* reasoning) that SH isn't the case.
- (2°) x couldn't know by E (plus *a priori* reasoning) that SH isn't the case.
- (C°) x couldn't know by E that p .

The brains-in-vats argument we presented earlier takes any of us as “ x ”, the whole set of our perceptual experiences as “ E ”, “we have hands” as “ p ” and “we are brains in vats” as “SH”.⁴⁷⁹

But obviously van Inwagen's anti-interoception argument is also of that form: we just have to substitute my experience of freedom for “ E ”, “I am free in performing action a ” for p , and the Martian device scenario for “SH”.

Maybe I have been too liberal in my formulation of the common form of closure-based arguments? Maybe I shouldn't have talked of “experience E ” in general, but only of “perceptual experiences” or “external perceptual experiences”? Well, if that were the case, we would have to say that closure-based arguments against memory rely on an essentially different intuition or argument form than the brains-in-vats arguments. But obviously they have just the same form, and this form is the one we just expressed. Just consider our argument type with the following parameters: let “ x ” be myself, let “ E ” be the set of my memorial experiences, let “ p ” be the proposition that I had coffee for breakfast this morning, and let “SH” be the hypothesis that a very powerful demon created me one second ago with just the set of memorial experiences which I happen to have. What we get is obviously a new instance of just the same sceptical argument. What is essential to this kind of closure-based sceptical arguments is that they threaten what I've called “knowledge by experience” (knowledge the content of which is the very content of the experience that grounds it).⁴⁸⁰ And we've seen that interoception was just an instance of that kind of knowledge.

I now come to a problem already alluded to in footnote.⁴⁸¹ In the formulation here given, the sceptical argument may seem to rely on the problematic contention that not only knowledge in general, but also particular knowledge modes are closed under (known) entailment. That is, it may seem to imply that whenever I know something *from source E* , then I know (or can come to know) the consequences *from source E as well*. This would be unfortunate, since the non-closure of knowledge modes is widely accepted.⁴⁸² So this might be sufficient to discredit both the argument type just mentioned, and van Inwagen's instance of that type. (Remember here that van Inwagen explicitly says that “if it were right [that we can find out whether we have free will by introspection], we could find out *by introspection* [sic] whether we were fitted out with Martian devices.” (van Inwagen 1983, 204), emphasis mine.) But it seems to me that this problem with van Inwagen's formulation is at the end of the day no more than a minor distraction.

A first thing to notice here is the difference between “knowing from source E ” and “knowing from source E plus *a priori* reasoning”. If anyone said that the former is closed under known entailment, this would indeed amount to a closure of the knowledge mode corresponding to source E . But saying that the latter is closed is really a claim about competent *a priori* deduction, and not about the particular knowledge mode corresponding to

⁴⁷⁹ More precisely, in order to satisfy the conditional (ii)-(b), SH must read: “we are brains in vats manipulated to have exactly the perceptual experiences E ”.

⁴⁸⁰ Namely, closure-based arguments which involve sceptical scenarios. This not the case for all closure-based arguments; for instance, “lottery arguments” do not involve sceptical scenarios (see Hawthorne 2004).

⁴⁸¹ See footnotes 10 and 20.

⁴⁸² Indeed, it is sufficiently uncontroversial for Dretske to take it as a premise in an argument against general closure. See (Luper 2005) on that point.

source E. Since van Inwagen obviously had in mind the deductive step from “we are free” to “we are not fitted out with MD”, it is clear that he never meant to assume anything like closure of knowledge modes, but only closure of Competent Deduction.

Now, a second remark is needed because it is still true that the most plausible versions of Competent Deduction Closure (including principle (CD-Closure) above) are unable to deliver *directly* premise (1°) and its instances. In order to get (1°) directly, a stronger principle would be needed:

(Strong CD-Closure) If x knows that p , and x is in a position to competently deduce q from p (due to *a priori* considerations), then x is in a position to know that q in this way.

Unfortunately, (Strong CD-Closure) doesn't seem to be true in general. There are cases in which I know a proposition p , and I can entertain the deduction from p to some entailed q , but I already know q and couldn't have come to know that p in the first place if I hadn't antecedently known q . In such cases, the deduction from p to q couldn't constitute a new way for me to know that q .⁴⁸³

Because of such cases (of essentially antecedent knowledge of the consequent), the most defensible versions of Competent Deduction Closure provide us with deductive knowledge *only* for those consequences (of our known beliefs) which we don't antecedently know.⁴⁸⁴ This qualification is of great importance for sceptical arguments because one possible response to scepticism is precisely to appeal to some antecedent (non perceptual, presumably *a priori*) knowledge of anti-sceptical propositions. If we do have such antecedent knowledge of $\sim$ SH, then (CD-Closure) trivially satisfied (by falsity of the antecedent), and doesn't deliver the result that everyday perceptual knowledge would set us in a position to acquire further deductive warrant of $\sim$ SH. In other words, it doesn't support premise (1°):

(1°) if x could know by E that p , then x could know by E (plus *a priori* reasoning) that SH isn't the case.

Therefore, if we have antecedent knowledge of $\sim$ SH, premise (1°) may be false without falsifying Competent Deduction Closure. This shows that premise (1°) really bears in itself two different assumptions essential to the sceptical reasoning: one is the epistemic closure intuition, and the other one is the impossibility of having antecedent knowledge of $\sim$ SH. For that reason, there are two different ways to resist premise (1°)... and therefore three ways to resist the closure-based argument type:

- a) resist premise (2°), *i.e.* accept that deduction from everyday perceptions is an acceptable way to come to know the negation of sceptical scenarios.
- b) resist premise (1°) by accepting antecedent knowledge of anti-sceptical propositions.

⁴⁸³ Suppose a reliable friend tells me that Mr John X is truthful and highly reliable regarding his family (the Xs), and I come to know this proposition (q) upon my friend's testimony. Suppose, furthermore, I hear Mr John X saying that *all* members of his family are truthful and highly reliable regarding the Xs, and I come to know this new proposition (p) upon Mr John X's testimony, which I know to be trustworthy because of my knowledge that q . Now, it is obvious that p entails q and I could easily bring my attention to the deduction of q from p . But it would be really odd to say that I would thereby acquire a new way of knowing that q – a way relying ultimately on my knowledge that p (plus *a priori* reasoning) – for I wouldn't know that p in the first place if I didn't antecedently know that q .

⁴⁸⁴ In the principle (CD-Closure) of this paper, this qualification is expressed by the condition that the knower x must be “in a position to *acquire* the belief that q ”. If x has antecedent knowledge of q , he is not in a position to acquire new belief (let alone new knowledge) that q , and therefore the principle is trivially satisfied for x doesn't satisfy the antecedent of the conditional.

c) resist premise (1°) by denying closure. This means accepting that we can know everyday perceptual propositions even though there is no way we can know some of its obvious logical consequences.

What solution is the most plausible? It is not the purpose of this paper to answer this question. What is important for us is only this: whatever solution is adopted for the sceptical argument-type, an instance of that solution will immediately follow to rebut van Inwagen's anti-interoception argument. If easy deductions like "this is a hand, therefore I am not a brain in a vat" are epistemically legitimate, then so are deductions like "I am free, therefore I am not fitted out with MD". If we have some antecedent, *a priori*, way to rule out brains-in-vats scenarios and demons scenarios, then surely we can also rule out Martian device scenarios. If Competent Deduction Closure fails, then there is no particular problem in "interocepting" Free Will even though we can't know that we are not fitted out with MD.

Though I (like van Inwagen) favour the answers which preserve epistemic closure, my general conclusion is only that anyone who rejects the sceptical argument-type (for whatever reason) should also reject (for just the same reason) van Inwagen's anti-introspection argument because it is just an instance of that type.

VI. Conclusions and Upshots

We have now reached the two intended conclusions. First, van Inwagen's anti-interoception argument is inconsistent with his overall treatment of the Existence Question, and more precisely with his defence of the knowledge-from-morality stance. For all his argument can show, the interoceptionist stance is no worse off than the knowledge-from-morality stance (it is rather, if any different, better off!). Some readers may want to read this conclusion in its negative form, namely that the knowledge-from-morality stance is "yet worse off" than the (already very badly off) interoceptionist stance. In any case, this goes against an assumption common to various philosophers of Free Will that *if* there is any hope for a libertarian stance, it could only be found in indirect strategies like the knowledge-from-morality stance.

Second, I argued that van Inwagen's argument is inconclusive for anyone not prepared to endorse general scepticism about perception. So not only does his argument fail to place the interoceptionist stance in a worse epistemic position than the knowledge-from-morality stance; it doesn't even place interoception in a worse epistemic position than everyday perceptual knowledge.

Some readers may find that I haven't managed to establish those conclusions. But in this last section, I want to address another kind of reaction, which we could express as follows.

Your discussion does succeed in showing that van Inwagen's argument, in its unfortunate historical form (with the sceptical MD scenario), doesn't pose a real threat to the libertarian interoceptionist stance. But a charitable reader should go beyond the arguments as explicitly stated and address their best reconstructions. And there is indeed a minimal revision of van Inwagen's argument that remains unscathed by anything you've said. Namely:

- (1r) If we could know by interoception that we are free, then we could know by interoception (plus *a priori* reasoning) that indeterminism is true.
 (2r) We cannot know by interoception (plus *a priori* reasoning) that indeterminism is true.
 (C) We cannot know by interoception that we are free.

Since *this* argument is obviously successful against libertarian interoceptionism, all you have proved at the end of the day is that van Inwagen was a bit careless in knocking down a position which almost nobody wants to hold anyway.

A first thing to notice is that this reaction changes nothing to the first conclusion we drew, *i.e.* the comparative conclusion that interoceptive knowledge isn't worse off than knowledge-from-morality. Indeed the "revised argument" here proposed renders the two arguments (anti-interoception and anti-knowledge-from-morality) yet closer, and therefore yet more likely to stand or fall together in an analogical reasoning with the sceptical argument. If this were van Inwagen's original intuition, it would devastate yet more clearly his own knowledge-from-morality stance.

For the same reason, this revised argument cannot be plausibly presented as a "charitable reconstruction" of what van Inwagen had in mind. Notice that what gives the revised argument a special force is that the threatening alternative in it isn't a pure invention made up for philosophical purposes and lacking any kind of positive support; it isn't a good old "sceptical scenario" like the MD scenario. And therefore, it is not clear that this argument can be considered as just one more instance of the sceptical argument type.⁴⁸⁵ But, van Inwagen shouldn't be happy to concede that the parameter of disanalogy with sceptical arguments which renders interoceptive knowledge unrescuable is precisely the presence of a non-farfetched alternative, for this very same parameter would yet more obviously undermine his knowledge-from-morality stance.⁴⁸⁶ For that reason, we should treat separately the anti-interoception argument with sceptical alternative (which is van Inwagen's) and the anti-interoception argument with Determinism alternative. Those are two completely different kinds of alternatives and generate completely different kinds of replies.

So, for the purposes of this paper, my two main reactions to the "revised argument" are that 1) it is irrelevant to the *comparative* conclusion that introspection is no worse off than knowledge-from-morality or other indirect strategies, and 2) though it may constitute a more

⁴⁸⁵ For sure, it keeps a strong *analogy* with sceptical arguments – this analogy is indeed what van Inwagen uses to defend knowledge-from-morality from the same threat of Determinism. But again, mere analogy may not be enough, and it may very well be that some plausible reply to the sceptical argument type is impossible or implausible for this argument.

⁴⁸⁶ There is a subtler and more important reason to think that the choice of the MD alternative (as opposed to a Determinism alternative), far from being an effect of carelessness, has an essential role in van Inwagen's strategy. Here is why. When the MD scenario first appears in van Inwagen's book (*i.e.* in the discussion of the Paradigm Case Argument (see van Inwagen 1983, chap. 4.2)), it isn't chosen first and foremost as a sceptical scenario, but rather as a scenario that precludes free will under any defensible construal of the term, *be it compatibilist or incompatibilist*. This feature makes it a good test case to eliminate theories or arguments that are unacceptable even by compatibilist lights. It seems to me that the same kind of dialectical intention is at play here: van Inwagen probably intended to dismiss introspection altogether, for compatibilists and libertarians alike. For if he succeeds in doing so, then the compatibilist is bound to follow him in accepting an indirect epistemology of free will, like knowledge-from-morality. Now, on such common ground, it is not clear at all that the libertarian has a harder epistemological time than the compatibilist, because indirect epistemologies are admittedly fairly mysterious for anyone. But if the compatibilist is still in a position to claim an introspective epistemology for *his* kind of free will, then maybe he will be in a less mysterious position (after all, could we not know by introspection, for instance, that our first order desires fit our second order desires, or that they fit our deep self?) Therefore, dismissing introspectionism *for everyone* sets van Inwagen in a far better position in the epistemological discussion with compatibilists. And only the MD scenario allows him to do that.

problematic threat to libertarianism than van Inwagen's argument, it cannot be considered a mere revision of van Inwagen's intuition, and therefore, it is also irrelevant to the positive conclusion that van Inwagen's argument raises no worry for the libertarian introspectionist stance.

Now, of course, if some other argument than van Inwagen's obviously defeated introspectionism, there would be little interest in responding to van Inwagen's argument in the first place. So, even though the proper treatment of the revised argument requires a proper and separate treatment, it may help to give here at least a first indication that the task of responding to it isn't totally hopeless. Fortunately, I think that the preceding discussion gives us some resources to give such an indication. We can see a first insight of a possible solution if we pay attention to the fundamental difference between van Inwagen's argument and the revised argument. As we've seen, the fundamental difference lies in the nature of scenarios. The most important aspect of sceptical alternatives is that our experiences are *in principle* unable to reliably track their being the case. Fortunately, the correlative weakness is that alternatives with such a feature seldom come up as live options, and sceptics have to *make them up*.⁴⁸⁷ So the "strength" of those alternatives is that, very intuitively, we cannot properly rule them out... but their weakness is that, being totally made up scenarios, lacking any kind of positive support, it is not so clear that they *have to* be ruled out, at least in any strong sense of the term. The Determinism scenario has exactly the reverse strength... and the reverse weakness! It is a live option with (at least some minimal) positive support, and this gives strength to the idea that a libertarian would have to rule out this alternative in some way or other... but it also has the reverse defect of not being "ineliminable in principle", or non-trackable by mere stipulation, because the nature of the "scenario" here is settled by our evidence from the external world, not stipulated by our dialectical needs as is the case for sceptical scenarios. For that very reason, the question whether our experiences could or couldn't track Determinism has to be *argued for*, not stipulated. This, I think, suggests a general strategy for how to tackle the new argument: one can put into doubt the assumption that our interoceptive experiences couldn't reliably track determinism. If we cannot exclude the possibility of such a tracking relation, why should (2r) be considered obviously true? There seems to be at least some room for discussion here.

Of course, I do not take this hint, as it stands, to be a full answer to the new argument. A full answer would need a paper of its own. My intention is only to suggest that rebutting van Inwagen's sceptical argument against interoception isn't just a local victory in a hopeless war; it is rather a first battle to clear the way for others.

To sum up the main upshots of this paper: I have tried to convey the general idea that there is no good reason for libertarians to despair of the traditional interoceptive epistemology and switch to an indirect one (like knowledge from morality); first because such strategies are not better off than interoception; and second, because it is far from clear that interoception faces decisive challenges. One of the rare explicit arguments actually proposed by an influential philosopher – that of van Inwagen – is actually a failure, and discussion of this argument also gives a hint of how to approach further issues.

⁴⁸⁷ See, for instance, the way in which Pritchard defines sceptical alternatives: "Since it is clearly the case that we are able to rule out some error-possibilities, in order to meet this challenge the sceptic therefore needs to further argue that there are error-possibilities within the class of relevant error-possibilities that are ineliminable—that is, error-possibilities which agents are unable to rule out. To this end, the sceptic typically adduces a special kind of error-possibility—viz. sceptical hypotheses. Crucially, this type of error-possibility has the special feature that, intuitively, no agent could ever rule it out." (Pritchard 2005, 24).

Bibliographie

- Adler, Mortimer. 1958. *The Idea of Freedom : A Dialectical Examination of the Conceptions of Freedom*. 1st ed. Garden City N.Y.: Doubleday.
- Albert, Hans. 1975. *Traktat über kritische Vernunft*. Tübingen: Mohr.
- Almeida, Claudio. 2001. « What Moore's Paradox Is about ». *Philosophy and Phenomenological Research* 62 (1): 33–58.
- Alston, William P. 1976. « Two Types of Foundationalism ». *The Journal of Philosophy* 73 (7) (April 8): 165.
- . 1986. « Epistemic Circularity ». *Philosophy and Phenomenological Research* 47 (1) (September): 1.
- . 1993. *The Reliability of Sense Perception*. Cornell University Press.
- . 2003. « Religious Language and Verificationism ». In *The Rationality of Theism*, edited by Paul Moser and Paul Copan, 26–34. New York: Routledge.
- . 2005. *Beyond "Justification" : Dimensions of Epistemic Evaluation*. Ithaca: Cornell University Press.
- Anscombe, G. E. M. 1957. *Intention*. Oxford: Blackwell.
- . 1971. *Causality and Determination: An Inaugural Lecture*. New York: Cambridge University Press.
- Aristote. 1959. *Éthique à Nicomaque*. Translated by Jules Tricot. Paris: Vrin.
- Armstrong, David. 1962. *Bodily Sensations*. London: Routledge & K. Paul.
- . 1968. *A Materialist Theory of the Mind*. London u.a.: Routledge & Kegan Paul.
- . 1973. *Belief, Truth and Knowledge*. London: Cambridge University Press.
- . 1989. *A Combinatorial Theory of Possibility*. Vol. 41. 164. Cambridge University Press.
- . 1999. *A World of States of Affairs*. Cambridge [etc.]: Cambridge University Press.
- Audi, Robert. 1993. « Modalities of Knowledge and Freedom ». In *Action, Intention, and Reason*, 253–280. Ithaca, N.Y.: Cornell University Press.
- . 1998. *Epistemology : A Contemporary Introduction to the Theory of Knowledge*. London; New York: Routledge.
- Ayer, A.J. 1954. « Freedom and Necessity ». In *Philosophical Essays*, 3–20. New York: St. Martin's Press.
- Balaguer, Mark. 2010. *Free Will as an Open Scientific Problem*. Cambridge: MIT Press.
- Banks, W. 2006. « Does Consciousness Cause Misbehavior ». In *Does Consciousness Cause Behavior*, edited by Susan Pockett, William P Banks, and Shaun Gallagher, 235–256. Cambridge, Mass.: MIT Press.
- Barberousse, Anouk, Max Kistler, and Pascal Ludwig. 2000. *La philosophie des sciences au XXe siècle*. [Paris]: Flammarion.
- Baumeister, Roy F., E. J. Masicampo, and C. Nathan DeWall. 2009. « Prosocial Benefits of Feeling Free: Disbelief in Free Will Increases Aggression and Reduces Helpfulness ». *Personality and Social Psychology Bulletin* 35 (2): 260–268.
- Bayne, Tim. 2006. « Phenomenology and the Feeling of Doing: Wegner on the Conscious Will ». In *Does Consciousness Cause Behavior*, edited by Susan Pockett, William P Banks, and Shaun Gallagher, 169–186. Cambridge, Mass.: MIT Press.
- . 2011a. « The Sense of Agency ». In *The Senses*, edited by Fiona Macpherson, Oxford University Press. New York.

- . 2011b. « Libet and the Case for Free Will Scepticism ». In *Free Will and Modern Science*, edited by Richard Swinburne, 25–46. Oxford: Oxford University Press.
- Bayne, Tim, and Neil Levy. 2006. « The Feeling of Doing: Deconstructing the Phenomenology of Agency ». *Disorders of Volition*: 49–68.
- Bealer, George. 1996. « A Priori Knowledge and the Scope of Philosophy ». *Philosophical Studies* 81 (2): 121–142.
- . 1999. « A Theory Of The A Priori ». *Nous* 33 (s13) (October): 29–55.
- . 2000. « A Theory of the A Priori ». *Pacific Philosophical Quarterly* 81 (1): 1–30.
- Bealer, George, and P. F. Strawson. 1992. « The Incoherence of Empiricism ». *Proceedings of the Aristotelian Society, Supplementary Volumes*: 99–143.
- Beebe, Helen. 2003. « Seeing Causing ». In *Proceedings of the Aristotelian Society (Hardback)*, 103:257–280. Wiley Online Library.
- . 2009. « Causation and Observation ». In *The Oxford Handbook of Causation*, edited by Helen Beebe, Christopher Hitchcock, and Peter Charles Menzies, 471–497. Oxford; New York: Oxford University Press.
- Benacerraf, Paul. 1965. « What Numbers Could Not Be ». *The Philosophical Review* 74 (1): 47.
- Bennett, Jonathan. 2003. *A Philosophical Guide to Conditionals*. Oxford; New York: Clarendon Press ; Oxford University Press.
- Bergmann, Michael. 2004. « Epistemic Circularity: Malignant and Benign ». *Philosophy and Phenomenological Research* 69 (3): 709–727.
- . 2006a. *Justification without Awareness : A Defense of Epistemic Externalism*. Oxford; New York: Clarendon Press ; Oxford University Press.
- . 2006b. « Epistemic Circularity and Common Sense: A Reply to Reed ». *Philosophy and Phenomenological Research* 73 (1): 198–207.
- Berofsky, Bernard. 2002. « Ifs, Cans, and Free Will: The Issues ». In *The Oxford Handbook of Free Will*, edited by Robert Kane, 181–201. New York: Oxford University Press.
- BonJour, Laurence. 1978. « Can Empirical Knowledge Have a Foundation? ». *American Philosophical Quarterly*: 1–13.
- . 1985. *The Structure of Empirical Knowledge*. Cambridge, Mass.: Harvard University Press.
- . 2001. « Toward a Defense of Empirical Foundationalism ». In *Resurrecting Old-Fashioned Empiricism*, edited by Michael DePaul, 21–38. New York: Rowman & Littlefield.
- BonJour, Laurence, and Ernest Sosa. 2003. *Epistemic Justification : Internalism vs. Externalism, Foundations vs. Virtues*. Oxford: Blackwell.
- Boswell, James. 1946. *The Life of Samuel Johnson*,. Edited by C. P Chadsey and Gordon Ross. Garden City, N.Y.: Doubleday and Co.
- Boulter, Stephen. 2007. *The Rediscovery of Common Sense Philosophy*. Basingstoke: Palgrave Macmillan.
- Bourget, David, and David J. Chalmers. 2013. « What Do Philosophers Believe? ». *Philosophical Studies*: 1–36.
- Brass, Marcel, and Patrick Haggard. 2007. « To Do or Not to Do: The Neural Signature of Self-Control ». *The Journal of Neuroscience* 27 (34): 9141–9145.
- Browning, D. 1964. « The Feeling of Freedom ». *The Review of Metaphysics*: 123–146.
- Buchak, Lara. 2013. « Free Acts and Chance: Why The Rollback Argument Fails ». *The Philosophical Quarterly* 63 (250): 20–28.
- Campbell, Charles Arthur. 1931. *Scepticism and Construction : Bradley's Sceptical Principle as the Basis of Constructive Philosophy*. London: G. Allen and Unwin.
- . 1951. « Is Freewill a Pseudo-Problem ». *Mind* 60 (240): 441–465.

- . 1957. *On Selfhood and Godhood: The Gifford Lectures Delivered at the University of St. Andrews during Sessions 1953-54 and 1954-55, Rev. and Expanded*. London: Allen & Unwin.
- Cartwright, N. 1979. « Causal Laws and Effective Strategies ». *Nous* 13 (4): 419–437.
- Cartwright, Nancy. 1993. « In Defence of 'This Worldly' Causality: Comments on van Fraassen's Laws and Symmetry ». *Philosophy and Phenomenological Research* 53 (2) (June): 423.
- Caruso, Gregg. 2010. « Consciousness and Free Will: A Critique of the Argument from Introspection ». *Southwest Philosophy Review* 24 (1) (July 21): 219–231.
- Chalmers, David John, David Manley, and Ryan Wasserman. 2009. *Metametaphysics: New Essays on the Foundations of Ontology*. Oxford; New York: Clarendon Press.
- Chisholm, Roderick M. 1964. « Human Freedom and the Self ». In *Lindley Lecture*. Department of Philosophy, University of Kansas.
- . 1976. « The Agent as Cause ». In *Action Theory*, edited by Douglas N Walton and Myles Brand, 199–211. D. Reidel Pub. Co.
- . 1982. « The Problem of the Criterion ». In *The Foundations of Knowing*, 61–75. Minneapolis: University of Minnesota Press.
- . 1989. *Theory of Knowledge*. Englewood Cliffs, N.J.: Prentice-Hall.
- Clarke, Randolph. 2000. « Modest Libertarianism ». *Noûs* 34 (s14): 21–45.
- . 2003. *Libertarian Accounts of Free Will*. Oxford; New York: Oxford University Press.
- Van Cleve, Jim. 2003. « Is Knowledge Easy-or Impossible? Extemalism as the Only Alternative to Skepticism ». In *The Skeptics: Contemporary Essays*, edited by Steven Luper, Ashgate, 45–59. Aldershot, UK.
- Cohen, Stewart. 2002. « Basic Knowledge and the Problem of Easy Knowledge ». *Philosophy and Phenomenological Research* 65 (2): 309–329.
- Coliva, Annalisa. 2008. « The Paradox of Moore's Proof of an External World ». *The Philosophical Quarterly* 58 (231): 234–243.
- . 2010. *Moore and Wittgenstein: Scepticism, Certainty, and Common Sense*. Houndmills Basingstoke Hampshire ;New York NY: Palgrave Macmillan.
- Conee, Earl Brink, and Richard Feldman. 2004. *Evidentialism: Essays in Epistemology*. Oxford; New York: Clarendon Press ; Oxford University Press.
- Dancy, Jonathan. 1986. *An Introduction to Contemporary Epistemology*. Oxford UK ;New York NY USA: B. Blackwell.
- Dancy, Jonathan, Ernest Sosa, and Matthias Steup, ed. 2010. *A Companion to Epistemology*. Chichester, West Sussex, U.K.; Malden, MA: Wiley-Blackwell.
- Davidson, Donald. 1973. « Freedom to Act ». In *Essays on Freedom of Action*, edited by Ted Honderich, 67–86. London: Routledge & Kegan Paul.
- . 1986. « A Coherence Theory of Knowledge and Truth ». In *Truth and Interpretation*, edited by Ernest LePore, 307–319. Oxford: Blackwell.
- Dayan, Peter, and L. F Abbott. 2001. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, Mass.: Massachusetts Institute of Technology Press.
- Deery, Oisin, Matt Bedke, and Shaun Nichols. 2013. « Phenomenal Abilities: Incompatibilism and the Experience of Agency ». In , 126–150. OUP.
- Dennett, Daniel C. 1978. « On Giving Libertarians What They Say They Want ».
- . 1984a. *Elbow room: the varieties of free will worth wanting*. Cambridge, Mass.: MIT Press.
- . 1984b. « I Could Not Have Done Otherwise--So What? ». *The Journal of Philosophy* 81 (10) (October 1): 553–565.

- . 2004. *Freedom Evolves*. Penguin UK.
- DeRose, Keith. 1991. « Epistemic Possibilities ». *The Philosophical Review* 100 (4): 581–605.
- . 1992. « Contextualism and Knowledge Attributions ». *Philosophy and Phenomenological Research* 52 (4): 913–929.
- Descartes, René. 1953. *Oeuvres et lettres*. Paris: Gallimard.
- . 1964. *OEuvres de Descartes*. Edited by Charles Adam and Paul Tannery. Paris: J. Vrin.
- Donagan, Alan. 1987. *Choice, the Essential Element in Human Action*. London ; New York: Routledge & Kegan Paul.
- Double, Richard. 1996. *Metaphilosophy and Free Will*. New York: Oxford University Press.
- Dretske, Fred. 1970. « Epistemic Operators ». *The Journal of Philosophy*: 1007–1023.
- . 1971. « Conclusive Reasons ». *Australasian Journal of Philosophy* 49 (1): 1–22.
- . 2005. « The Case against Closure ». In *Contemporary Readings in Epistemology*, edited by Matthias Steup and Ernest Sosa, 13–25. Malden, MA: Blackwell Pub.
- Ducasse, Curt J. 1926. « On the Nature and the Observability of the Causal Relation ». *The Journal of Philosophy* 23 (3): 57–68.
- Duns Scotus, John. 1987. *Philosophical writings: a selection*. Translated by Allan Bernard Wolter. Indianapolis: Hackett Pub. Co.
- Durkheim, Émile. 2002. *Cours de Philosophie Fait Au Lycée de Sens En 1883-1884 Notes Prises Entre 1883-84 Par Le Philosophie Français André Lalande*. Chicoutimi : J.-M. Tremblay,.
- Ekstrom, Laura. 2002. « Libertarianism and Frankfurt-Style Cases ». In *The Oxford Handbook of Free Will*, edited by Robert Kane, 309–322. New York: Oxford University Press.
- Engel, Pascal. 2007. *Va Savoir! : De La Connaissance En Général*. Paris: Hermann.
- Fales, Evan. 1990. *Causation and Universals*. London; New York: Routledge.
- Feldman, Richard. 1999. « Contextualism and Skepticism ». *Noûs* 33: 91–114.
- Feldman, Richard, and Earl Conee. 1985. « Evidentialism ». *Philosophical Studies* 48 (1) (July): 15–34.
- . 2001. « Internalism Defended ». *American Philosophical Quarterly* 38 (1): 1–18.
- Firth, Roderick. 1978. « Are Epistemic Concepts Reducible to Ethical Concepts? ». In *Values and Morals: Essays in Honor of William Frankena, Charles Stevenson, and Richard Brandt*, edited by Jaegwon Kim and Alvin Goldman, 215–229. Dordrecht: Kluwer.
- Fischer, John Martin. 1994. *The Metaphysics of Free Will : An Essay on Control*. Cambridge, Mass.: Blackwell.
- . 1999. « Recent Work on Moral Responsibility ». *Ethics* 110 (1) (October 1): 93–139.
- . 2000. « Problems with Actual-Sequence Incompatibilism (Comments on Robert Kane's Presentation) ». *The Journal of Ethics* 4 (4): 323–328.
- . 2006. *My Way Essays on Moral Responsibility*. New York: Oxford University Press.
- Fischer, John Martin, Robert Kane, Derk Pereboom, and Manuel Vargas. 2007. *Four Views on Free Will*. Great Debates in Philosophy. Malden, MA: Blackwell.
- Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge; New York: Cambridge University Press.
- Flint, Thomas P. 1998. *Divine Providence : The Molinist Account*. Ithaca, N.Y.: Cornell University Press.
- Foley, Richard. 1987. *The Theory of Epistemic Rationality*. Cambridge, Mass.: Harvard University Press.
- . 2005. « Justified Belief as Responsible Belief ». In *Contemporary Debates in Epistemology*, edited by Matthias Steup and Ernest Sosa, 313–325.

- Frankfurt, Harry G. 1969. « Alternate Possibilities and Moral Responsibility ». *The Journal of Philosophy* 66 (23): 829–839.
- . 1971. « Freedom of the Will and the Concept of a Person ». *The Journal of Philosophy* 68 (1) (January 14): 5.
- Franklin, Christopher Evan. 2011. « Farewell to the Luck (and Mind) Argument ». *Philosophical Studies* 156 (2): 199–230.
- Fumerton, Richard. 1985. *Metaphysical and Epistemological Problems of Perception*. Lincoln: University of Nebraska Press.
- . 1995. *Metaepistemology and Skepticism*. Lanham, Md.: Rowman & Littlefield.
- Gettier, Edmund L. 1963. « Is Justified True Belief Knowledge? ». *Analysis* 23 (6) (June): 121.
- Ginet, Carl. 1966. « Might We Have No Choice ». In *Freedom and Determinism*, edited by Keith Lehrer, 87–104. New York: Random House.
- Globus, Gordon. 1995. « ‘Kane on Incompatibilism: An Exercise in Neurophilosophy’ ». *Unpublished Paper*.
- Goldman, Alvin. 1967. « A Causal Theory of Knowing ». *The Journal of Philosophy* 64 (12) (June 22): 357.
- . 1976. « Discrimination and Perceptual Knowledge ». *The Journal of Philosophy* 73 (20) (November 18): 771.
- . 1979. « “What Is Justified Belief?” ». In *Justification and Knowledge*, edited by George Pappas, 1–25. Boston: D. Reidel.
- . 1986. *Epistemology and Cognition*. Cambridge Mass.: Harvard University Press.
- . 1993. « The Psychology of Folk Psychology ». *Behavioral and Brain Sciences* 16 (01): 15–28.
- Le Gousse, Ghislain. 2011. « Liberté, Responsabilité Morale et Déterminisme Causal, Enquête Sur Le Pouvoir D’agir Autrement ». Nantes.
- Grice, H. Paul. 1975. « Logic and Conversation ». In *Studies in the Way of Words*, 22–40. Cambridge, Mass.: Harvard University Press.
- Grünbaum, Adolf. 1952. « Causality and the Science of Human Behavior ». *American Scientist* 40 (4): 665–689.
- Guillon, Jean-Baptiste. 2014. « Van Inwagen on Introspected Freedom ». *Philosophical Studies* 168 (3) (April 1): 645–663.
- Haggard, Patrick. 2005. « Conscious Intention and Motor Cognition ». *Trends in Cognitive Sciences* 9 (6): 290–295.
- Haggard, Patrick, and Martin Eimer. 1999. « On the Relation between Brain Potentials and the Awareness of Voluntary Movements ». *Experimental Brain Research* 126 (1): 128–133.
- Haji, Ishtiyaque. 2001. « Control Conundrums: Modest Libertarianism, Responsibility, and Explanation ». *Pacific Philosophical Quarterly* 82 (2): 178–200.
- Hameroff, Stuart R., and Roger Penrose. 1996. « Conscious Events as Orchestrated Space-Time Selections ». *Journal of Consciousness Studies* 3 (1): 36–53.
- Harré, Rom, and Edward H Madden. 1975. *Causal Powers: A Theory of Natural Necessity*. Totowa, N.J.: Rowman and Littlefield.
- Hasker, William. 1989. *God, Time, and Knowledge*. Ithaca: Cornell University Press.
- Hawthorne, John. 2004. *Knowledge and Lotteries*. Oxford; New York: Clarendon ; Oxford University Press.
- . 2005. « The Case for Closure ». In *Contemporary Readings in Epistemology*, edited by Matthias Steup and Ernest Sosa, 26–43. Malden, MA: Blackwell Pub.
- Hobart, Richard E. 1934. « Free Will as Involving Determination and Inconceivable without It ». *Mind* 43 (169): 1–27.

- Hobbes, Thomas. 1998. *Leviathan*. Edited by John Charles Addison) Gaskin. Oxford: Oxford University Press.
- Hodgson, David. 2002. « Quantum Physics, Consciousness, and Free Will ». In *The Oxford Handbook of Freewill*.
- Holton, Richard. 2006. « The Act of Choice ». *Philosophers' Imprint* 6 (3): 1–15.
- Honderich, Ted. 1988. *A Theory of Determinism: The Mind, Neuroscience, and Life-Hopes*. Oxford [England]; New York: Clarendon Press ; Oxford University Press.
- Horgan, Terry. 1979. « “could”, Possible Worlds, and Moral Responsibility ». *The Southern Journal of Philosophy* 17 (3): 345–358.
- . 2011a. « The Phenomenology of Agency and Freedom: Lessons from Introspection and Lessons from Its Limits ». *Humana Mente* 15: 77–97.
- . 2011b. « Causal Compatibilism about Agentive Phenomenology ». In *A Festschrift for Jaegwon Kim*, edited by Terry Horgan, M. Sabates, and D. Sosa. Cambridge, Mass.: MIT Press.
- Horgan, Terry, John Tienson, and George Graham. 2003. « The Phenomenology of First-Person Agency ». In *Physicalism and Mental Causation: The Metaphysics of Mind and Action*, 323.
- Horgan, Terry, and Mark Timmons. 2011. « Introspection and the Phenomenology of Free Will: Problems and Prospects ». *Journal of Consciousness Studies* 18 (1): 180–205.
- Huemer, Michael. 2007. « Compassionate Phenomenal Conservatism ». *Philosophy and Phenomenological Research* 74 (1): 30–55.
- . 2010. « Foundations and Coherence ». In *A Companion to Epistemology*, edited by Jonathan Dancy, Ernest Sosa, and Matthias Steup, 22–33. Chichester, West Sussex, U.K.; Malden, MA: Wiley-Blackwell.
- . 2011. « The Puzzle of Metacoherence ». *Philosophy and Phenomenological Research* 82 (1): 1–21.
- Hughes, G. E, and M. J Cresswell. 1996. *A New Introduction to Modal Logic*. London; New York: Routledge.
- Hume, David. 1978. *A Treatise of Human Nature*. Edited by L. A Selby-Bigge and P. H Nidditch. Oxford; New York: Clarendon Press ; Oxford University Press.
- . 1983. *Enquête sur l'entendement humain*. Edited by Michelle BeyssadeTranslated by André Leroy. Paris: Flammarion.
- . 2007. *An Enquiry Concerning Human Understanding*. Edited by P. F Millican. Oxford; New York: Oxford University Press.
- Hurlburt, Russell, and Sarah A. Akhter. 2006. « The Descriptive Experience Sampling Method ». *Phenomenology and the Cognitive Sciences* 5 (3-4): 271–301.
- Hurlburt, Russell, and Eric Schwitzgebel. 2007. *Describing Inner Experience?: Proponent Meets Skeptic*. Cambridge, Mass.: MIT Press.
- Van Inwagen, Peter. 1975. « The Incompatibility of Free Will and Determinism ». *Philosophical Studies* 27 (3): 185–199.
- . 1983. *An Essay on Free Will*. Oxford: Clarendon Press.
- . 2000. « Free Will Remains a Mystery: The Eighth Philosophical Perspectives Lecture ». *Noûs* 34 (s14): 1–19.
- . 2008. « How to Think about the Problem of Free Will ». *The Journal of Ethics* 12 (August 21): 327–341.
- Jacobs, Jonathan D., and Timothy O'Connor. 2013. « Agent Causation in a Neo-Aristotelian Metaphysics ». In *Mental Causation and Ontology*. Oxford: OUP.
- James, William. 1897. « The Will to Believe ». In *The Will to Believe and Other Essays in Popular Philosophy*, 145–183. New York: Longmans, Green, and Co.

- . 1912. *Essays in Radical Empiricism*. Edited by Ralph Barton Perry. New York: Longmans, Green, and co.
- . 1956. « The Dilemma of Determinism ». In *The Will to Believe*, 145–183. New York: Dover Publications.
- . 1983. *The Principles of Psychology*. avec une introduction de George A. Miller. Cambridge: Harvard university press.
- Jaurès, Jean. 1891. « De primis socialismi germanici lineamentis apud Lutherum, Kant, Fichte et Hegel ». Tolosae: A. Chauvin et fils.
- Jeannerod, Marc. 2006. « From Volition to Agency: The Mechanism of Action Recognition and Its Failures ». In *Disorders of Volition*, edited by Natalie Sebanz and Wolfgang Prinz, 175–192. Cambridge: MIT Press.
- Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
- Kane, Robert. 1985. *Free Will and Values*. Albany, NY: State University of New York Press.
- . 1989. « Two Kinds of Incompatibilism ». *Philosophy and Phenomenological Research*: 219–254.
- . 1996. *The significance of free will*. New York: Oxford University Press.
- . 1999. « Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism ». *The Journal of Philosophy* 96 (5): 217–240.
- . 2005. *A Contemporary Introduction to Free Will*. Oxford: Oxford University Press.
- Kant, Emmanuel. 1907. *Fondements de la métaphysique des mœurs*. Translated by Victor Delbos. Paris: Librairie Delagrave.
- Kapitan, Tomis. 2002. « A Master Argument for Incompatibilism? ». In *The Oxford Handbook of Free Will*, edited by Robert Kane, 127–157. Oxford: OUP.
- Kelly, Thomas. 2005. « Moorean Facts and Belief Revision, or Can the Skeptic Win? ». *Philosophical Perspectives* 19 (1) (December): 179–209.
- Klein, Peter. 1995. « Skepticism and Closure: Why the Evil Genius Argument Fails ». *Philosophical Topics* 23 (1): 213–236.
- . 1999. « Human Knowledge and the Infinite Regress of Reasons ». *Noûs* 33: 297–325.
- . 2008. « Skepticism: Ascent and Assent? ». In *Ernest Sosa and His Critics*, edited by John Greco, 112–125. Blackwell Publishing Ltd.
- . 2011. « Skepticism ». *The Stanford Encyclopedia of Philosophy* (Summer 2011 Edition). <http://plato.stanford.edu/archives/sum2011/entries/skepticism/>.
- Koechlin, Etienne, Chrystèle Ody, and Frédérique Kouneiher. 2003. « The Architecture of Cognitive Control in the Human Prefrontal Cortex ». *Science* 302 (5648) (November 14): 1181–1185.
- Korcz, K. A. 2000. « The Causal-Doxastic Theory of the Basing Relation ». *Canadian Journal of Philosophy* 30 (4): 525–550.
- Kornhuber, Hans H., and Lüder Deecke. 1965. « Hirnpotentialänderungen Bei Willkürbewegungen Und Passiven Bewegungen Des Menschen: Bereitschaftspotential Und Reafferente Potentiale ». *Pflüger's Archiv Für Die Gesamte Physiologie Des Menschen Und Der Tiere* 284 (1): 1–17.
- Kripke, Saul A. 1980. *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.
- Lakatos, Imre. 1978. *The Methodology of Scientific Research Programmes*. Cambridge; New York: Cambridge University Press.
- Lehrer, Keith. 1960. « Can We Know That We Have Free Will by Introspection? ». *The Journal of Philosophy* 57 (5): 145–157.
- . 1968. « Cans without Ifs ». *Analysis* 29 (1): 29–32.
- Lehrer, Keith, and Stewart Cohen. 1983. « Justification, Truth, and Coherence ». *Synthese*: 191–207.

- Leibniz, Gottfried. 1969. *Essais de Théodicée : Sur La Bonté de Dieu, La Liberté de L'homme et L'origine Du Mal*. Paris: Garnier-Flammarion.
- Lemos, Noah Marcelino. 2004. *Common Sense: A Contemporary Defense*. New York: Cambridge University Press.
- Lewis, David. 1973. *Counterfactuals*. Cambridge: Harvard University Press.
- . 1986. *On the Plurality of Worlds*. Oxford, UK; New York, NY, USA: B. Blackwell.
- . 1996. « Elusive Knowledge ». *Australasian Journal of Philosophy* 74 (4): 549–567.
- . 2004. « Causation as Influence ». In *Causation and Counterfactuals*, edited by John David Collins, Edward Jonathan Hall, and Laurie Ann Paul, 75 – 106. MIT Press.
- Libet, Benjamin. 1985. « Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action ». *Behavioral and Brain Sciences* 8 (04): 529–539.
- . 1999. « Do We Have Free Will? ». *Journal of Consciousness Studies* 6 6 (8): 47–57.
- . 2004. *Mind Time the Temporal Factor in Consciousness*. Cambridge, Mass.: Harvard University Press.
- Libet, Benjamin, Curtis A. Gleason, Elwood W. Wright, and Dennis K. Pearl. 1983. « Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (readiness-Potential) The Unconscious Initiation of a Freely Voluntary Act ». *Brain* 106 (3): 623–642.
- Lindberg, David C. 1976. *Theories of Vision from Al-Kindi to Kepler*. Chicago: University of Chicago Press.
- Locke, John. 1690. *An Essay Concerning Humane Understanding in Four Books*. London: Thomas Basset.
- . 1979. *The Correspondence of John Locke 4. Letters*. Oxford: Clarendon Pr.
- Lowe, E. 2002. *A Survey of Metaphysics*. Oxford ; New York: Oxford University Press.
- . 2008. *Personal Agency : The Metaphysics of Mind and Action*. Oxford ; New York: Oxford University Press.
- Malebranche, Nicolas. 1979. *Oeuvres. I*. Edited by Geneviève Rodis-Lewis and Germain Malbreil. Paris: Gallimard.
- Markosian, Ned. 1999. « A Compatibilist Version Of The Theory Of Agent Causation ». *Pacific Philosophical Quarterly* 80 (3) (September 1): 257–277.
- . 2012. « Agent Causation as the Solution to All the Compatibilist's Problems ». *Philosophical Studies* 157 (3): 383–398.
- Martin, Charles B. 1993. « Power for Realists ». In *Ontology, Causality and Mind*, edited by John Heil, 175–186. Dordrecht: Kluwer.
- McGrew, Timothy J, and Lydia McGrew. 2007. *Internalism and Epistemology the Architecture of Reason*. London; New York: Routledge.
- Mele, Alfred R. 1999. « Ultimate Responsibility and Dumb Luck ». *Social Philosophy and Policy* 16: 274–293.
- . 2008. « RECENT WORK ON FREE WILL AND SCIENCE (English) ». *Am. Philos. Q. (Oxf.)* 45 (2): 107–130.
- . 2009. *Effective Intentions: The Power of Conscious Will*. Oxford University Press, USA.
- Mercier, Désiré-Joseph. 1911. *Cours de philosophie. Vol. IV, Critériologie générale ou traité général de la certitude*. 6th ed. Louvain: Institut Supérieur de Philosophie.
- Michon, Cyrille. 2011. *Qu'est-ce que le libre arbitre ?*. Paris: J. Vrin.
- Michotte, A. 1954. *La perception de la causalité*. Louvain: Publications universitaires de Louvain.
- Milhet, S., and A. Streri. 1988. « Equivalences Intermodales de La Forme Des Objets Entre La Vision et Le Toucher Chez Les Bébés de 2 Mois ». *L'année Psychologique* 88 (3): 329–341.

- Mill, John Stuart. 1865. *An Examination of Sir William Hamilton's Philosophy and of the Principal Philosophical Questions Discussed in His Writings*. 2e ed. London: Longmans, Green, Reader, & Dyer.
- . 1973. *Collected Works of John Stuart Mill*. Toronto ;London: Univ. of Toronto P. ;;Routledge & Kegan Paul.
- . 1974. *A System of Logic, Ratiocinative and Inductive : Being a Connected View of the Principles of Scientific Investigation*. Toronto ;London: University of Toronto Press;Routledge & Kegan Paul.
- . 1981. *Autobiography and Literary Essays*. Toronto ;;Buffalo ;London: University of Toronto Press ;;Routledge & Kegan Paul.
- Molnar, George. 2006. *Powers : A Study in Metaphysics*. Edited by Stephen Mumford. Oxford: Clarendon.
- Moore, G. E. 1939. « Proof of an External World ». In *Proceedings of the British Academy*. Vol. 25. London: Published for the British Academy by the Oxford University Press.
- . 1942. « A Reply to My Critics ». In *The Philosophy of G.E. Moore*, edited by P. Schilpp, 535–677. La Salle, Ill: Open Court.
- . 1962. *Commonplace Book, 1919-1953*. London; New York: Allen & Unwin; Macmillan.
- . 1993a. « Moore's Paradox ». In *Selected Writings*, edited by Thomas Baldwin, 207–212. London: Routledge.
- . 1993b. « A Defence of Common Sense ». In *Selected Writings*, edited by Thomas Baldwin. London; New York: Routledge.
- . 2005. *Ethics*. Edited by William H Shaw. Oxford: Clarendon.
- Mumford, Stephen. 2008. « Powers, Dispositions, Properties or a Causal Realist Manifesto ». In *Revitalizing Causality: Realism about Causality in Philosophy and Social Science*, 139–151. New York: Routledge.
- Mumford, Stephen, and Rani Lill Anjum. 2011. *Getting Causes from Powers*. Oxford; New York: Oxford University Press.
- Nahmias, Eddy. 2002. « When Consciousness Matters: A Critical Review of Daniel Wegner's The Illusion of Conscious Will ». *Philosophical Psychology* 15 (4): 527–541.
- . 2010. « Scientific Challenges to Free Will ». In *A Companion to the Philosophy of Action*, edited by Timothy O'Connor and Sandis Constantine, 345–356. Malden, MA: Wiley-Blackwell.
- Nahmias, Eddy, D. Justin Coates, and Trevor Kvaran. 2007. « Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions ». *Midwest Studies in Philosophy* 31 (1): 214–242.
- Nahmias, Eddy, Stephen Morris, Thomas Nadelhoffer, and Jason Turner. 2004. « The Phenomenology of Free Will ». *Journal of Consciousness Studies*, 11 7 (8): 162–179.
- . 2005. « Surveying Freedom: Folk Intuitions about Free Will and Moral Responsibility ». *Philosophical Psychology* 18 (5): 561–584.
- . 2006. « Is Incompatibilism Intuitive? ». *Philosophy and Phenomenological Research* 73 (1): 28–53.
- Nahmias, Eddy, and D. Murray. 2011. « Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions ». *New Waves in Philosophy of Action*: 189–217.
- Nelkin, Dana Kay. 2011. *Making Sense of Freedom and Responsibility*. Oxford University Press.
- Nelkin, D.K. 2004. « The Sense of Freedom ». In *Freedom and Determinism*, edited by Joseph Keim Campbell, Michael O'Rourke, and David Shier, 105–134. Cambridge Mass.: MIT Press.

- Nichols, Shaun. 2004. « The Folk Psychology of Free Will: Fits and Starts ». *Mind and Language* 19 (5) (November): 473–502.
- Nichols, Shaun, and Joshua Knobe. 2007. « Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions ». *Nous* 41 (4): 663–685.
- Nida-Rümelin, Martine. 2007. « Doings and Subject Causation ». *Erkenntnis* 67 (2) (September 6): 255–272.
- Nozick, Robert. 1981. *Philosophical Explanations*. Cambridge, Mass.: Harvard University Press.
- O'Connor, Timothy. 1995. « Agent Causation ». In *Agents, Causes, and Events : Essays on Indeterminism and Free Will*, 173–200. New York: Oxford University Press.
- . 2000. *Persons and Causes the Metaphysics of Free Will*. New York: Oxford University Press.
- . 2011. « Agent-Causal Theories of Freedom ». In *Oxford Handbook of Free Will*, edited by Robert Kane, 2e édition, 309–328. Oxford: OUP.
- Olsson, Erik. 2012. « Coherentist Theories of Epistemic Justification ». In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2012.
- Oppy, Graham Robert. 1995. *Ontological Arguments and Belief in God*. Cambridge; New York: Cambridge University Press.
- Pacherie, Elisabeth. 2006. « Toward a Dynamic Theory of Intentions ». In *Does Consciousness Cause Behavior*, edited by Susan Pockett, William P Banks, and Shaun Gallagher, 145–167. Cambridge, Mass.: MIT Press.
- . 2007. « The Anarchic Hand Syndrome and Utilization Behavior: A Window Onto Agentive Self-Awareness ». *Functional Neurology* 22 (4): 211–217.
- . 2008. « The Phenomenology of Action: A Conceptual Framework ». *Cognition* 107 (1) (April): 179–217.
- Peacocke, Christopher. 1999. *Being Known*. Oxford; Oxford; New York: Clarendon Press ; Oxford University Press.
- Pereboom, Derk. à paraître. « The Phenomenology of Agency and Deterministic Agent-Causation ». In *Horizons of Authenticity in Phenomenology, Existentialism, and Moral Psychology - Essays in Honor of Charles Guignon*, edited by Hans Pedersen and Megan Altman. New York: Springer.
- . 2001. *Living without Free Will*. Cambridge; New York: Cambridge University Press.
- Perry, Ralph Barton. 1936. *The Thought and Character of William James as Revealed in Unpublished Correspondence and Notes, Together with His Published Writings*. Vol. I. Boston: Little, Brown & C0.
- Petitmengin, Claire. 2006. « Describing One's Subjective Experience in the Second Person: An Interview Method for the Science of Consciousness ». *Phenomenology and the Cognitive Sciences* 5 (3-4): 229–269.
- Plantinga, Alvin. 1993. *Warrant : The Current Debate*. New York: Oxford University Press.
- . 1996. « Respondeo : Ad Feldman ». In *Warrant in Contemporary Epistemology*, edited by Jonathan L Kvanvig, 357–361. Lanham, Md.: Rowman & Littlefield Publishers.
- . 2000. *Warranted Christian Belief*. New York: Oxford University Press.
- Pockett, Susan, William P Banks, and Shaun Gallagher. 2006. *Does Consciousness Cause Behavior?*. Cambridge (Massachusetts): MIT.
- Pollock, John L, and Joseph Cruz. 1999. *Contemporary Theories of Knowledge*. Lanham, Md.: Rowman & Littlefield Publishers.
- Popper, Karl R. 1963. *Conjectures and Refutations. The growth of scientific knowledge. (Essays and lectures.)*. London: Routledge & Kegan Paul.

- . 1977. *The Open Society and Its Enemies 2. The High Tide of Prophecy: Hegel, Marx, and the Aftermath*. London [u.a.]: Routledge & Kegan Paul.
- Post, J. F. 1980. « Infinite Regresses of Justification and of Explanation ». *Philosophical Studies* 38 (1): 31–52.
- Pouivet, Roger. 2003. *Qu'est-ce que croire?*. Paris: Vrin.
- Prinz, Wolfgang, Daniel Dennett, and Natalie Sebanz. 2006. « Toward a Science of Volition. ».
- Pritchard, Duncan. 2005. *Epistemic Luck*. New York: Oxford University Press.
- Proust, Joëlle. 2005. *La nature de la volonté*. [Paris]: Gallimard.
- . 2008. « Réponses À Mes Critiques ». *Philosophiques*: 35 (1): 139.
- Pruss, Alexander. 2006. *The Principle of Sufficient Reason, a Reassessment*. Cambridge: Cambridge University Press.
- Pryor, James. 2000. « The Skeptic and the Dogmatist ». *Nous* 34 (4) (December): 517–549.
- . 2001. « Highlights of Recent Epistemology ». *The British Journal for the Philosophy of Science* 52 (1): 95.
- . 2004. « What's Wrong with Moore's Argument? ». *Philosophical Issues* 14 (1): 349–378.
- . 2006. « XIII*—Hyper-Reliability and A Priority ». In *Proceedings of the Aristotelian Society*, 106:329–346.
- . 2007. « When Warrant Transmits ». In *Meaning, Epistemology and Mind. Themes from the Philosophy of Crispin Wright*. Oxford: Oxford University Press.
- Pust, Joel. 2000. *Intuitions as Evidence*. New York: Garland Pub.
- . 2012. « Intuition ». In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2012.
- Quine, W. V. 1951. « Two Dogmas of Empiricism ». *The Philosophical Review* 60 (1) (January): 20.
- . 1960. *Word and Object*. [Cambridge: Technology Press of the Massachusetts Institute of Technology.
- Quine, W. V., and J. S Ullian. 1978. *The Web of Belief*. New York: Random House.
- Ragland, Clyde Prescott. 2006. « Is Descartes a Libertarian? ». In *Oxford Studies in Early Modern Philosophy*, 3:57–90.
- Rea, Michael C. 2002. *World without Design: The Ontological Consequences of Naturalism*. Oxford; New York: Clarendon Press ; Oxford University Press.
- Reed, B. 2002. « How to Think about Fallibilism ». *Philosophical Studies* 107 (2): 143–157.
- Reid, Thomas. 1764. *An Inquiry into the Human Mind on : The Principles of Common Sense*. Dublin: Alexander Ewing.
- . 1785. *Essays on the Intellectual Powers of Man*. Edinburgh: Printed for John Bell, and G.G.J. & J. Robinson, London.
- . 1788. *Essays on the Active Powers of Man By Thomas Reid*. Edinburgh: Printed for John Bell and G.G.J. & J. Robinson London.
- Roessler, Johannes, and Naomi Eilan. 2003. *Agency and Self-Awareness: Issues in Philosophy and Psychology*. Clarendon Press.
- Roskies, Adina. 2010. « Why Libet's Studies Don't Pose a Threat to Free Will ». *Conscious Will and Responsibility: A Tribute to Benjamin Libet*: 11–22.
- Roskies, Adina, and Shaun Nichols. 2008. « Bringing Moral Responsibility down to Earth ». *The Journal of Philosophy* 105 (7): 371–388.
- Russell, Bertrand. 2009. *Human Knowledge Its Scope and Limits*. Edited by John G Slater. London; New York: Routledge.
- Rysiew, Patrick. 2011. « Epistemic Contextualism ». In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2011.

- Salles, Ricardo. 2005. *The Stoics on Determinism and Compatibilism*. Aldershot, Hants, England; Burlington, VT: Ashgate Pub.
- Sanger, Lawrence Mark. 1999. « Epistemic Circularity: An Essay on the Problem of Meta-Justification ». Ohio State University.
- Sarkissian, Hagop, Amita Chatterjee, Felipe De Brigard, Joshua Knobe, Shaun Nichols, and Smita Sirker. 2010. « Is Belief in Free Will a Cultural Universal? ». *Mind & Language* 25 (3) (June 1): 346–358.
- Schlick, Moritz. 1939. « When Is a Man Responsible? ». In *Problems of Ethics*, 143–56.
- Schmitt, Frederick F. 2004. « What Is Wrong with Epistemic Circularity? ». *Philosophical Issues* 14 (1): 379–402.
- Schooler, Jonathan W. 2010. « What Science Tells Us about Free Will ». In *Free Will and Consciousness*, edited by Roy F. Baumeister, Alfred R. Mele, and Kathleen D. Vohs, 191–218. Oxford ; New York: Oxford University Press, UK,.
- Scott, George P. 1991. « Dissipative Structures and the Mind-Body Problem ». *Scott (ed.) Time, Rhythms and Chaos in the New Dialogue with Nature (Iowa State)*.
- Searle, John R. 1984. *Minds, Brains, and Science*. Cambridge, Mass.: Harvard University Press.
- Sebanz, Natalie, and Wolfgang Prinz. 2006. *Disorders of Volition*. Cambridge, Mass.: MIT Press.
- Shariff, Azim F., Jonathan Schooler, and Kathleen D. Vohs. 2008. « The Hazards of Claiming to Have Solved the Hard Problem of Free Will ». In *Are We Free? Psychology and Free Will*, edited by J. Baer, J. Kaufman, and R. Baumeister, 181. New York: OUP.
- Shoemaker, Sydney. 1995. « Moore's Paradox and Self-Knowledge ». *Philosophical Studies* 77 (2-3) (March): 211–228.
- Siegel, Susanna. 2008. « The Phenomenology of Efficacy ». *Philosophical Topics* 33 (1): 265–284.
- . 2011. « The Epistemic Impact of the Etiology of Experience ». *Philosophical Studies*.
- Smilansky, Saul. 2000. *Free Will and Illusion*. Oxford; New York: Clarendon Press ; Oxford University Press.
- . 2003. « Compatibilism: The Argument from Shallowness ». *Philosophical Studies* 115 (3): 257–282.
- Soames, Scott. 2003. *Philosophical Analysis in the Twentieth Century Vol. 1, The Dawn of Analysis*. Princeton, N.J., [etc.]: Princeton University Press.
- Soon, Chun Siong, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. 2008. « Unconscious Determinants of Free Decisions in the Human Brain ». *Nature Neuroscience* 11 (5): 543–545.
- Sosa, Ernest. 1991. *Knowledge in Perspective : Selected Essays in Epistemology*. Cambridge [England]; New York: Cambridge University Press.
- . 2000. « Skepticism and Contextualism ». *Philosophical Issues* 10 (1): 1–18.
- . 2007. « Moore's Proof ». In *Themes from GE Moore: New Essays in Epistemology and Ethics*, edited by Susana Nuccetelli and Gary Seay, 49–61. New York: Oxford University Press.
- Spence, Sean A., and Chris D. Frith. 1999. « Towards a Functional Anatomy of Volition ». *Journal of Consciousness Studies* 6 (8-9): 8–9.
- Spinoza, Baruch. 1986. *Traité Politique Lettres*. Nouveau tirage. Paris: Garnier-Flammarion.
- Stalnaker, Robert. 1968. « A Theory of Conditionals ». *Studies in Logical Theory* 2: 98–112.
- Stanley, J., and Z. Gendler Szabó. 2003. « On Quantifier Domain Restriction ». *Mind & Language* 15 (2-3): 219–261.
- Steward, Helen. 2012. *A Metaphysics for Freedom*. Oxford: Oxford University Press.

- Strawson, Galen. 1986. *Freedom and Belief*. Oxford [Oxfordshire]; Oxford; New York: Clarendon Press; Oxford University Press.
- . 1994. « The Impossibility of Moral Responsibility ». *Philosophical Studies* 75 (1): 5–24.
- Strawson, Peter Frederick. 2008. « Freedom and Resentment ». In *Freedom and Resentment and Other Essays*, 1–28. New York: Routledge.
- Stroud, Barry. 1984. *The Significance of Philosophical Scepticism*. Oxford [Oxfordshire]; New York: Clarendon Press ; Oxford University Press.
- Taylor, Richard. 1963. *Metaphysics*. Englewood Cliffs, N.J.: Prentice-Hall.
- . 1991. *Metaphysics*. 4th edition. Englewood Cliffs, N.J.: Prentice-Hall.
- Tegmark, Max. 2000. « Importance of Quantum Decoherence in Brain Processes ». *Physical Review E* 61 (4): 4194.
- Timpe, Kevin. 2008. *Free Will Sourcehood and Its Alternatives*. London; New York: Continuum.
- Turri, John. 2010. « On the Relationship between Propositional and Doxastic Justification ». *Philosophy and Phenomenological Research* 80 (2): 312–326.
- Unger, Peter. 1968. « An Analysis of Factual Knowledge ». *The Journal of Philosophy* 65 (6) (March 21): 157.
- . 1975. *Ignorance : A Case for Scepticism*. Oxford [Eng.]: Clarendon Press.
- Vaidya, Anand. 2011. « The Epistemology of Modality ». In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2011.
- Vargas, Manuel. 2011. « The Revisionist Turn: A Brief History of Recent Work on Free Will ». In *New Waves in the Philosophy of Action*, edited by Jesús H Aguilar, Andrei A Buckareff, and Keith Frankish. Houndmills, Basingstoke, Hampshire; New York: Palgrave Macmillan.
- Vihvelin, Kadri. 2004. « Free Will Demystified: A Dispositional Account ». *Philosophical Topics* 32.
- . 2011. « Arguments for Incompatibilism ». In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2011.
- Vivian, Frederick. 1964. *Human Freedom and Responsibility*. London: Chatto & Windus.
- Vogel, Jonathan. 2000. « Reliabilism Leveled ». *The Journal of Philosophy* 97 (11) (November 1): 602–623.
- . 2009. « Epistemic Bootstrapping ». *The Journal of Philosophy* 105 (9): 518–539.
- Vohs, Kathleen D., and Jonathan W. Schooler. 2008. « The Value of Believing in Free Will Encouraging a Belief in Determinism Increases Cheating ». *Psychological Science* 19 (1): 49–54.
- Watson, Gary. 2003. « Introduction ». In *Free Will*, edited by Gary Watson, 2e ed., 1–25. Oxford: Oxford university press.
- Wegner, Daniel M. 2002. *The illusion of conscious will*. 1 vols. Bradford books. Cambridge, Mass.: MIT Press.
- Wegner, Daniel M., and Thalia Wheatley. 1999. « Apparent Mental Causation: Sources of the Experience of Will. ». *American Psychologist* 54 (7): 480.
- Whitehead, Alfred North. 1923. *Symbolism, Its Meaning and Effect: Barbour-Page Lectures, University of Virginia 1927*. Cambridge: Cambridge University Press.
- Wilder, Billy. 1959. *Some like It Hot*. Comédie. United Artists Corporations.
- Williams, Michael. 1991. *Unnatural Doubts: Epistemological Realism and the Basis of Scepticism*. Oxford, UK; Cambridge, USA: B. Blackwell.
- Williamson, Timothy. 2000. *Knowledge and Its Limits*. Oxford ; New York: Oxford University Press.
- . 2007. *The Philosophy of Philosophy*. Malden, MA: Blackwell Pub.

- Wittgenstein, Ludwig. 2005. *Recherches philosophiques*. Edited by Élisabeth Rigal. [Paris]: Gallimard.
- Wolf, Susan R. 1990. *Freedom within Reason*. New York: Oxford University Press.
- Wood, L. 1941. « The Free-Will Controversy ». *Philosophy* 16 (64): 386–397.
- Wright, Crispin. 2000. « Cogency and Question-Begging: Some Reflections on McKinsey's Paradox and Putnam's Proof ». *Philosophical Issues* 10 (1): 140–163.
- . 2004. « Warrant for Nothing (and Foundations for Free)? ». In *Aristotelian Society Supplementary Volume*, 78:167–212.
- . 2007. « The Perils of Dogmatism ». In *Themes from GE Moore: New Essays in Epistemology and Ethics*, edited by Susana Nuccetelli and Gary Seay, 25–48. New York: Oxford University Press.
- Yaffe, Gideon. 2004. *Manifest Activity : Thomas Reid's Theory of Action*. Oxford ;New York: Clarendon Press ;Oxford University Press.
- Yaffe, Gideon, and Ryan Nichols. 2000. « Thomas Reid ». *Stanford Encyclopedia of Philosophy*. < [Http://plato. Stanford. Edu/entries/reid](http://plato.stanford.edu/entries/reid).
- Zalabardo, José L. 2005. « Externalism, Skepticism, and the Problem of Easy Knowledge ». *The Philosophical Review*: 33–61.

Table des matières générale

Introduction : la question épistémologique	2
I. Le libre arbitre : concept et problèmes	4
II. Connaissance et Justification : définitions	16
III. Le Problème d'Intégration du Libre arbitre	27
IV. Plan général de cette étude	32
 PREMIERE PARTIE : PRESOMPTION DE CONNAISSANCE DU LIBRE ARBITRE	39
 Chapitre 1 : L'expérience commune et la définition de « libre arbitre »	40
I. Difficultés pour définir le concept de « libre arbitre »	41
II. Le conflit intuitif originel : solution descriptiviste pour fixer la référence du terme « libre arbitre »	49
III. L'existence du conflit intuitif, et l'expérience du libre arbitre	63
IV. Conclusion : L'expérience du Libre arbitre comme point de départ.....	76
 Chapitre 2 : Justification <i>prima facie</i> à croire que nous sommes libres	79
I. Introspectionnisme vs Anti-Introspectionnisme.....	81
II. Trois interprétations de l'objection anti-introspection	99
III. Conclusion : la justification indexicale de premier ordre en faveur du Libre Arbitre	170
 Chapitre 3 : Justification <i>prima facie</i> à croire que nous savons que nous sommes libres	178
I. Deux types d'objections méta-épistémiques à l'introspectionnisme	181
II. L'objection de l'impression non étayée	189
III. L'objection de l'impression non étayable	217
IV. Conclusion : la justification indexicale de second ordre et le « point de départ épistémologique » du problème du libre arbitre.....	235

DEUXIÈME PARTIE : ABSENCE DE RÉFUTATIONS	247
Chapitre 4 : Le contenu incompatible de notre expérience du libre arbitre.....	248
I. Le conflit conceptuel comme meilleure explication du conflit original	251
II. Comment être (et comment ne pas être) compatible	255
III. Echec des théories compatibles du conflit illusoire	280
IV. Analyse phénoménologique de notre expérience du libre arbitre	307
V. Conclusion.....	344
Chapitre 5 : Absence de réfutation directe du libertarianisme.....	346
I. Réfutations a priori	348
II. Réfutations empiriques.....	381
III. Conclusion : une « information recevable ».....	398
Chapitre 6 : Absence de réfutation fondationnelle : la croyance libertarienne comme connaissance	400
I. Réfutations par la source faible	402
II. Réfutation par le contenu alourdi : le Problème d’Intégration.....	417
Conclusion : le libertarianisme, à l’égal des croyances perceptives.....	466
I. Camp de base et positions de repli	466
II. N’y a-t-il vraiment aucune différence (pertinente) entre perception et libertarianisme ?	477
ANNEXES	489
Annexe 1 : Pourquoi la notion internaliste de « Justification » ne préjuge pas de l’Internalisme ou de l’Externalisme	490
Annexe 2 : Métaphilosophie ou Epistémologie de la philosophie ?.....	495
Annexe 3 : Table récapitulative des objections épistémologiques à l’introspectionnisme libertarien	500
Annexe 4 : Les grandes positions possibles dans le débat sur le libre arbitre	504
Annexe 5 : Van Inwagen on Introspected Freedom	505
Bibliographie	520
Table des matières générale.....	534

Thèse de Doctorat

Résumé

La philosophie analytique a réalisé, depuis une cinquantaine d'années, des progrès considérables en philosophie du libre arbitre et en philosophie de la connaissance (épistémologie). Jusqu'à présent, ces différents résultats n'ont pas été réunis pour proposer une épistémologie du libre arbitre. Pourtant, la question du libre arbitre soulève, depuis l'antiquité mais de manière plus pressante aujourd'hui, un problème qui est proprement épistémologique. On peut le formuler comme un trilemme : (1) il semble que nous savons que nous sommes libres, (2) il semble également que le libre arbitre requière des indéterminismes causaux, mais pourtant (3) il semble impossible de connaître l'existence d'indéterminismes causaux sans une étude scientifique approfondie. C'est ce que Christopher Peacocke a appelé un « défi d'intégration » : comment « intégrer » notre métaphysique du libre arbitre avec une épistémologie qui nous permette de le connaître ? Dans cette étude, je m'efforce de reconfigurer le débat contemporain sur le libre arbitre en présentant les différentes thèses existantes (compatibilisme, libertarisme, indéterminisme « dur ») comme différentes réponses à ce défi d'intégration. Ultimement, je défends une réponse « libertarienne introspectionniste » : (1) nous connaissons effectivement l'existence du libre arbitre sur la base de l'expérience que nous en faisons communément (introspectionnisme), (2) le libre arbitre requière effectivement des indéterminismes causaux (incompatibilisme), mais (3) nous pouvons donner un modèle épistémologique dans lequel notre expérience commune du libre arbitre constitue une véritable source de connaissance de l'indéterminisme causal.

Mots clés

Libre arbitre, scepticisme, expérience, justification, introspection, incompatibilisme, indéterminisme, fiabilisme, connaissance, clôture épistémique.

Abstract

In the last fifty years, the philosophy of free will, and the philosophy of knowledge (epistemology) have made considerable progress in analytic philosophy. Nevertheless, those results haven't been brought together to provide an epistemology of free will. This is a pity since the issue of free will raises a properly epistemological problem, a problem that is traditional but still pressing nowadays. The problem can be presented as a trilemma: (1) it seems that we know that we have free will, (2) it also seems that free will requires causal indeterminisms, and yet (3) it seems that we couldn't possibly know about causal indeterminisms without deep scientific investigation. This is what Christopher Peacocke has called an "Integration Challenge": how can we "integrate" our metaphysics of free will with an epistemology that would let us *know* about it?

In this study, my purpose is to reshape the state of the contemporary debate, showing that every existing thesis can be presented as a possible response to the Integration Challenge. Ultimately, I defend the "libertarian introspectionist" response: (1) we do know the existence of free will from the experience we commonly have of it in action (introspectionism), (2) free will does entail the existence of causal indeterminisms (incompatibilism), but (3) we can make up an epistemological model in which our common experience of free will gives us genuine knowledge of causal indeterminism.

Key Words

Free Will, scepticism, experience, justification, introspection, incompatibilism, indeterminism, reliabilism, knowledge, epistemic closure.