N° d'ordre : 2269

Thèse
présentée à

L'Université des Sciences et Technologies de Lille

en vue de l'obtention du titre de
Docteur de l'Université de Lille
Discipline : Informatique

présentée et soutenue publiquement par

Bruno Marchal

Calculabilité, Physique et Cognition

Date de soutenance : mardi 2 juin 1998

Jury :

Max DAUCHET	Université de Lille I	Président
Jean-Paul DELAHAYE	Université de Lille I	Directeur
Serge GRIGORIEFF	Université Paris VII	Rapporteur
Patrice ENJALBERT	Université de Caen	Rapporteur
Paul GOCHET	Université de Liège	Rapporteur
François DENIS	Université de Lille III	Examinateur
Philippe SMETS	Université Libre de Bruxelles	Examinateur

IRIDIA/Bruxelles, LIFL/Lille

Table des matières

Remerciements

Le présent travail est le fruit d'une recherche de très longue haleine. J'ai posé en 1963, lors d'un exposé scolaire intitulé "L'amibe, l'euglène et la paramécie" la question principale qui est à l'origine de la démonstration exposée ici dans les trente premières pages. En gros il s'agit de la question "Quel effet cela fait-il d'être une amibe qui se divise en deux?" Jusqu'en 1971, j'hésiterai entre les outils de la biologie et les outils de la chimie pour aborder cette question. En 1971, essentiellement grâce au petit livre de Nagel et Newman sur le théorème de Gödel, je réaliserai la possibilité d'aborder cette question et d'autres questions autoréférentielles, de façon mathématique. C'est ce qui est reflété dans les trente dernières pages du présent travail (annexes non comprises).

J'ai reçu des encouragements multiples et constant à l'Université Libre de Bruxelles et dans d'autres universités, et j'aurai la chance de bénéficier de nombreuses discussions encourageantes avec beaucoup de personnes. Il ne me sera pas possible de les remercier toutes et j'espère avoir l'indulgence de ceux que j'oublie de mentionner.

Il va de soi que ces remerciements ne signifient pas que ces personnes apprécient automatiquement les hypothèses du travail, ni qu'elles apprécient nécessairement les conclusions, ni qu'elles sont persuadées qu'ils ne subsistent pas quelques maladresses ou quelques erreurs dans la démonstration et dans la prospection proposée.

Je remercie la regrettée Mademoiselle Lucia de Brouckère, de l'Université Libre de Bruxelles, et mon maître Jean Robberecht, de l'Athénée Robert Catteau, de m'avoir inculqué l'esprit scientifique (je veux dire l'amour de la clarté, de la rigueur et l'amour de la liberté de la pensée) et m'avoir fait apprécier la chimie moderne. C'est grâce à mon goût pour la chimie que je n'ai jamais pris la notion de matière pour argent comptant, et que j'ai suivi de près les questions d'interprétation de la mécanique quantique.

Je remercie Messieurs Jean Rommelaere et René Thomas, de l'Université Libre de Bruxelles, pour m'avoir accueilli à différents moments au laboratoire de biologie moléculaire de Rhode-St-Genèse. Je remercie particulièrement René Thomas pour les conversations sur Lewis Carroll et sur la logique formelle.

Je remercie Monsieur Jean Ladrière, de l'Université Catholique de Louvain, pour m'avoir offert un exemplaire de sa merveilleuse thèse sur les limitations internes des formalismes. Cela m'a permis de rendre définitivement "mon" ex-

emplaire à la Bibliothèque Nationale et de continuer à consulter régulièrement cet ouvrage qui reste ma référence préférée pour la période prélöbienne des phénomènes d'incomplétude. Le chapitre où Monsieur Jean Ladrière expose ses suggestions philosophiques m'a profondément inspiré. Je le remercie aussi pour m'avoir encouragé à l'époque, dans mes réflexions sur Gödel, sur la logique modale, et sur la biologie, en m'invitant notamment à Louvain pour exposer les travaux "logico-génétique" de René Thomas.

Je remercie Monsieur François Englert pour m'avoir accueilli deux années au laboratoire de cosmologie de l'Université Libre de Bruxelles. Cela m'a permis d'approfondir les fondements de la mécanique quantique et d'étudier les formidables problèmes posés par l'utilisation de la mécanique quantique en cosmologie.

Je remercie Monsieur Georges Papy pour m'avoir permis d'effectuer mon service civil d'objecteur de conscience au service d'Algèbre de l'Université Libre de Bruxelles. J'ai pu y enseigner la programmation à des enfants handicapés et à des professeurs de l'enseignement secondaire. J'ai pu enseigner, le soir, la programmation fonctionnelle et la programmation logique. J'ai pu organiser un club informatique à l'athénée Maïmonide et surtout, j'ai pu développer une méthodologie pour expliquer le fonctionnement logique de l'ordinateur à de très jeunes enfants. C'est ce travail qui m'a conduit à voir l'ordinateur comme un graphe, et c'est dans les locaux du service d'Algèbre que j'ai développé l'argument du graphe filmé. Je remercie Madame Frédérique Papy pour sa gentillesse et pour ses conseils pertinents concernant les difficultés cognitives des enfants mentalement handicapés.

Je remercie particulièrement Monsieur Paul Gochet, de l'Université de Liège, pour m'avoir sérieusement encouragé dans mon approche des questions cognitives, et pour m'avoir incité à la fin des années 80 à publier une partie de mes travaux. Grâce à ces publications j'ai été invité plusieurs fois en France où j'ai pu exposer mes résultats à un public chaleureux et encourageant, notamment par Monsieur Jacques Pitrat à Paris, Monsieur Mario Borillo à Toulouse et Monsieur Jean-Louis Chrétien à Grenoble. Je les remercie tous pour ces heureux moments.

Je remercie chaleureusement Monsieur Philippe Smets pour m'avoir accueilli à l'IRIDIA (Institut de Recherche Interdisciplinaires et de Développements en Intelligence Artificielle, ULB). Je le remercie pour les nombreuses conversations passionnantes sur le calcul des probabilités et sur les fonctions de croyances, et je tiens à le remercier aussi pour m'avoir aidé à financer une version préliminaire du présent travail, notamment au moyen du projet national Verhofstadt pour la recherche fondamentale en Intelligence Artificielle et au moyen des projets internationaux BELON et ESPRIT. L'IRIDIA fut pour moi un véritable havre de paix et d'écoute.

Je remercie très chaleureusement Monsieur Jean-Paul Delahaye pour avoir accepter de diriger la présente thèse à l'Université des Sciences et Technologies

de Lille, et d'avoir rendu possible la soutenance. Je le remercie de m'avoir encouragé à être le plus court et le plus clair possible, et de m'avoir poussé à mettre en relief la contribution originale au problème du corps et de l'esprit. Sa vigilance critique et amicale m'a encouragé à isoler la présentation hypothético-déductive de l'argumentation et m'a permis d'éviter de nombreux écueils. Son article dans "POUR LA SCIENCE" m'a permis d'avoir une correspondance électronique importante et intéressante sur mon travail avec des lecteurs du journal, que je remercie aussi pour leurs remarques.

Je remercie Messieurs Serge Grigorieff, de l'Université de Paris VII, Patrice Enjalbert de l'Université de Caen, et Paul Gochet pour avoir accepté de rapporter ce travail. Je les remercie pour leur diligence, et aussi de m'avoir fait part de dernières corrections et remarques pertinentes.

Je remercie Messieurs Max Dauchet et François Denis de l'Université des Sciences et Technologies de Lille, ainsi que Monsieur Philippe Smets pour avoir accepté d'examiner le travail. Je remercie en particulier Monsieur Max Dauchet pour me faire l'honneur de présider le jury.

Je tiens encore à remercier tout ceux qui m'ont soutenu de près ou de loin dans le développement de ce travail. Je remercie chaleureusement Hugues Bersini pour son grand soutien moral. Je remercie mon ami Robert Kennes (pour son assistance LaTeX, mais aussi pour avoir découvert l'article de Maudlin de 1989 à la bibliothèque de la VUB !), je remercie Adrien Sluys, Frédéric Janssens, Daniel Lehman, Philippe Grotard (qui m'a abreuvé ces dernières années de nouveaux articles sur les fondements de la physique quantique !), Je remercie Coby et Nicole Avidar et leurs enfants, Anne de Rudder, Althea Williams, Brigitte Horlait, Alice Horlait, Lina Baré et André Haucotte (qui m'a fait découvrir et partager le charme du TRS 80 !), Edwin Zaccaï, Nathalie Reyners, Charles et Antoinette Nzajyibwami, pour leur patience et leur gentillesse. Je remercie Vincent Detours (pour son indéfectible soutien OFFICIEL et électronique), Georges Miedzianagora (pour les innombrables discussions à la cafetaria BEPPINO depuis si longtemps et pour sa compassion dans les moments difficiles), Serge Pahaut, Georges Elencwajg, la regrettée Viviane Munyandamutsa, Isabelle Stengers, Mony et Olga Elkaïm, Christoffe Schiller, Monique Rémy, Paul et Roland Van Praag, Henry Thieren, Paul-Louis Van Berg, et Alessandro Saffiotti (notamment pour son humour tendre et poétique). Je remercie mes amis et collègues de l'IRIDIA, Marco Saerens, Moustapha Hamzaoui, Philip Miller, Christine Defrise, Paul Magrez, Jean-Pierre Nordvik, Jacques Marée, Hong Xu, Karsten Jöred, Marc Klein, Selwyn Piramuthu, Tristan Salomé, Thierry Van de Merckt, Stéphane Amarger, Gianni Di Caro, Jorge Gasós, Alain Soquet, Willy Serniclaes, Francesco Allevi, Masaaki Minagawa, Gianluca Bontempi, Antoine Duchâteau, Emanuele Persico, Nick Bradshaw, Mauro Birattari, Vera Calenbuhr, Vittorio Gorrini, Elizabeth Umkehrer, Youbin Peng, Marco Dorigo, Ben Burdsall, Yen-Teh Hsia, Laurence Vignollet, Truong Quoc Dung, Philippe Besnard, Salem Benferhat, et Victor Poznan-

sky. Je remercie Michel Bardiaux pour les nombreuses conversations sur l'IA pratique et théorique.

Je remercie mes vieux amis Dominique Thirion et Pierre Barbier pour avoir supporté la genèse et le développement de mes obsessions amibiennes (pardon), ainsi que mon vieux copain Denis Goldschmidt pour m'avoir longtemps prêté son extraordinaire microscope, sans lequel je n'aurais pu si bien contempler autant de divisions cellulaires : ma véritable source d'inspiration.

Je remercie ceux qui m'ont invité ces trois dernières années à exposer mes travaux, en particulier je remercie Vincent Rialle, Michel Elias, Francis Rousseau, Jérôme Grynpas, Julien Friedler, Hélène Weemaes, Josette Hector, Claude-Yves Baum, Philippe van Ham, Armand de Callataÿ, Serge Lesens, Fernand Schmetz, Marianne Rooman, Lambros Couloubaritsis, Axel Cleeremans, Jean-Noël Missa, Philippe Van Ham, Paul Jorion, Alexandre Wajnberg, Gérard Pinson. Je remercie Monsieur Albert Visser pour m'avoir reçu à Utrecht.

Je remercie encore chaleureusement Philippe Van Ham et Christine Decaestecker pour leur solide soutien.

Je remercie Giovanna Colombetti de s'intéresser de si près à mon argumentation, et je regrette de ne pas avoir pu tenir davantage compte de ses très nombreuses et judicieuses remarques sur le présent travail (ce n'est que partie remise !).

Je remercie Muriel Decreton, secrétaire de l'IRIDIA, pour la chasse aux fautes d'orthographe, son soutien logistique, son sourire et sa bonne humeur.

Je remercie mon frère et ma soeur, ma nièce et mes neveux, pour leur grande patience et leur support. Je remercie profondément mes regrettés parents pour m'avoir toujours témoigné leur confiance.

Bruxelles, le 18 avril 1998

Introduction

L'hypothèse computationnaliste, ou plus simplement le *mécanisme*, que je considère ici, est l'hypothèse selon laquelle *je* suis une machine, ou *vous* êtes une machine. De façon précise je m'intéresse à l'hypothèse selon laquelle nous pourrions survivre, non seulement avec un coeur artificiel ou un rein artificiel, etc., mais aussi avec un cerveau artificiel digital (finiment descriptible) pour autant qu'il soit configuré convenablement au niveau adéquat.

Le but n'est pas de défendre cette hypothèse, mais d'en étudier les conséquences, notamment concernant le problème du corps et de l'esprit.

En particulier je vais montrer que, contrairement à une idée *très largement* répandue aussi bien chez les philosophes, les physiciens que chez l'homme de la rue, le mécanisme est incompatible avec le matérialisme.

Je vais démontrer que le mécanisme est incompatible avec le monisme matérialiste selon lequel il existe exclusivement un univers substantiel descriptible en principe entièrement en termes physiques. Je vais incidemment démontrer aussi que le mécanisme est incompatible avec le dualisme selon lequel il existe à la fois un univers concret (décrit par les lois de la physique), *et* un univers spirituel. En fait le dualisme peut être vu comme un double matérialisme puisque le dualisme tente de substantialiser aussi bien le corps-matière que l'âme-esprit.

Je vais donc démontrer que le mécanisme nécessite un idéalisme moniste incompatible avec toute forme de substantialisme. Cette démonstration ne va pas résoudre le problème du corps et de l'esprit, mais va conduire vers une nouvelle formulation du problème. En effet, avec l'hypothèse du computationnalisme, le problème du corps et de l'esprit va se transformer nécessairement en la recherche des dérivations

1. d'une phénoménologie de l'esprit —capable d'expliquer l'origine et la nature des savoirs et des croyances, et

2. d'une phénoménologie de la matière, capable d'expliquer l'origine et la nature de nos observations et de nos croyances physiques.

Le point 1. peut difficilement être considéré comme original. Avec le computationnalisme, la psychologie est trivialement réductible *en principe* à l'informatique théorique. Ce qui est original, c'est de démontrer que pour

résoudre le problème du corps et de l'esprit, on est obligé de dériver la phénoménologie de la matière à partir de la phénoménologie de l'esprit. Cela fait de la physique une branche, *en principe*, de la psychologie.

C'est précisément l'inverse des tentatives de réduction habituelle où l'on essaye de rendre compte des qualités psychologiques à partir de la matérialité cérébrale, corporelle ou même cosmique ou universelle.

Au contraire, le mécanisme nécessite un psychologisme éliminant toute ontologie substantielle plutôt qu'un physicalisme éliminant l'ontologie spirituelle.

Le mécanisme nécessite donc de faire de la physique une branche de la psychologie, elle-même branche de l'informatique théorique, elle-même branche de la théorie des nombres. Le terme "branche" est utilisé dans un sens un peu plus général que d'habitude, cela se précisera de soi-même au fur et à mesure de la démonstration.

Un logicien attentif constatera que la matière n'est pas *logiquement* éliminée. Mais elle va perdre *tout* pouvoir explicatif y compris pour rendre compte aussi bien des sensations physiques que de la science physique.

On remarquera une certaine ironie dans cette situation. En effet, le mécanisme est en général évoqué par des matérialistes réductionnistes (si pas éliminativistes) pour désubstantialiser l'esprit et contrer le dualisme ou d'autres spiritualismes. Et cela marche en effet, mais si on y regarde de plus près (ce qui est proposé ici) la désubstantialisation ne peut pas s'arrêter à l'esprit, mais s'étend sur le corps, la matière et l'univers.

Le travail n'est pas spéculatif. Il s'agit bien d'une démonstration, ou d'une argumentation hypothético-déductive: SI le mécanisme est correct ALORS la physique *doit* être réduite à la psychologie. Je précise ce point au chapitre 2.

Remarque méthodologique Afin d'aider le lecteur à ne pas perdre le fil de la démonstration, j'ai décidé d'être le plus court possible. La démonstration, qui commence au chapitre 3, est terminée à la fin du chapitre 4. Elle ne nécessite aucune connaissance particulière, à l'exception d'une connaissance passive de la thèse de Church et, bien sûr, du minimum de bagage en philosophie classique, comme celle requise dans l'enseignement secondaire supérieur en France (de bons manuels sont (Huisman and Vergez, 1966), ou (Nagel, 1987)). L'annexe D fournit une introduction au problème du corps et de l'esprit ainsi que quelques précisions supplémentaires sur la notion de niveau adéquat de mécanisme.

Le chapitre 1, qui énonce de façon précise les hypothèses *de tout* le travail, mentionne des points techniques qui ne sont pas utilisés dans la démonstration. Inutile de trop s'y attarder avant le chapitre 5.

Le chapitre 5 illustre la recherche d'une solution au problème du corps et de l'esprit à la lumière de la démonstration donnée. A la différence de la démonstration, cette recherche prospective nécessite un certain nombre de prérequis techniques. On peut consulter directement le rapport technique (Marchal,

1995) ou les annexes en partie tirées de ce rapport, ou certains ouvrages particulièrement pertinents comme (Boolos, 1993), (Webb, 1980), ainsi que (Albert, 1992) et (Maudlin, 1994) pour la physique.

Chapitre 1

L'hypothèse du computationnalisme

L'hypothèse philosophique que j'appelle "computationnalisme", ou plus simplement "mécanisme", est la conjonction des trois hypothèses suivantes :

- Le mécanisme digital et indexical
- La thèse de Church
- Le réalisme arithmétique

1) L'hypothèse du *mécanisme digital et indexical* est l'hypothèse selon laquelle *je* (aspect indexical) peux survivre, non seulement avec un coeur artificiel, un rein artificiel, etc., mais aussi avec un cerveau artificiel, en l'occurrence constitué d'une machine universelle digitale, c'est-à-dire un ordinateur, convenablement "programmé" à partir d'une description d'un état instantané du cerveau saisi à un niveau adéquat (aspect digital).

La finitude mise à part, je ne place *aucune* restriction sur le niveau de description du cerveau: l'hypothèse du mécanisme digital reste correcte dans le cas où il s'avérerait qu'il faille décrire l'état quantique de tout l'univers pour décrire adéquatement l'état du cerveau. En ce sens l'hypothèse peut être considérée comme particulièrement faible (au sens logique). J'appellerai quelque fois "cerveau généralisé" la partie de l'univers apparent qu'il faut reproduire pour survivre à une reconstitution.

Le mécanisme tel qu'on l'aborde ici, en particulier l'aspect indexical, présuppose un minimum de "psychologie populaire". On reconnaît la présence d'une conscience privée chez l'autre. En particulier on admet qu'une expression du genre "Paul estime avoir survécu à la greffe de coeur" puisse avoir un sens.

2) La *thèse de Church*, énoncée la première fois par Post dans les années 20 et indépendamment par Church et Turing une bonne décennie plus tard, est

une hypothèse plus technique apparue dans le fondement des mathématiques (Post, 1922; Church, 1936; Turing, 1936). La thèse de Church, ou plutôt une version anachronique mathématiquement équivalente, dit que toute fonction (intuitivement) calculable est programmable (calculable par un ordinateur). Une version intensionnelle (mathématiquement équivalente aussi, voir annexe B) de la thèse de Church rend possible l'existence d'une numérotation de toutes les (descriptions des) machines digitales possibles: $\{M_1, M_2, M_3, ...\}$. Par définition ces machines disposent d'autant de temps et d'espace mémoire qu'il est nécessaire pour mener à bien leur activité.

C'est la thèse de Church qui confère à la notion de machine universelle (ordinateur abstrait) un statut extrêmement général, quoique non trivial. En effet, avec la thèse de Church on peut démontrer que toute machine universelle consistante est obligatoirement silencieuse sur certaines questions, notamment la question de savoir *quelle* machine elle est (Benacerraf, 1967). La thèse de Church protège le mécanisme digital des réfutations Gödéliennes comme celles de Lucas et de Penrose (Lucas, 1961; Penrose, 1989). Cela est examiné au chapitre 5. La thèse de Church protège plus généralement le mécanisme contre les réductionnismes formels (Webb, 1980; Marchal, 1995; Wang, 1974). Voir annexe B.

3) L'hypothèse du *réalisme arithmétique* consiste à admettre que la vérité des propositions arithmétiques est indépendante de moi (de vous, de l'humanité, de l'univers, ...). On admet par exemple que la proposition "il n'existe pas de plus grand nombre premier" est absolument vraie, que la vie soit ou non apparue sur la terre. Avec la thèse de Church, le réalisme arithmétique permet d'affirmer que toute machine sur toute donnée va s'arrêter ou, ... ne va jamais s'arrêter. Le réalisme arithmétique définit l'ontologie non substantielle acceptée ici. Depuis le bien connu travail de Gödel 1931 (bien que cela se trouve déjà chez Post 1922), on sait qu'il n'existe pas de théorie complète et décidable capable d'axiomatiser l'ensemble des vérités arithmétiques (Gödel, 1931; Post, 1922). Heureusement, car si ce n'était pas le cas, le mécanisme pourrait constituer un réductionnisme formel, en contradiction avec la thèse de Church.

Chapitre 2

Comment fonctionne la démonstration ?

L'argument du déployeur universel permet de démontrer assez rapidement le résultat principal, à savoir que le computationnalisme nécessite une phénoménologie de la matière :

$$COMP \Rightarrow PhMat$$

Malheureusement l'argument utilise en dernier recours une hypothèse que l'on peut juger *extravagante* HE, si bien que l'argument du déployeur universel montre seulement:

$$COMP + HE \Rightarrow PhMat$$

L'argument du graphe filmé sert essentiellement à éliminer l'hypothèse extravagante. Je procède ainsi afin de modulariser les difficultés. L'argument du graphe filmé apporte des informations complémentaires cependant, et d'une certaine façon récapitule l'argumentation du déployeur.

Pour le besoin de la démonstration et la facilité des expériences par la pensée, je vais supposer que le niveau de substitution se situe au niveau du cerveau, par exemple au niveau de la constitution biochimique du cerveau (concentration locale des ions et des molécules).

Avec le déployeur universel nous verrons que cette hypothèse n'est en rien limitative. La démonstration que je propose reste en effet valide quel que soit le niveau de substitution exigé, fut-ce l'état quantique de l'univers, pourvu que cet état soit calculé par une fonction partielle calculable. Dans le cas contraire le computationnalisme est faux, et nous sortons du cadre de notre hypothèse. J'appellerai "cerveau généralisé" la portion d'univers qu'il faut dupliquer pour me reconstituer —autrement dit la portion d'univers (finiment descriptible par hypothèse mécaniste) nécessaire pour véhiculer mon expérience privée. On peut consulter l'annexe D pour plus de détails.

Pour une explication détaillée sur les types d'argumentations philosophiques (déductives et inductives) et de leurs types d'expériences par la pensée afférents, on peut consulter (Brown, 1991). En gros, une argumentation déductive (à valeur démonstrative) est une argumentation telle que ceux qui admettent (momentanément pour le raisonnement) les hypothèses et n'admettent pas les conclusions sont tenus de trouver une erreur dans l'argumentation (par exemple sous la forme d'une hypothèse manquante ou d'une déduction non valide, etc.). La seule façon de critiquer une argumentation (philosophique, scientifique) de nature inductive, est de proposer une "meilleure" argumentation. Si les chapitres 3 et 4 constituent une argumentation déductive, le chapitre 5 peut être considéré comme une argumentation inductive en faveur du computationnalisme.

Je voudrais aussi insister sur le fait que la frontière entre la science et la philosophie est vague et relative. Des hypothèses "philosophiques" comme le sont la thèse de Church, ou le principe de réalité locale d'Einstein, peuvent avoir des conséquences "scientifiques" vérifiables (confirmables, réfutables,...). Par exemple je montre dans les annexes que la thèse de Church entraîne l'incomplétude gödélienne (annexe B), et je montre que des propositions d'Einstein (longtemps jugée *philosophiques*) ont pu être réfutées expérimentalement (annexe C). A ce sujet, on peut dire que, grâce au travail de 1964 de Bell (voir (Bell, 1964) contenu dans (Bell, 1987)), un véritable champ de "philosophie expérimentale" est apparu.

Dans le même esprit, je démontre dans ce travail que l'hypothèse "philosophique" du computationnalisme entraîne des conséquences concrètes et testables.

La science et la philosophie sont inextricablement liées. Elles résultent chacune en grande partie, du dialogue entre ceux qui sincèrement posent la question "vois-tu ce que je vois?" et ceux qui sincèrement posent la question "crois-tu ce que je crois?". Un scientifique qui prétend ne pas faire de philosophie, est, dans le meilleur des cas un philosophe positiviste (malgré lui), et, dans le pire des cas un philosophe incapable de remettre ses hypothèses philosophiques (souvent héritées inconsciemment) en question.

Chapitre 3

L'argument du déployeur universel

Dans son petit livre "le philosophe et son scalpel" Stéphane Ferret semble ne pas craindre l'intrusion du scalpel philosophique partout dans son corps, à la notable exception du cerveau (Ferret, 1993). Le *philosophe* mécaniste au contraire, ne privilégie aucune de ses parties, cerveau compris, si bien, que sur son lit d'hôpital avec une tumeur au cerveau, lorsque son médecin lui apprend qu'il n'en a plus que pour une semaine de vie sauf, peut-être, s'il accepte une greffe d'un cerveau digital, il se demande : "Pourquoi pas ?" L'hypothèse du mécanisme est que cette greffe est en principe possible à *un certain niveau*. Je propose alors l'expérience par la pensée consistant à se mettre à la place du philosophe subissant à l'hôpital une greffe effectuée, par hypothèse encore, au bon niveau. Comme l'opération se fait sous anesthésie, l'hypothèse du mécanisme rend cette expérience équivalente à celle d'un quelconque séjour à l'hôpital. Après un mois de convalescence, le philosophe remercie son médecin prétendant qu'il lui a sauvé la vie, et rentre chez lui vaquer à ses occupations habituelles.

Le point capital, à présent, est que si on a accepté le scénario précédent, autrement dit si on accepte l'hypothèse computationnelle, on est forcé d'accepter la possibilité du scénario suivant.

A l'hôpital on découvre que le cerveau était sain. Ne disposant pas des informations adéquates (suite aux négligences du département informatique de l'hôpital) et croyant bien faire, une équipe de chirurgiens reconstituent un corps au philosophe. Grâce aux renseignements génétiques extraits de cellules du cerveau, le nouveau corps artificiel est semblable à son corps naturel. Ainsi, deux mois après son entrée, le philosophe rentre chez lui, une fois de plus (?), avec son cerveau original, mais avec un corps artificiel.

Appelons P_1 le philosophe qui est rentré après un mois. Son corps est naturel, son cerveau est artificiel. Appelons de même P_2 le philosophe qui est rentré après deux mois. Son corps est artificiel, son cerveau est naturel.

Appelons P le philosophe avant son hospitalisation.

Avec l'hypothèse computationnelle et parce que nous supposons le niveau de substitution adéquat, ni P_1, ni P_2 ne peuvent se douter de la présence de leur *doppelgänger*[1] avant leur rencontre. Après leur rencontre tout deux ont raison de revendiquer un statut "original". Nozick propose pour résoudre le problème de l'identité personnelle une théorie dite du continuateur "le plus proche" (Nozick, 1981). Mais un tel continuateur est ambigu : P_1 est plus proche de P dans le temps, il dispose aussi de son corps biologique original, mais P_2 dispose du cerveau original. En fait, avec le mécanisme il n'existe pas de critère objectif (communicable) capable de définir un continuateur plus proche univoque. En effet, si un tel critère existait, on pourrait l'utiliser pour dupliquer ce continuateur. Par construction il perdrait son univocité.

On peut tirer trois leçons de cette expérience:

1. Ce qu'on vient d'illustrer : le mécanisme nous rend, à la façon des unicellulaires, essentiellement duplicables. Il s'agit d'un important possible *effet secondaire de la greffe artificielle*.

2. Le computationnalisme est une affaire d'opinion personnelle : il est immoral de vous y contraindre et il *pourrait* être immoral de vous l'interdire.

3. Le mécanisme entraîne une forme forte d'indéterminisme.

On comprend le point 2 en se plaçant à la place de P_2. Il estime avoir survécu grâce à la négligence du département informatique de l'hôpital. En effet, si celle-ci ne s'était pas produite, son cerveau original aurait été détruit et il estime *rétrospectivement* qu'il aurait été tué puisqu'il voit bien à présent que celui disposant d'un cerveau artificiel *est un autre*, une sorte de jumeau imposteur qui tente de lui voler sa place. P aurait-il eu la moindre méfiance vis-à-vis du mécanisme, que P_2 aurait ressenti cette suspicion croître en lui. Si P bénéficie d'assez de dispositions introspectives, P_1 peut comprendre qu'en aucune façon il ne peut convaincre son double P_2 d'avoir, lui P, survécu, en P_1, à cette expérience. Le computationnalisme, s'il est correct, n'est donc pas démontrable (bien qu'on verra qu'il peut être réfuté). Du coup, il est criminel de contraindre une personne à une greffe de cerveau artificiel, car cela *pourrait être* un meurtre. De même, si la technologie rend une telle greffe possible, il pourrait, en l'absence de réfutation du mécanisme, être criminel de l'interdire à celui qui la demande car cela aussi *pourrait être* un meurtre par euthanasie passive. Un computationnalisme bien compris, je veux dire consistant, force ainsi le respect de la position non-computationnaliste.

Le point 3 est capital pour le besoin de notre démonstration et fait l'objet de la section suivante.

[1]Le terme allemand "doppelgänger" désigne le double ou le sosie. Dans la psychiatrie anglo-saxonne il désigne aussi l'expérience de rencontrer son double (Gregory, 1987).

3.1 Le mécanisme entraîne une forme forte d'indéterminisme phénoménal

L'indéterminisme découle de la possibilité de se multiplier et de la non-univocité du continuateur le plus proche. Le domaine sur lequel porte cet indéterminisme est donné par la collection des continuateurs les plus proches. Pour bien comprendre ce point il est pédagogiquement opportun de partir d'une duplication a priori symétrique, où on ne dispose clairement d'aucun critère pour sélectionner un continuateur univoque. Nous verrons plus tard que la symétrie ne joue aucun rôle.

Le computationnalisme rend possible le télétransport, et le "mécaniste pratiquant" est un adepte de ce mode de locomotion. Se téléporter de Bruxelles à Moscou (par exemple) revient à se faire annihiler à Bruxelles et à se faire reconstituer à Moscou à partir d'une description digitale de l'état instantané du corps avant l'annihilation. Une telle expérience est équivalente à une greffe de corps et de cerveau. Avec le mécanisme, la probabilité de survivre à Moscou à cette expérience de télétransport est égale à 1. Il ne s'agit pas d'une certitude absolue, mais bien d'une certitude relativisée à l'hypothèse computationnelle et à l'adéquation du niveau choisi pour la substitution.

A présent, si on survit, *en vertu du mécanisme*, à un télétransport, on survivra à une duplication symétrique où, par exemple, on est reconstitué simultanément à Moscou et à Washington (par exemple). Cela est dû au fait que la duplication a été effectuée au niveau adéquat et que les reconstitutions sont indépendantes. En particulier cela entraîne que les deux reconstitués auront des expériences a priori indépendantes, de la même façon que P_1 et P_2.

J'utiliserai les termes "subjectif", "phénoménal", "privé", ou encore l'expression "de la première personne" de façon synonyme. Ces termes se diront d'une expérience telle qu'un individu la ressent. Par exemple si Untel se plaint d'une rage de dent, son discours est de la première personne. Par opposition, un discours communicable "de, ou à, la troisième personne" est dit d'un discours objectif. Par exemple lorsque le dentiste de Untel montre des photographies de la bouche de Untell à un congrès international, son discours, en l'occurrence non verbal, participe de la troisième personne.

Le problème du corps et de l'esprit consiste en grande partie à isoler un cadre ontologique minimal capable de justifier les discours correctes ou possibles des premières et troisièmes personnes (voir annexe D). On a :

Proposition 1 *Le mécanisme entraîne un indéterminisme "de la première personne"*

Preuve. Considérons un individu *Jules* qui admet l'hypothèse computationnaliste, mais qui ne croit pas à l'indéterminisme. Je vais montrer que Jules est contraint soit à confondre la première et la troisième personne, soit

à octroyer arbitrairement un statut "original" à une des ses copies futures. Posons à Jules la question suivante:

> Où va-tu *phénoménalement* te retrouver après l'expérience d'annihilation à Bruxelles et de reconstitutions simultanées à Washington et à Moscou?

Comme les reconstitutions sont indépendantes, et que Jules est mécaniste (et donc croit survivre au télétransport) Jules ne peut pas répondre qu'il ne se retrouvera *ni* à Washington, *ni* à Moscou. Comme on suppose que la duplication est faite au niveau adéquat, Jules, qui est mécaniste, ne peut pas répondre qu'il se retrouvera à Washington *et* à Moscou. Il peut bien sûr dire correctement (avec le mécanisme): "Vous me verrez à Washington *et* à Moscou". Mais cette réponse n'est pas satisfaisante car Jules y parle de lui à la troisième personne, et la question porte sur sa future expérience *phénoménale*, à la première personne. La duplication, *au niveau adéquat*, ne permet pas de se sentir à deux endroits à la fois. Jules, mécaniste et déterministe, est donc obligé de choisir sa réponse parmi "Washington" et "Moscou". Soit il reconnaît que son choix est arbitraire. Dans ce cas il n'est pas difficile de se convaincre que s'il ne confond pas la première et la troisième personne, c'est qu'il octroie un statut "original" arbitraire à une de ses copies. Soit il choisit "Washington" (par exemple) prétendant que son choix n'est pas arbitraire. Il s'identifie donc *exclusivement* à celui qui va se sentir reconstitué à Washington. Dans ce cas il doit reconnaître qu'il serait tué si la reconstitution à Washington n'est pas effectuée, et donc qu'il ne survivrait pas à un télétransport Bruxelles-Moscou, et donc il doit abandonner l'hypothèse computationnelle.

3.2 La sélection ne dépend pas des délais de reconstitutions

Supposons que lors d'un télétransport de Bruxelles à Moscou, la reconstitution moscovite soit postposée d'un délai d'une heure. Je suppose que vous êtes le sujet de cette expérience. Vous est-il possible de vous apercevoir du délai? Oui, bien sûr, si vous vous êtes fait encodé à Bruxelles accompagné d'une horloge, il vous suffira de comparer votre horloge avec l'horloge de Moscou (en tenant compte du décalage horaire et du temps de reconstitution). Vous est-il possible de vous en apercevoir sans consulter d'horloge? Il n'est pas difficile de se convaincre que cela est impossible puisque pendant le délai rien ne vous permet de mesurer le temps écoulé: en fait la reconstitution rétablira la perception subjective du temps lors de votre encodage. Et ceci reste vrai que le délai soit d'une heure ou d'un googol d'heures (10^{100} heures).

De ceci il résulte que, quelle que soit la façon dont on tente de mesurer ou de quantifier ou d'évaluer la chance de se retrouver à Moscou ou à Washington

dans une expérience de multiplication de soi, cette façon donnera des résultats invariants selon qu'on ajoute ou non des délais aux reconstitutions, en l'occurrence à Washington ou à Moscou.

Définition J'appelle ***relation de sélection***, ou simplement *sélection* une telle façon de quantifier l'indéterminisme mécaniste pour une expérience d'automultiplication de soi.

Cette relation de sélection est une inconnue. Elle reste inconnue au cours de ce travail, même si on met en évidence des propriétés qualitatives de cette relation. La plupart de ces propriétés qualitatives sont décrites par des résultats d'invariance. Par exemple, on vient de montrer que la relation de sélection est invariante pour la position dans le temps (et l'espace) de la reconstitution.

Notre but n'est pas ici de calculer ni même seulement de définir de façon précise cette relation de sélection. Notre but est seulement de démontrer qu'avec l'hypothèse du mécanisme, il est impossible de résoudre le problème du corps et de l'esprit sans parvenir à justifier les lois de la physique ou les croyances en les lois de la physique, exclusivement à partir de cette relation.

3.3 Le mécanisme entraîne une forme forte de non-localité phénoménale

Supposons que vous vous téléportiez de Bruxelles à Moscou. Supposons que l'annihilateur ne fonctionne pas. Et supposons que vous soyez prévenu à l'avance de ce qu'il ne fonctionne pas. Et supposons que la relation de sélection soit définie par une distribution de probabilité uniforme placée sur l'ensemble des reconstitutions (ce qui semble, au moins, raisonnable lorsque les reconstitutions sont distinctes et en nombres finis). Que vaut dans ce cas la probabilité de se sentir être reconstitué à Moscou ?

Réponse: *dans ce cas* la probabilité vaut 1/2. Ne pas annihiler l'original (vu de la troisième personne) revient à transformer une expérience de téléportation en expérience d'autoduplication (avec délai).

Supposons à présent que vous décidiez d'aller de Bruxelles à Moscou par voie de chemin de fer. Appelons X une description de l'état de votre cerveau à Bruxelles au moment précis où vous approchez du train pour Moscou, et supposons qu'à un moment ou à un autre un accident cosmique lointain reconstitue votre cerveau, dans l'état X, dans un environnement viable. Appelons cet environnement *Isbrahül*, pour fixer les idées. Et supposons encore que vous soyez prévenu à l'avance de l'existence de cette reconstitution lointaine. Et supposons encore que la relation de sélection soit donnée par une distribution de probabilité définie sur l'ensemble des reconstitutions. Que vaut dans ce cas, à Bruxelles, la probabilité de vous retrouver à Moscou (ou à Isbrahül) ?

On aimerait répondre 1/2, mais on réalise avec cet accident lointain qu'une supposition de plus est nécessaire. Il faudrait qu'on puisse vous prévenir à l'avance non seulement de la reconstitution lointaine, mais aussi de l'inexistence de reconstitution lointaine de l'état X du cerveau dans tout autre environnement viable dans tout lieu et pendant toute l'histoire de l'univers.

Ceci montre qu'avec le mécanisme, la sélection est nécessairement non-locale même pour une expérience aussi peu "science-fictionnesque" que l'usage d'un chemin fer. Avec le mécanisme, en toute circonstance, pour prédire les chances de ses propres futurs discours à la première personne, il ne suffit pas de garantir une collection d'événements (physiques ou cérébraux par exemple) dans un domaine précis, il faut encore connaître l'existence de non-événements partout dans l'univers, ou si on préfère, il faut être assuré de l'inexistence de certains événements partout dans l'univers.

3.4 La sélection ne dépend pas de la nature des reconstitutions

On trouve aujourd'hui sur le marché des logiciels et des matériels permettant à un utilisateur de s'interfacer à un ordinateur de façon qu'il ait, partiellement et à des degrés divers, l'impression de visiter une réalité artificielle (appelée aussi virtuelle). Il peut par exemple poursuivre des monstres dans un labyrinthe ou apprendre à piloter un hélicoptère.

Voici un autre résultat d'invariance qualitative de la relation de sélection: celle-ci ne dépend pas de la nature réelle ou virtuelle de la reconstitution. Cela découle immédiatement du fait que si le cerveau est digitalisable à un certain niveau n (ce qui est assuré par l'hypothèse computationnelle), alors il est possible de digitaliser les entrées sensorielles au niveau n. Dans ce cas le sujet d'une telle reconstitution ne peut pas remarquer la différence de nature, virtuelle ou réelle de son environnement. Ici, la machine universelle fait office de malin génie cartésien.

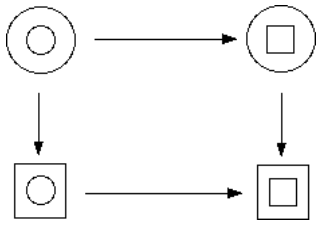


Figure 3.1: Deux chemins pour *un grand plongeon*

Si un tel sujet dispose au départ d'un cerveau artificiel, il se retrouvera momentanément "plongé" dans un environnement exclusivement digital en ce sens qu'aussi bien sa personne que son environnement local résulte de l'activité

computationnelle d'une machine universelle. Dans la figure 3.1, les carrés représentent des artefacts, les ronds des choses naturelles, les petites figures représentent des cerveaux, les grandes figures représentent des environnements. La figure 3.1 illustre le fait qu'un tel plongement peut s'effectuer de deux façons selon qu'on dispose ou non, au départ d'un corps-cerveau artificiel. Le résultat facile mais capital pour la suite est le suivant : quelle que soit la façon dont on quantifie l'indéterminisme mécaniste pour une expérience d'automultiplication, c'est-à-dire : quelle que soit la relation de sélection, le résultat ne peut pas dépendre de la nature réelle ou virtuelle des environnements des éventuelles reconstitutions. En particulier si on admet $P(Washington) = P(Moscou) = 1/2$ pour une expérience de duplication concrète (avec reconstitution "réelle") alors $P(Washington) = P(Moscou) = 1/2$ pour une expérience de duplication avec deux reconstitutions virtuelles.

Il en est de même pour les reconstitutions, avec ou sans délais, hybrides. Par exemple, si on admet l'existence d'une distribution de probabilité uniforme sur un ensemble fini de n reconstitutions distinctes R_i d'un certain état computationnel x, alors la probabilité de se sentir être reconstitué en $R_k = 1/n$ quelle que soit la nature, réelle ou virtuelle, de *chacune* des reconstitutions R_i.

3.5 L'argument du déployeur universel

Définition Un déployeur universel est un programme qui exécute tous les programmes possibles dans tous les langages de programmation possibles.

L'existence *arithmétique* d'un tel programme découle du réalisme arithmétique. Il est facile (pour ceux qui apprécie "tâter la bécanne") d'écrire un programme A en FORTRAN capable de générer et numéroter tous les programmes FORTRAN par ordre de longueur (et par ordre alphabétique ceux de mêmes longueurs): $F_1, F_2, F_3, ...$ Je me limite aux programmes sans entrées. Il est facile d'écrire un programme B capable de simuler un interprète FORTRAN sur un nombre fini d'étapes. Il ne suffit plus alors que d'écrire une routine DU qui va zigzaguer sur toutes les exécutions finies de tous les programmes. Par exemple en appliquant B (destructivement) sur les programmes, générés avec A, dans l'ordre suivant 1, 1, 2, 1, 2, 3, 1, 2, 3, 4, 1, etc.

Il est facile de démontrer que, malgré les limitations apparentes, le déployeur DU se déploie sur tous les programmes, qu'il soit écrit en FORTRAN, LISP, JEU-DE-LA-VIE, etc, et ça qu'il soit à 0 entrée, 1 entrée, n entrées, entrées infinies et générables, etc. Par exemple parmi les programmes FORTRAN à 0 entrée figure un simulateur d'un déployeur des programmes LISP ayant pour entrée les segments initiaux des nombres réels. DU, notre déployeur universel concret se déploie sur tous les programmes possibles, interfacés de toutes les façons possibles, dans tous les environnements digitalisables possibles.

Qu'un tel déployeur *DU* se déploie sur *tous* les langages de programmation *possibles* découle de la thèse de Church.

3.5.1 L'hypothèse extravagante

L'hypothèse extravagante HE affirme qu'un déployeur universel est concrètement et *intégralement* exécuté dans l'univers. Notons que cette hypothèse repose elle-même sur une hypothèse *apparemment* bénigne HU selon laquelle il existe un univers concret.

Un déployeur universel a été exécuté pendant un moment à Bruxelles dans le courant de l'année 1991. Cela n'a rien d'extravagant, le propos était illustratif. Le caractère extravagant de l'hypothèse HE est dû à l'exigence que le déployeur universel soit *intégralement* exécuté. Comme il s'agit d'une exécution infinie, cela nécessite que l'univers concret soit lui-même infini dans le temps et l'espace et suffisamment robuste dans ses relations causales et historiques.

Remarques

1. L'hypothèse extravagante élimine les objections "techniques". Même s'il est impossible de capturer l'état computationnel d'un être humain, le déployeur va générer un jour ou l'autre cet état. Et c'est tout ce dont on aura besoin pour la démonstration.

2. De la même façon on peut montrer que l'hypothèse extravagante élimine des objections plus conceptuelles comme l'évocation de la mécanique quantique pour réfuter la possibilité de la duplication quantique sans annihilation de l'original (cette impossibilité repose sur un travail de Bennett, voir (Yam, 1993)). Le déployeur universel se déploie, en effet, sur tous les états quantiques calculables. Et si l'hypothèse computationnelle est neutre au sujet du caractère calculable de l'état quantique de l'univers, elle exige que votre propre état (peut-être quantique) puisse être le produit d'un calcul.

Bien sûr l'hypothèse HE est, ou en tout cas peut raisonnablement être considérée comme, extravagante et devra être éliminée à son tour. C'est l'objet de l'argument du graphe filmé esquissé à la section suivante. Nous verrons que le graphe filmé éliminine HE, mais éliminine aussi HU par la même occasion. Toutefois la grosse part de l'élimination de HU résulte déjà de l'argument du déployeur universel.

3.5.2 Une phénoménologie de la matière est nécessaire

A présent on peut démontrer la version affaiblie (par HE) du résultat principal du travail:

Théorème 2 $COMP + HE \Longrightarrow PhMat$

En français: l'hypothèse computationnelle accompagnée de l'hypothèse extravagante nécessite d'extraire une phénoménologie de la matière sans recourir à l'existence de l'univers ou d'un univers. L'hypothèse de l'existence d'un univers (HU) est nécessaire *seulement* pour justifier l'existence de la simulation initiale du déployeur universel. Cependant, étant donné l'existence de cette simulation, je vais démontrer qu'avec le mécanisme, il n'est pas possible d'utiliser l'hypothèse de l'existence de l'univers pour justifier les croyances et les connaissances physiques. Parmi ces dernières figurent aussi bien les sensations physiques que les lois de la physique.

Preuve: considérons une expérience de physique élémentaire, par exemple l'expérience consistant à lâcher une craie et à la suivre des yeux. Supposons pour fixer les idées que vous soyez l'auteur de cette expérience. Quelle est la probabilité que vous voyiez la craie tomber? Notons bien que je ne vous demande pas d'évaluer la chance, à la troisième personne, que la craie tombe, je vous demande explicitement d'évaluer la chance que *vous la voyiez tomber*. La question est donc posée à la première personne, en l'occurrence vous.

La méthode traditionnelle pour répondre à cette question, ou à des questions similaires repose sur l'inférence inductive. Chaque fois que vous avez lâché un objet, vous l'avez vu tomber. Par inférence inductive vous évaluez la probabilité à 1. Une méthode plus sophistiquée consiste à utiliser les lois de la physique. Notons que ces dernières ont aussi été inférées inductivement. Comme la craie est un objet macroscopique la physique classique s'impose. La physique classique permet de dériver la proposition selon laquelle tout corps suffisamment proche de la terre tombe sur la terre avec une probabilité 1 (la physique classique est déterministe: idéalement toutes les probabilités valent 1).

Nous savons cependant qu'avec l'hypothèse computationnelle, par indéterminisme et non localité mécaniste, la probabilité sera proche de 1 seulement si la probabilité d'un accident cosmique —reconstituant l'état de votre cerveau dans un environnement virtuel où la craie ne tombe pas— est proche de 0. Nous pourrions en déduire avec le computationnalisme que de tels événements accidentels cosmiques lointains sont effectivement rares. C'est ici qu'intervient l'hypothèse extravagante. En effet, quel que soit le niveau de substitution du mécanisme, celui-ci est finiment descriptible (par hypothèse du mécanisme digitale). Or l'état de conscience (de la première personne donc) correspondant à l'expérience consistant à voir la craie ne pas tomber existe logiquement et correspond donc à un état computationnel possible. Or le déployeur universel génère tous les états computationnellement possibles. Le fait qu'il génère ces états de façon lente et disparate dans l'histoire de l'univers concret n'a aucune influence sur la relation de sélection puisque celle-ci est invariante pour les

délais et les lieux de reconstitution. De même, le fait qu'il s'agit de reconstitutions virtuelles n'entre pas en ligne de compte comme on l'a justifié plus haut. Donc le domaine de reconstitution relatif à votre expérience du lâcher de la craie contient un nombre gigantesque, au moins, de reconstitutions aberrantes (où, par exemple, les craies ne tombent pas). Le déployeur génère en quelque sorte une infinité de malins génies. Privilégier la "reconstitution physique naturelle" correspondant à votre continuation dans l'environnement physique réel (non virtuel) revient à octroyer *arbitrairement*, avec l'hypothèse mécaniste, un statut original à un de vos doppelgängers. Et cela n'est justement pas permis par le mécanisme. En particulier, supposer que notre environnement-univers réel et concret est non digitalisable ne change rien au problème car l'existence de vos reconstitutions virtuelles dans le déploiement ne repose que sur votre propre digitalité, laquelle est assurée par l'hypothèse computationnelle.

L'unique façon de conserver les probabilités 1 de la "vie de tous les jours" consiste donc à justifier l'usage des probabilités 1 pour les doppelgängers reconstitués virtuellement par le déployeur universel. Cela revient à justifier, sans invoquer de réalités substantielles, que la collection des "malins génies" virtuels ou des "accidents lointains" virtuels est négligeable ou de mesure nulle, pour une mesure définie sur les états ou sur les suites d'états computationnels apparaissant dans le déploiement. Et c'est tout ce que nous voulions démontrer.

Pour conclure cette démonstration, il n'a pas été nécessaire de définir de façon précise, ni la relation de sélection ni la mesure capable de rendre négligeable la collection des malins génies et des accidents lointains, ni de définir précisément l'espace sur lequel porte cette mesure, ni même ce qu'est exactement un état computationnel vu de la *première personne*. Ce qui a été démontré (avec HE), c'est que si le computationnalisme est correct, alors cette mesure doit exister et elle ne peut pas être définie au moyen de prédicats du genre "*réel*(x)" ou "*virtuel*(x)" ou "*A-tel-moment*(x)" ou "*A-tel-endroit*(x)", (x représentant un état computationnel de la première personne). Au contraire, nous sommes obligés de rendre compte de prédicats du genre "*réel*(x)" ou "*virtuel*(x)" "*A-tel-moment*(x)" ou "*A-tel-endroit*(x)" par la règle de sélection et sa mesure associée.

Remarques

- D'un certain point de vue la présente contribution est modeste. Non seulement le problème du corps et de l'esprit n'est pas résolu, mais je montre qu'avec le mécanisme, le problème est plus complexe qu'on ne l'imagine habituellement. Non seulement une phénoménologie de l'esprit ou de la conscience doit être apportée (ceci n'est pas original), mais surtout je démontre qu'une phénoménologie de la matière et de l'univers doit *nécessairement* accompagné la phénoménologie de la conscience. De

façon un peu négative, je montre qu'avec le mécanisme, le problème du corps et de l'esprit est *deux fois* plus compliqué.

- D'un autre point de vue la présente contribution est *littéralement* renversante. En effet, je démontre qu'avec le mécanisme, une solution au problème du corps et de l'esprit apporte nécessairement une solution à ce qu'on pourrait appeler le problème dur de la matière. Ce problème concerne la question de l'origine de la matière-univers ou de l'origine de nos croyances dans la matière-univers, ainsi que la question des relations entre cette matière apparente et les mathématiques (Einstein, 1989; Wigner, 1967b). Je démontre qu'avec le mécanisme *toute* solution au problème de la conscience nécessite une réduction de la physique à la psychologie (c'est ça qui est "renversant" par rapport aux approches fonctionnalistes et matérialistes). Le terme "psychologie" doit être pris dans un sens très général mais non trivial (et dépendant de la thèse de Church). Il doit être pris dans le sens de sciences des croyances et savoirs possibles des machines universelles. Nous verrons plus loin comment trouver des interprétations arithmétiques d'axiomatiques modales des croyances ou de la connaissance. La mesure sur le déploiement doit justifier l'existence d'histoires computationnelles partiellement partageables pour de vastes collections de machines. C'est la thèse de Church qui rend non trivial l'espace des histoires computationnelles possibles, et qui permet d'espérer isoler une notion de mesure qui soit machine-indépendante ou système formel-indépendant.

- Certains se demanderont si on n'a pas simplement démontré la négation de l'hypothèse extravagante. Ce serait un résultat assez extraordinaire puisqu'on aurait, par des considérations purement analytiques et psychologiques, démontré une propriété purement physique de l'univers, à savoir l'inexistence d'un déploiement. Mais ce mouvement sera justement interdit par l'argument du graphe filmé. Celui-ci montre que l'exécution physique du déployeur n'entre pas en ligne de compte pour poser le problème de la mesure associée à la sélection.

- D'autres estimeront peut-être que l'hypothèse computationnelle doit être abandonnée. En absence de claire contradiction, cette option est prématurée. Par contre il est correct de voir à travers la présente démonstration un argument comme quoi le mécanisme est réfutable. En effet le mécanisme pourrait être réfuté à terme de plusieurs manières:

  mathématiquement En montrant qu'il n'existe pas de mesures "sur le déploiement universel" qui satisfassent les contraintes du mécanisme.

semi-empiriquement-mathématiquement En montrant, après avoir isolé cette mesure, que la physique théorique dérivable par le mécanisme diverge de la (des) physiques théoriques dérivées (actuellement) de l'observation.

empiriquement En découvrant des divergences expérimentales entre la non-localité quantique (d'Einstein-Podolski-Rosen-Bell-Bohm-Clauser-Horn-Shimony-Holt-Aspect pour citer les principaux protagonistes, et la non-localité du mécanisme (Einstein et al., 1935; Bohm, 1951; Bell, 1964; Clauser et al., 1969; Aspect, 1976). Une formulation précise de non-localité "quantique" extraite d'une phénoménologie arithmétique de la matière est proposée plus loin.

Avant d'aborder le graphe filmé et l'élimination de HE, et l'élimination "définitive" de HU, regardons à quoi peut ressembler la phénoménologie de la matière.

3.6 A quoi peut ressembler une phénoménologie computationnelle de la matière ?

Nous savons déjà que cette phénoménologie existe nécessairement (avec l'hypothèse computationnelle) et nous savons déjà à quoi elle ressemble : une relation de sélection définie par une mesure sur les états appartenant au déploiement universel.

Il est possible cependant de détailler davantage les aspects qualitatifs de la phénoménologie.

1. La relation de sélection est conditionnelle, la mesure est définie sur des états relatifs. Par exemple dans l'expérience du lâcher de la craie à Bruxelles la mesure est définie sur des paires

$$< A \mid B... >$$

où A représente l'état à la première personne d'un continuateur computationnel le plus proche, et B représente l'état de la troisième personne à Bruxelles.

2. On a alors une correspondance entre chaque état computationnel possible et l'ensemble de ses continuateurs les plus proches (ou plutôt leurs reconstitutions virtuelles apparaissant dans le déploiement). On peut démontrer que de tels domaines de reconstitution universels relatifs sont toujours infinis. Cela peut sembler évident car on sait que le déployeur atteint tous les états accessibles une infinité de fois. Ce n'est pas évident

car on peut montrer que du point de vue d'une première personne le digitalisme entraîne qu'il ne peut exister qu'un nombre fini de continuateurs le plus proche.

3. En itérant les expériences d'automultiplication, on peut justifier l'indéterminisme mécaniste par des sondages sur la population résultante. Il est clair cependant que cet indéterminisme est purement de la première personne. Dans le cas où le domaine de reconstitution est fini, on peut argumenter en faveur de l'existence d'une distribution uniforme de probabilités sur le domaine. Il est impossible de justifier les probabilités par l'approche usant de la notion de pari. On peut cependant retrouver des justifications probabilistes par paris en multipliant, non plus des individus, mais des populations d'individus et en étudiant les sondages portant sur les sondages au sein de chaque population reconstituée. Cela permet d'introduire une curieuse "personne" située à mi-chemin entre la première et la troisième personne, et que l'on peut raisonnablement considéré comme la première personne du pluriel. *Au sein* de chaque population reconstituée, l'indéterminisme mécaniste est communicable à la troisième personne: l'indéterminisme mécaniste est communicable à la première personne *du pluriel*. Ceci permet, à partir d'une solution au problème du corps et de l'esprit, de résoudre le problème de "l'autre esprit" et cela protège a priori l'idéalisme objectif du computationnalisme de l'idéalisme subjectif du solipsisme.

4. En exploitant le point précédent, et le fait que le déployeur universel génère des populations virtuelles issues de calcul très long à partir de programmes relativement courts, on peut montrer que la normalité des états (d'esprits virtuels) relatifs est partiellement justifiée par l'existence de "petits" programmes générant une infinité non dénombrable d'histoires infinies. De telles histoires produisent des objets computationnels relativement profonds (au sens de (Bennett, 1988), voir aussi (Delahaye, 1994)). L'hypothèse mécaniste accompagnée de la thèse de Church entraîne l'existence d'une topologie non triviale sur cette collection d'histoires. Je mentionne dès à présent la ressemblance intuitive entre cette notion de normalité profonde qui émerge ici, et la notion de "quantum-depth" proposée par Deutsch pour capturer une notion épistémologique de connaissance en mécanique quantique (Deutsch, 1985).

5. Nous avons utilisé abondamment les notions d'état computationnel du point de vue de la première personne et du point de vue de la troisième personne. Pour arriver à la formulation précise (arithmétique) des phénoménologies de l'esprit et de la matière, et donc pour isoler la topologie sur les histoires computationnelles, il faudra arriver à définir de façon

purement impersonnelle, c'est-à-dire à la troisième personne, la notion de première personne. La logique de l'autoréférence, connue aussi sous le nom de *logique de la prouvabilité* suggère une voie pour procéder à une telle *désindexicalisation* de la première personne (Smoryński, 1985; Boolos, 1979; Boolos, 1993). Comme il s'agit du chemin que j'ai choisi pour chercher à isoler la formulation précise du problème du corps et de l'esprit promise par le computationnalisme, j'énoncerai dans la dernière section les résultats partiels obtenus. L'idée est d'étudier les discours stables des machines autoréférentiellement correctes. Les grandes catégories issues de la philosophie de l'esprit, y compris l'esprit et la matière, vont apparaître comme des *variations* sur les modalités de l'autoréférence.

Nous verrons alors comment la logique de la prouvabilité et ses variantes *Théétètiques* donnent des outils pour isoler la topologie et/ou la mesure des histoires infinies.

Chapitre 4

L'argument du graphe filmé

Avant d'aborder l'usage des modalités pour isoler les phénoménologies de l'esprit et de la matière, nous devons encore, pour terminer complètement la démonstration, éliminer l'utilisation de l'hypothèse extravagante.

Je vais d'abord décrire où et comment intervient le graphe filmé dans l'élimination de l'hypothèse extravagante. Ensuite je vais présenter l'argument du graphe filmé proprement dit (Marchal, 1988). J'examinerai alors quelques objections et quelques raffinements.

4.1 Les thèses de supervénience

Dès qu'on admet que les discours de la première personne peuvent avoir un sens, et en particulier que des termes comme "conscience" ont une référence, se pose le problème de trouver les termes de la troisième personne correspondants (causalement, épiphénoménalement, peu importe) à la présence de la production de la conscience.

4.1.1 La supervénience physique

En général les matérialistes (non éliminativiste) attribuent la production de la conscience d'un certain sujet X à l'activité physique d'un cerveau (ou d'un corps ou d'un univers, ou, avec le computationnalisme, d'un ordinateur adéquat). Ces matérialistes admettent ce que j'appellerai la thèse de *supervénience physique*, en abrégé SUP-PHYS. Par exemple, avec le computationnalisme, la supervénience physique, SUP-PHYS, associe

(la sensation de ma douleur) en l'espace-temps (X,T)

avec

(l'état d'un ordinateur adéquat) en l'espace-temps (X,T).

Dans chaque cas, on se réfère à des token, selon la distinction token/type introduite initialement par Peirce au sujet des mots, mais qui correspond, en philosophie de l'esprit anglo-saxonne, la distinction classique particulier/universel. Un token est donc une instantiation concrète, particulire, d'un type. Par exemple le chien particulier *Médor* instantie les types Chien, Animal, Être Vivant, etc. Il existe des token de-la-première-personne, comme "la rage de dent que j'ai ressentie l'autre jour", et des types de-la-première-personne, comme les types "états joyeux", "états douloureux", etc. "La sensation de ma douleur" dénote ici un token de la première personne. "L'état de l'ordinateur adéquat" dénote un token de la troisième personne.

On peut remplacer la sensation par un flux de sensations et l'état de l'ordinateur par l'activité de l'ordinateur sur un certain intervalle de l'espace-temps, pour obtenir des versions dynamiques correspondantes.

La thèse de supervénience physique suppose l'existence d'(au moins) un monde physique concret. Cette thèse présuppose donc l'hypothèse de l'existence de l'univers HU.

4.1.2 La supervénience computationnelle

La supervénience computationnelle (SUP-COMP) internalise *d'office* la physicalité et enlève directement toute capacité explicative à la notion d'univers concret. Elle associe en effet

la sensation de (ma douleur en l'espace-temps (X,T))

avec

un (type d') état computationnel relatif

L'espace-temps est internalisé au même titre que la douleur. Une version "plus dynamique" de la supervénience computationnelle, associe

la sensation du (flux de ma douleur en l'espace-temps (ΔX, ΔT))

avec

un type d'histoire computationnelle relative

où un "type d'histoire computationnelle relative" désigne une classe d'exécutions d'un programme (d'une machine) relativement à un ensemble d'environnements "assez similaires". Ce qui est remarquable avec la supervénience computationnelle, c'est qu'un type computationnel, ou un type d'histoire computationnelle, trimbale avec lui les "vérités contrefactuelles". La conscience n'est plus alors attachée à l'activité d'une machine concrète dans un environnement, mais au type (abstrait) d'activités possibles de cette machine dans un certain ensemble d'environnements.

4.2 L'éliminination de l'hypothèse extravagante

En internalisant la physicalité, SUP-COMP entraîne automatiquement l'élimination de HE et de HU. Avec le réalisme arithmétique et avec la thèse de Church, l'existence de la machine universelle abstraite (ou du déployeur universel abstrait) définit tous les types d'histoire computationnelle relative. Avec SUP-COMP, on n'a plus besoin de l'hypothèse de l'existence d'un univers, même pour la simulation initiale du déployeur universel. Le déployeur universel ne doit plus a priori être exécuté pour faire apparaître des "états de conscience", car ceux-ci, ne superviennent pas sur l'activité physique d'une machine. Au contraire, avec SUP-COMP et avec l'existence arithmétique du déployeur universel, c'est le concept "d'activité physique d'une machine" qui doit être associé à un ensemble abstrait d'états computationnels possibles *à la première personne du pluriel*. Cet ensemble est abstrait dans le sens qu'il est immatériel, comme tous les objets mathématiques. Avec la thèse de Church, il peut cependant être considéré comme concret dans le sens qu'il est parfaitement bien défini. Et, il existe indépendamment de nous avec le réalisme arithmétique. Bien sûr la notion de similarité entre état (histoire) computationnel(le) n'est pas clairement définie, et est en relation avec le problème de mesure posé par le déployeur. Résoudre le problème du corps et de l'esprit, et en particulier, isoler la phénoménologie de la matière, doit en grande partie, avec COMP, reposer sur l'identification de cette relation de proximité.

Pour éliminer HE (et HU), il suffit donc de démontrer que COMP entraîne SUP-COMP.

4.2.1 COMP entraîne SUP-COMP

On doit d'abord se convaincre, avec l'hypothèse computationnelle, que si la thèse de supervénience physique SUP-PHYS est fausse, on est forcé d'admettre la thèse de supervénience computationnelle SUP-COMP :

Proposition 3 *COMP* + ¬ *SUP-PHYS* ⇒ *SUP-COMP*

Cela n'est pas aussi évident que cela en a peut-être l'air. Il est en effet possible d'affaiblir considérablement la notion de supervénience physique —avec des notions de types d'état physique (et les relations de proximités éventuelles)— sans pour autant aboutir à la supervénience computationnelle. Mais *avec l'hypothèse du computationnalisme*, il est évident que ces notions de supervéniences physiques faibles sont des cas particuliers de supervéniences computationnelles, puisqu'on survit, *par hypothèse*, à la capture computationnelle de ces types physiques.

4.2.2 Le graphe filmé: COMP $\Rightarrow \neg$ SUP-PHYS

Ce qui termine la démonstration et est sans doute la chose étonnante, c'est que l'hypothèse computationnelle COMP est incompatible avec la thèse de la supervénience physique SUP-PHYS. L'argument du graphe filmé sert effectivement à démontrer la proposition suivante:

Théorème 4 *COMP* $\Rightarrow \neg$ *SUP-PHYS*

On peut consulter (Marchal, 1988) pour l'argument original. Une démonstration logiquement équivalente a été donnée indépendamment par Maudlin en 1989 (Maudlin, 1989). La démonstration de Maudlin donne cependant plus d'informations. Elle met explicitement en relief l'importance du *contrefactuel* dans le computationnel en ce qui concerne la supervénience. Pour une comparaison et traduction directe de l'argument de Maudlin avec le graphe filmé on peut consulter mon rapport technique (Marchal, 1995).

Notons que le graphe filmé et l'argument de Maudlin démontre l'incompatibilité entre le computationnalisme et la thèse de supervénience physique. Comme le computationnalisme est notre hypothèse (de travail), je propose d'abandonner la supervénience physique, ce qui permet d'éliminer l'hypothèse extravagante ainsi que l'hypothèse de l'existence d'un univers concret. Maudlin semble vouloir conserver la thèse de la supervénience physique et présente donc sa démonstration comme une réfutation (au moins partielle) du computationnalisme.

Preuve. La démonstration va se faire par l'absurde. Je vais à nouveau procéder en modularisant les difficultés. Je vais d'abord présenter l'argument original (Marchal 1988). Je vais ensuite critiquer cet argument en soulevant quelques objections. Ensuite, en m'inspirant du travail de Maudlin 1989, je vais éliminer ces objections. J'évoquerai pour terminer une objection contre l'argument de Maudlin proposée par Barnes (Barnes, 1991).

Si on admet simultanément le computationnalisme *et* la thèse de la supervénience physique on admet qu'un flux de conscience —une suite d'expériences de la première personne— est associée (causalement, épiphénoménalement, peu importe la nature de l'association) à l'activité physique —en principe descriptible à la troisième personne— d'un cerveau généralisé. Le cerveau généralisé désigne la partie finiment descriptible de l'univers qu'il est nécessaire d'émuler pour qu'une première personne estime (sur)vivre (voir annexe D). Cette partie existe avec COMP. Le cerveau généralisé peut donc être considéré comme isolé. Dans la figure 4.1 le dessin du cerveau représente un cerveau généralisé.

Pour fixer les idées, je supposerai que ce cerveau émule l'expérience correspondant à un rêve, en l'occurrence le sujet rêve qu'il est en train de voler. On peut aussi remplacer le cerveau par un simulateur de vol et supposer (ce

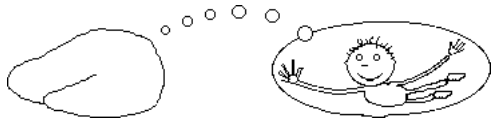


Figure 4.1: La supervénience physique

qui avec COMP ne restreint en rien la généralité de la démonstration) que le sujet a effectué un "grand plongeon". A présent, l'activité locale (sur un interval borné de l'espace-temps) d'une machine universelle ne nécessite pas une capacité de mémoire infinie. Dans la figure 4.2, le cube représente ainsi un dispositif fini entièrement digitalisable.

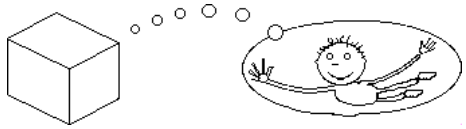


Figure 4.2: supervénience physique avec le computationnalisme

La thèse de la supervénience physique SUP-PHYS suppose l'existence d'un univers physique. Avec cette thèse, l'association entre l'expérience du flux de la conscience et l'activité physique et computationnelle du cerveau généralisé peut être pensée comme une association en temps réel ou en *espace-temps réel*. Bien sûr nous savons déjà qu'avec le computationnalisme, le sujet n'a pas de connaissance directe sur ce temps réel. En particulier, si on ralentit ou si on accélère le travail du dispositif digital, le sujet ne peut pas s'en rendre compte. Néanmoins, la thèse de supervénience physique SUP-PHYS associe des expériences de conscience *de la première personne*, comme le vertige ou le plaisir de voler, à des événements physiques comme ceux (descriptibles à la troisième personne) se produisant dans le cerveau généralisé.

A présent, avec le computationnalisme, au niveau de substitution adéquat, la nature des entités physiques et des événements physiques réalisant l'exécution digitale n'importe pas. Que le cube digitale soit émulé par une machine de Babbage, par une machine de Turing manipulée manuellement ou par un ordinateur électronique de von Neumann, ne change rien à l'expérience de la conscience du sujet. On peut donc supposer que le cube digital est réalisé par une machine de von Neumann, et en particulier une machine de von Neumann *tout à fait particulière*, comme celle que je vais décrire plus loin.

Je rappelle qu'une machine de von Neumann est constituée essentiellement d'un graphe booléen, c'est-à-dire un graphe dont les sommets sont constituées de portes électroniques réalisant les opérations logiques du ET, du OU et du NON, et dont les arêtes sont constituées de "fils électriques" (on peut consulter (Marchal, 1983) pour une illustration concrète).

Comme la nature des entités physiques et des événements physiques réalisant l'exécution digitale n'importe pas, on peut supposer que les entités substituables, comme les portes logiques ou les bus électroniques, ne sont pas conscientes —ne véhiculent pas d'expérience de la première personne. Au cas où elles seraient "accidentellement conscientes", *leur* conscience est supposée ne pas jouer de rôle pertinent pour l'expérience subjective du sujet. Dans leur réponse à l'argument de la chambre chinoise de Searle, Dennett et Hofstadter sont particulièrement clairs sur ce point (Dennett and Hofstadter, 1981). Pour être tout à fait clair, si cette conscience des composants élémentaires devait jouer un rôle dans la conscience du sujet cela signifierait, évidemment (avec COMP), que le niveau de substitution n'a pas été adéquatement choisi.

Supposons à présent que lors de la *nième* étape de l'exécution particulière d'une machine de von Neumann sur laquelle le rêve de vol supervient, une (ou plusieurs : cela ne change rien au raisonnement) porte logique soit défectueuse. A cette étape la porte logique p aurait dû envoyer une impulsion à une certaine autre porte logique q. Juste après l'étape n, je suppose que la porte logique p est immédiatement réparée. Malgré cette réparation, la machine n'est pas computationnellement équivalente à la machine de départ, et, a priori, la (ou les) défectuosité(s) entraîne(nt) à partir de l'étape n un changement dans l'expérience du rêve.

Supposons cependant que par un concours de circonstances extraordinaires, un accidentel (et heureux) rayon cosmique, par chance, vient exciter la porte logique q qui aurait dû (en l'absence de défectuosité) recevoir l'impulsion de la porte p. Le rayon cosmique supplée donc, au moment de la panne, et pour cette exécution particulière, à la défectuosité de la panne. Voir figure 4.3.

Figure 4.3: Un heureux rayon cosmique supplée

On sait, avec l'hypothèse du computationnalisme, que le sujet, au cas où on l'interroge ultérieurement, ne sera pas au courant de cette défectuosité temporaire. Si on admet que la conscience supervient sur l'activité *physique* particulière d'une machine (SUP-PHYS) *du fait* que cette machine réalise à chaque instant une suite d'états computationnels adéquats (COMP), on doit admettre que la conscience doit supervenir aussi sur l'exécution particulière de cette machine en présence de portes défectueuses, au cas où, pour une raison externe et extraordinaire, une intervention accidentelle supplée aux défectuosités.

Je dirai que la thèse de la supervénience physique entraîne la *thèse de la supervénience accidentelle active (SAC)*.

De même, supposons que lors de la *nième* étape de l'exécution particulière de la machine de von Neumann une certaine porte logique ne soit pas utilisée. Lors de cette étape cette porte logique est *physiquement inactive*. Avec la thèse de la supervénience physique on doit admettre que si la conscience supervient sur l'exécution particulière d'une machine, la conscience supervient sur l'exécution de cette machine où on retire les pièces inactives au moment où elles sont inactives pour cette exécution particulière.

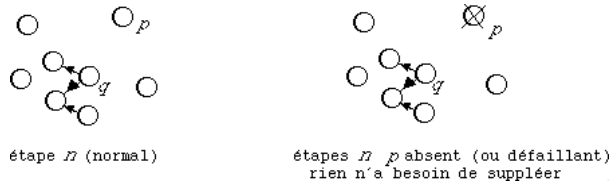


Figure 4.4: Rien ne supplée ... ni n'a besoin de suppléer

Je dirai que la thèse de la supervénience physique entraîne la *thèse de la supervénience accidentelle passive (SAP)*. Voir figure 4.4

Cette forme de supervénience accidentelle passive n'est pas utilisée dans l'argument du graphe filmé proprement dit, mais sera utilisée dans le raffinement proposé plus loin.

Venons-en à notre machine de von Neumann *très particulière*.

La machine (très particulière) de von Neumann que je considère n'est pas une machine électronique, mais est une machine où l'information est traitée optiquement. Elle est constituée de portes et de fils *optiques*. En outre, pour les besoins de la démonstration j'aimerais que le graphe booléen réalisant le réseau optique (correspondant au cube digital) soit inscriptible dans un plan. La seule difficulté technique consiste à faire croiser deux fils dans le plan en utilisant seulement les portes logiques. Ce problème a été posé par Dewdney aux lecteurs du *Scientific American*, lesquels lecteurs ont proposé diverses solutions. La figure 4.5 représente une telle solution (Dewdney, 1988).

Les modules sont, par définition, constitués des fils (les arcs) du graphe, et des connecteurs. Les fils du graphe sont des fibres optiques directionnelles et capables d'être excitées par une source extérieure perpendiculaire. Ce qu'illustre la figure 4.6.

Pour le bus on utilise des fibres bidirectionnelles. De même les modules "$\wedge$", "$\vee$", "$\neg$" sont supposés être sensibles à la lumière lorsque la direction de

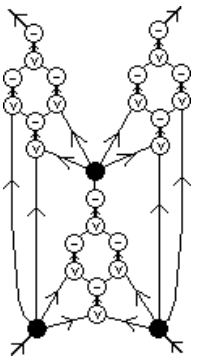


Figure 4.5: Croisement plan de Dewdney

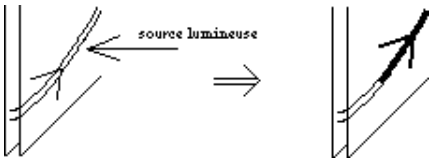


Figure 4.6: activation externe d'une connexion optique

la source est approximativement perpendiculaire au graphe.

Par exemple dans la figure 4.7, le "∧" (& dans le dessin) est activable aussi bien par une fibre optique interne au graphe que par une source lumineuse externe.

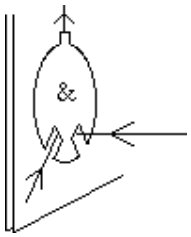


Figure 4.7: activation externe du module ET (&, ∧)

Il en est de même pour le "ou" (∨) et le "non" (¬).

Le graphe booléen est supposé être plongé dans de la fumée semi-transparente serrée entre deux vitres. De cette façon il nous est possible à la fois de contempler l'activité du graphe et d'éventuellement le perturber de l'extérieur. Macbeth est le nom que je donne à ce dispositif physique.

Le graphe booléen, Macbeth donc, est initialisé de telle façon qu'il soit computationnellement équivalent à l'état instantané d'un cerveau généralisé (au niveau supposé adéquat: qu'il soit physique, chimique, au besoin quantique), d'une personne que j'appelle Hamlet (afin de prévenir la confusion entre le cerveau généralisé, Macbeth, et la (première) personne, Hamlet, dont la conscience est associée à ce cerveau). L'initialisation s'effectue en début de phase

de sommeil paradoxal (de rêve), au niveau adéquat (existant par hypothèse).

Avec l'hypothèse du computationnalisme COMP accompagné de l'hypothèse de supervénience physique SUP-PHYS, pendant une exécution de Macbeth, Hamlet rêve. Si on parvient à capturer l'état du graphe à la fin du rêve et à l'implémenter dans le cerveau original de Hamlet, celui-ci sera à même de raconter son rêve, et son souvenir sera aussi pertinent qu'après un éveil normal, et cela quel que soit le temps réel mis par le graphe à "exécuter" l'activité du rêve.

Pendant le sommeil paradoxal le système sensoriel est inhibé (ou si on préfère: le cerveau généralisé est isolé, le sujet a commis un grand plongeon). Avec COMP et SUP-PHYS, chaque ré-initialisation et ré-exécution de Macbeth sera donc à l'origine du même rêve. On utilise bien sûr le fait que l'activité du graphe est entièrement déterministe, du point de vue d'une troisième personne. Cependant, toutes les fois qu'on interroge Hamlet au sujet de l'originalité de son rêve (soit directement si on a connecté des entrées sensorielles à Macbeth, soit par implémentation de l'état final de Macbeth dans le cerveau original de Hamlet) il répondra affirmativement qu'il a commis ce rêve pour la première fois.

Considérons à présent les expériences suivantes.

Première expérience.

La première expérience consiste à filmer Macbeth lorsqu'il exécute le rêve. On utilise une caméra capable de restituer des images très précises dans lesquelles on distingue les composants élémentaires (fils et connecteurs) du graphe booléen. Par un film-de-Macbeth j'entends (un token d') une projection sur écran du graphe filmé. Voici la première question posée:

La conscience supervient-elle sur un film-de-Macbeth ?

Il semble à première vue que répondre *oui* à cette question revienne à commettre l'erreur la plus grave qu'un philosophe puisse faire : *confondre la réalité et un dessin animé*. Dans un dessin animé un mouvement (apparent) d'un projectile (apparent) envoyé (apparemment) sur une vitre (apparente) n'est pas la cause de l'éclatement (apparent) de la vitre (apparente). Le dessin animé est "acausal", et il en est de même du film-de-Macbeth.

Deuxième expérience.

Nous enlevons un module (un "ET" par exemple) de Macbeth dans son état initial du début du rêve. Appelons cette nouvelle version de Macbeth Macbeth^{-1}. Macbeth^{-1} n'est plus computationnellement équivalent à Macbeth. Qu'à cela ne tienne, nous lançons Macbeth^{-1} et nous projetons le film de Macbeth obtenu tout à l'heure, en "temps et espace réel" sur Macbeth^{-1}.

Supposons qu'en cours du procès deux rayons lumineux transportés par les fibres internes arrivent au "ET" manquant. Celui-ci, étant absent, n'est pas à même de réaliser sa fonction, mais au même moment, parce que le film du graphe est projeté sur le graphe et qu'il y a correspondance lumineuse grâce au déterminisme objectif (à la troisième personne) du mécanisme, le film enverra de façon appropriée le signal de sortie aux connecteurs du voisinage du "ET" défaillant.

Je rappelle qu'aucun parmi ces connecteurs n'a jamais été supposé être à même de reconnaître l'origine de la source lumineuse. Le contraire signifierait que les modules sont capables de précognition, mais au niveau adéquat du computationnalisme les modules ne sont pas même supposés capables de cognition. Comme l'explique Dennett, une théorie de l'intelligence ou de la conscience, ne peut pas postuler de l'intelligence ou de la conscience au départ dans ses éléments constitutifs, et c'est là la motivation philosophique principale pour le mécanisme dans les sciences cognitives (Dennett, 1978). Le film joue le rôle, pour cette exécution particulière, de répertoire d'heureux rayons cosmiques suppléant à l'absence du module. Pour cette exécution particulière Macbeth^{-1}, accompagné du film projeté sur le graphe optique, est physiquement équivalent à l'activité de Macbeth. Si on admet le principe de supervénience physique, nous devons admettre que le rêve de Hamlet supervient sur Macbeth^{-1} couplé à la projection du film en "espace-temps" réel. Cela découle de suite de la supervénience accidentelle active (SAC).

Le raisonnement ne dépend pas du nombre de modules absents. Le même raisonnement peut être explicitement fait pour Macbeth dont on ôte d'abord deux modules, puis trois modules, etc. Si n = le nombre total de modules de Macbeth, avec le computationnalisme n est fini, et on a démontré ainsi que, si le rêve supervient sur l'activité du graphe booléen Macbeth, alors il supervient sur le film de Macbeth. Ainsi, si on admet la thèse de la supervénience physique, qui concerne des exécutions particulières de machines, on doit reconnaître que le film véhicule, en espace et temps réel les expériences subjectives correspondantes. Ce qui est absurde.

4.2.3 Objections et raffinements

Et si le film rêvait?

On pourrait se demander si on est vraiment arriver à une absurdité. Pourquoi la conscience ne pourrait-elle pas supervenir sur le film du graphe? Cela est logiquement possible en effet, mais la conscience ne peut pas *supervenir* sur le film *en vertu du computationnalisme*. En effet, aucune activité computationnelle ne peut être associée à la projection du film. Si on déroule une pellicule suffisamment grande circulairement autour d'un stroboscope envoyant de la lumière dans toutes les directions, un observateur en mouvement autour de

cette pellicule observerait le film (voir la figure 4.8a). Mais avec le mécanisme, la présence de l'observateur n'est pas pertinente. Et sans l'observateur, il n'y a plus de film "en temps réel" (voir figure 4.8b). En fait la pellicule, étant immobile, n'a aucun rôle computationnel possible et peut donc (avec la supervénience accidentelle passive) être retirée. La conscience de tous les rêves possibles devrait supervenir sur l'activité d'un stroboscope ou même sur *rien du tout* car le stroboscope lui-même n'a pas d'activité computationnelle et pourrait donc être supprimé à son tour.

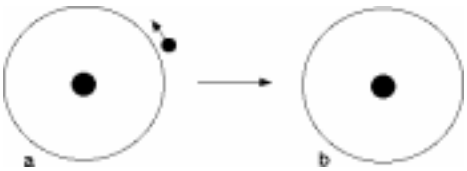
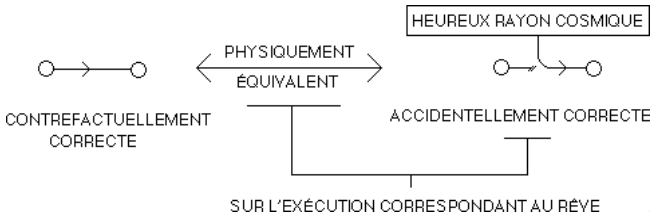


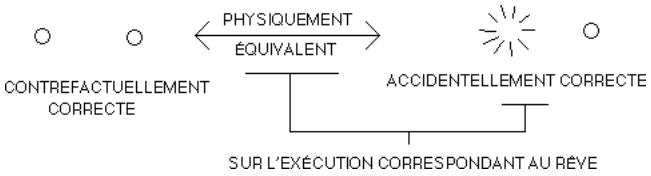
Figure 4.8: absurdité de la supervénience sur un film

Contrefactuellement correcte/accidentellement correcte

Si on admet la supervénience physique de la conscience sur l'activité computationnelle du graphe booléen Macbeth, sommes-nous forcés d'admettre la supervénience de la conscience sur l'activité physique de Macbeth^{-1}. Après tout, il y a une différence fondamentale entre Macbeth et Macbeth^{-1}. Leurs activités physiques ne sont équivalentes, en présence de l'heureux rayon cosmique ou en présence de la projection du film, que pour une exécution particulière. La moindre pertubation externe rend le rayon cosmique ou la projection du film inopérant. On peut dire que Macbeth est *contrefactuellement correct*, en ce sens que Macbeth fonctionne pour toutes les exécutions possibles alors que "Macbeth^{-1} + le film" ne fonctionne correctement que pour une exécution particulière.



Ce qui est mis en doute ici est la proposition selon laquelle la conjonction du computationnalisme et de la supervénience physique entraîne la supervénience accidentelle active. Une telle objection met en doute de la même manière la proposition selon laquelle la conjonction du computationnalisme et de la supervénience physique entraîne la supervénience accidentelle passive. La machine dont on retire une pièce inutile pour un ensemble fini d'exécutions n'est pas non plus contrefactuellement correcte.



Si on reconnaît ces objections, on admet qu'il y aurait supervénience physique de la conscience sur des processus physiques non actualisés. Ceci est raisonnable, d'autant plus qu'aujourd'hui nous savons que des résultats de mesure, en mécanique quantique, peuvent effectivement dépendre d'observations potentielles non actualisées et donc dépendre de vérités contrefactuelles. Mais dans ce cas, on ne peut plus parler de supervénience physique, qui, je le rappelle, concerne nécessairement des *token*. On est obligé de faire supervenir la conscience sur les activités *potentielles* de la machine. On doit dès lors associer la conscience à des types ou des collections d'activités physiques *possibles*. Mais alors, avec le computationnalisme, on doit, dans ce cas, associer la conscience à des *types* d'activités computationnelles possibles, donc à des ensembles d'exécutions possibles de machines universelles, mais ceci revient à remplacer la supervénience physique par la supervénience computationnelle.

On réalise l'importance de la notion de "contrefactuel" pour toute notion de supervénience computationnelle. J'ai réalisé l'importance du contrefactuel concernant la supervénience à la lecture de l'article de Maudlin 1989.

En résumé, la conscience relative à un type d'état computationnel d'une machine universelle supervient sur la collection des extensions "immédiates" possibles de ces états apparaissant dans l'entièreté du déploiement universel. Ceci résulte à la fois de l'argument du déployeur universel et du graphe filmé. Dans la recherche prospective des phénoménologies de l'esprit et de la matière, on tentera de capturer (modéliser) le "possible" par l'arithmétiquement consistent (voir aussi (Goldblatt, 1978; Goldblatt, 1993)). Il s'agit d'une forme de "darwinisme arithmétique": les sondages sur les avis des machines vont être restreints aux états des machines consistantes (survivantes aux reconstitutions). On utilise ici explicitement l'aspect *indexical* de notre hypothèse du mécanisme illustré au chapitre précédent (voir aussi annexe D).

Barnes et Malcolm

Je termine en mentionnant le travail de Barnes qui a tenté de sauver la conjonction du computationnalisme et de la supervénience physique en réfutant l'argument de Maudlin (Barnes, 1991).

Sa réfutation repose en partie sur l'usage que fait Maudlin de la supervénience accidentelle passive (SAP) et stricto sensu ne s'applique pas à l'usage de la supervénience accidentelle active (SAC) utilisée dans l'argument du graphe filmé. Mais ceci est sans doute accidentel, l'argument de Barnes devrait pouvoir être reconstruit avec l'usage que je propose de SAC.

Barnes met en évidence le rôle capital de l'histoire computationnelle ayant mené aux états instantanés capables d'être capturés par des descriptions de machines. Cependant, pour pouvoir réfuter Maudlin et garder le mécanisme avec la supervénience physique, Barnes est obligé de prêter au sujet conscient une capacité de distinguer une histoire "réelle" d'une histoire "rêvée" ou d'une histoire "virtuelle". Ce point contrarie la proposition selon laquelle un sujet, capable de survivre à une expérience de télétransport, est incapable de distinguer la réalité du virtuel, ou du rêve. En effet, nous avons vu au chapitre précédent qu'avec le mécanisme, la machine universelle joue le rôle du malin génie cartésien et empêche toute capacité de distinguer *à coup sûr* la réalité d'un rêve (même si dans certaines circonstances, le mécanisme permet au sujet de distinguer un rêve de la réalité, cf. (Marchal, 1995)).

En fait, l'argumentation de Barnes est équivalente à l'argumentation de Malcolm qui réfute aussi bien la possibilité du mécanisme que la possibilité du rêve (au sens de la psychologie populaire) en utilisant cette qualité, qu'il attribue au sujet humain, de distinguer le souvenir de la réalité du souvenir du rêve (Malcolm, 1959; Malcolm, 1968; Marchal, 1992a). Pour une réfutation computationnaliste détaillée de Malcolm on peut consulter (Marchal, 1995). Pour une critique neurophysiologique de l'argument de Malcolm, on peut consulter (Laberge, 1991), voir aussi (Dement, 1972; Dement and Kleitman, 1957; Jouvet, 1992).

Chapitre 5

Opinions et silences de la machine löbienne

Nous savons à présent que le mécanisme est incompatible avec le matérialisme. Nous savons que le mécanisme nécessite un psychologisme plutôt qu'un physicalisme pour fonder les expériences cognitives. Nous savons que le prix du computationnalisme, pour résoudre le problème du corps et de l'esprit, est de dériver la physique de la psychologie. La conscience ne peut plus être considérée comme émergeant de l'activité de la matière, au contraire nous devons, pour résoudre le problème du corps et de l'esprit *avec l'hypothèse computationnelle*, parvenir à expliquer comment les *apparences* d'univers et de matière émergent des expériences possibles de la conscience, c'est-à-dire, avec l'hypothèse mécaniste, des histoires computationnelles possibles. C'est à cette tâche que je propose de nous consacrer.

La partie démonstrative du travail est terminée. Ce qui suit est plus prospectif.

Nous allons tenter de dériver d'abord une phénoménologie de l'esprit à partir de l'informatique théorique et de la logique de la prouvabilité des machines. Ensuite nous allons tenter de dériver une phénoménologie de la matière à partir de la phénoménologie de l'esprit, comme le computationnalisme nous y contraint.

L'idée de base est simple, pour ne pas dire naïve. Je propose en effet d'interroger les machines elles-mêmes. Plus précisément je propose d'étudier certains discours possibles des machines digitales universelles. Toutes les machines n'ont pas nécessairement des choses intéressantes à dire, ce qui explique l'aspect plus prospectif. Je propose de nous limiter aux discours des machines qui sont autoréférentiellement correctes —c'est-à-dire qui n'énoncent que des propositions (arithmétiquement) vraies sur elles-mêmes— et qui possédent des "capacités introspectives" suffisamment élaborées. L'aspect prospectif n'est pas lié exclusivement au choix de ce type de machine, mais aussi aux variantes que nous ferons d'une des définitions de la *connaissance* que Théétète propose

à Socrate (Platon, 1950).

Je serai concis. A la différence de ce qui précède ce qui suit nécessite une certaine familiarité avec la logique mathématique et les logiques philosophiques. On consultera les annexes ou le rapport technique "autocontenu" détaillé (Marchal, 1995). On trouvera aussi dans ce rapport une collection de motivations, notamment basées sur l'usage du rêve en philosophie pour l'approche théétètique de la connaissance (voir aussi (Caillois, 1956), ainsi que les méditations dans (Descartes, 1953)).

5.1 Une "toute petite théorie de la conscience"

On se rappelle de la frayeur introspective et rétrospective du philosophe P_2 lorsqu'il réalise que son cerveau *aurait dû être détruit*. Cela illustre que le computationnalisme, paramétré à un niveau n, n'est pas communicable à la troisième personne (démontrable). Notons COMP_n la proposition selon laquelle je survis à une substitution opérée au niveau n. On a justifié (avec un minimum de psychologie populaire)

$$\mathrm{COMP}_n \Rightarrow \neg\Box\mathrm{COMP}_n$$

où "$\Box$" désigne une modalité de preuve ou de communication convainquante. Cela a-t-il un sens ? Peut-on résoudre logiquement "une équation" de la forme $x \rightarrow \neg\Box x$? Une solution facile, et valable pour de nombreuses logiques, est $x = \bot$, où "$\bot$" désigne votre proposition fausse préférée. Un peu de logique modale permet de trouver une solution plus intéressante: $x = \neg\Box\bot = \Diamond\top$. Cette solution, ou plus exactement la formule C :

$$\Diamond\top \rightarrow \neg\Box\Diamond\top$$

est valable dans tous les mondes appartenant à une vaste classe de référentiels de Kripke. Il suffit qu'à partir de chaque monde on puisse toujours accéder à un dernier monde. La formule C, ou sa généralisation avec une formule quelconque A à la place de $\top$, caractérise cette sorte de référentiel. Si on interprète la relation d'accessibilité par l'usage du télétransport, la formule en question rappelle que le mécanisme interdit toute garantie de survie : on peut accéder à un dernier monde (mourir) à chaque "voyage".

De façon précise, voilà une présentation de la logique TPTC (Toute Petite Théorie de la Conscience):

AXIOMES :	$\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$	K
	$\Diamond A \rightarrow \neg\Box\Diamond A$	C
RÈGLES :	$\frac{A \ , \ A \rightarrow B}{B}$	MP
	$\frac{A}{\Box A}$	NEC

Définition. Un état (ou un monde) est transitoire ssi il n'est pas un dernier monde. Un référentiel (W,R) est réaliste si tous les états transitoires aboutissent à un dernier monde.

On peut démontrer avec les techniques habituelles (voir annexe A et (Marchal, 1995))

Proposition 5

$$\text{TPTC} \vdash A \quad \text{ssi} \quad A \text{ est respectée par tous les référentiels réalistes}$$

5.2 La prouvabilité formelle

Le mécanisme permet de donner une motivation ou même une justification plus directe et n'usant pas de la moindre goutte de psychologie populaire pour la formule C. En effet, si on interprète $\Box p$ par $\text{Bew}(\ulcorner p \urcorner)$ où $\ulcorner p \urcorner$ désigne le nombre de Gödel d'une description formelle de la proposition p et où "Bew" (Bew*eisbar*) désigne le *prédicat* de prouvabilité de Gödel, alors, la formule C correspond au second théorème d'incomplétude de Gödel (Gödel, 1931).

On peut justifier, comme de nombreux auteurs, que le théorème de Gödel s'applique (au moins) aux machines universelles "suffisamment riches" (capables de prouver leur propre Σ_1-complétude). Et on peut conclure, avec le mécanisme, ou à partir de réflexions concernant les fondements des mathématiques comme celles de Webb et de Myhill (Webb, 1980; Myhill, 1952), que le (second) théorème de Gödel s'applique à nous. Il est en quelque sorte le premier théorème de "psychologie exacte". Il affirme, en terme de machine, qu'aucune machine (potentiellement) universelle, consistante et ayant les "facultés introspectives" suffisantes pour prouver sa "propre" Σ_1-complétude, n'est capable de prouver sa "propre" consistance. Je qualifierai une telle machine de *löbienne* en référence à l'article de Löb (Löb, 1955). Voir plus bas.

Le terme "propre" est mis entre guillemets car l'autoréférence gödélienne, étant complètement arithmétisable, ne permet pas de distinguer une "machine originale" d'une copie. Il s'agit donc d'une autoréférence *à la troisième personne*. Avec la prouvabilité formelle, la machine démontre des propositions la concernant, mais concernant tout autant une machine quelconque parmi ses doppelgängers. Gödel (1933) avait déjà constaté qu'il n'était pas possible d'utiliser la prouvabilité formelle pour formaliser la *connaissance* qui par nature est de la première personne. En effet, il faudrait qu'une machine consistante puisse communiquer les équivalents arithmétiques des formules $\Box A \rightarrow A$ (généralement admises pour la connaissance), mais cette communication contredirait le second théorème d'incomplétude. Cela découle directement des équivalences classiques :

$$\Diamond\top \leftrightarrow \neg\Box\bot \leftrightarrow (\Box\bot \rightarrow \bot)$$

Il n'en demeure pas moins qu'apparaît ici une phénoménologie naturelle de la communication (correcte et à la troisième personne) par une machine au sujet d'une version d'elle-même.

A partir de ce moment, il n'est pas possible de ne pas tenir compte d'une histoire (essentiellement hollandaise, italienne, soviétique, américaine) où, afin d'être court, je mentionne deux événements charnières:

- La réponse de Löb au problème de Henkin: que peut-on dire d'une formule arithmétique énonçant sa propre prouvabilité? Löb démontre qu'une telle formule est (paradoxalement) nécessairement vraie et prouvable (Löb, 1955).
- Les deux théorèmes de complétude de Solovay (Solovay, 1976).

  1. Solovay découvre que la formule (dite de Löb, c'est la version modale de la formalisation arithmétique du théorème de Löb)

     $$\Box(\Box A \to A) \to \Box A$$

     permet, accompagnée de la traditionnelle formule de Kripke

     $$\Box(A \to B) \to (\Box A \to \Box B)$$

     et des règles d'inférence du modus ponens ($\frac{A \; A \to B}{B}$) et de la nécessitation ($\frac{A}{\Box A}$) d'axiomatiser complètement la part prouvable *par la machine* de la logique de la prouvabilité formelle (arithmétisable) *de la machine* (précisément la logique modale ou le carré $\Box$ modélise le *Beweisbar* de Gödel). G désigne (souvent) la théorie obtenue.
  2. Dans le même article, Solovay démontre que si on adjoint aux théorèmes de G les formules de réflexion $\Box A \to A$, et que l'on ferme le système exclusivement par la règle du modus ponens (en abandonnant donc la règle de nécessitation) on obtient une théorie (souvent appelée G^*) qui axiomatise complètement la part *vraie* (et non pas simplement la part prouvable *par la machine*) de la logique de la prouvabilité formelle (arithmétisable) *de la machine*. En particulier G^* prouve $\Diamond\top$. En outre G et G^* sont décidables.

$G^* \setminus G$, la part vraie mais incommunicable (à la troisième personne) de la machine Löbienne constitue une curieuse logique décidable et fermée, comme G^*, pour le modus ponens et la "possibilisation" $\frac{p}{\Diamond p}$ (voir plus loin). Avec le computationnalisme, $G^* \setminus G$ axiomatise en quelque sorte un réservoir de propositions seulement inférables et jamais communicables.

De façon précise, voici une présentation de la logique G:

AXIOMES :	$\Box(A \to B) \to (\Box A \to \Box B)$	K
	$\Box A \to \Box\Box A$	4
	$\Box(\Box A \to A) \to \Box A$	L
RÈGLES :	$\frac{A \; A \to B}{B}$	MP
	$\frac{A}{\Box A}$	NEC

Définition R désignant une relation d'accessibilité (voir 1.2), une *échelle infinie* est une suite de mondes $a_0, a_1, a_2, a_3, ...$ tel que $a_0 R a_1 R a_2 R a_3 R ...$

Définition R est *bien chapeautée* ssi il n'existe pas d'échelles infinies.

On peut démontrer que (W, R) respecte les formules $\Box p \rightarrow \Box\Box p$ et $\Box(\Box p \rightarrow p) \rightarrow \Box p$ ssi R est transitive et bien chapeautée.

En fait on a le résultat de complétude suivant

Proposition 6

G $\vdash A$ ssi A est respectée par tous les référentiels transitifs et bien chapeautés

De même, voici G*

$$\begin{array}{lll} \text{AXIOMES :} & \text{les théorèmes de G,} & \\ & \Box A \rightarrow A & \\ \text{RÈGLES :} & \frac{A \;\; A \rightarrow B}{B} & \text{MP} \end{array}$$

G* n'est pas fermé pour la règle de nécessitation, ni même pour la règle de monotonie. Il n'admet donc ni sémantique de Kripke, ni sémantique de Scott-Montague. Boolos, à partir d'une idée de Solovay, a néanmoins trouvé une intéressante sémantique de Kripke *variable* pour G* (Boolos, 1980c; Solovay, 1976), voir aussi (Boolos, 1980a; Marchal, 1995).

Pour énoncer de façon rigoureuse les résultats de complétude *arithmétique* de Solovay, il faut préciser la façon dont les formules modales sont interprétées arithmétiquement. A cette fin on définit une "réalisation" F qui associe à chaque variable propositionnelle p_i un énoncé arithmétique, ou dans notre contexte un énoncé dans le langage de la machine löbienne M. L'interprétation arithmétique des formules se définit alors par induction sur la complexité des formules. Pour une formule modale le carré $\Box$ est interprété par le Bew arithmétique de Gödel. J'appelle "morphisme de Magari-Boolos" cette interprétation arithmétique de la logique modale (Magari, 1975).

$$\begin{array}{l} MB_F(p_i) = F(p_i) \\ MB_F(A \vee B) = MB_F(A) \vee MB_F(B) \\ MB_F(A \wedge B) = MB_F(A) \wedge MB_F(B) \\ MB_F(\neg A) = \neg MB_F(A) \\ MB_F(\Box A) = \text{Bew}(\ulcorner MB_F(A) \urcorner) \end{array}$$

Les formules $\neg A$, $\top$, $A \vee B$, $A \wedge B$, $A \leftrightarrow B$, $\Diamond A$, sont considérées, ici, comme des abréviations de $A \rightarrow \bot$, $\neg\bot$, $\neg A \rightarrow B$, $\neg(A \rightarrow \neg B)$, $(A \rightarrow B) \wedge (B \rightarrow A)$, $\neg\Box\neg A$ respectivement.

Le premier théorème de complétude de Solovay devient:

Théorème 7

G $\vdash A$ ssi pour toute réalisation F, M $\vdash MB_F(A)$

et le second devient:

Théorème 8

G* $\vdash A$ ssi pour toute réalisation F, $MB_F(A)$est vrai

Autrement dit, si la traduction d'une formule modale est vraie (arithmétiquement, ou dans le langage de M, etc.) alors G* la démontre.

Solovay a montré que G est décidable. De même il a montré en introduisant une transformation modale particulièrement intéressante que la décidabilité de G* se ramène à celle de G:

Proposition 9

$$\mathrm{G}^* \vdash A \text{ ssi } \mathrm{G} \vdash \mathrm{SOL}(A)$$

avec $\mathrm{SOL(A)} = (\bigwedge_{B_i \in S^{\Box}(A)} \Box B_i \rightarrow B_i) \rightarrow A$.

où $S^{\Box}(A)$ désigne l'ensemble des formules B_i avec $\Box B_i$ sous-formules de A.

De cette façon la formule $\mathrm{SOL}(A)$ permet de considérer un *philosophe mécaniste* assez "prudent". Pour choisir un niveau de description, il exige non pas une preuve (impossible) de l'adéquation du niveau, (c'est-à-dire une preuve qu'il ne va jamais finir (mourir) dans un monde accessible avec le télétransport), mais une preuve de l'adéquation du niveau pour chacune de ses parties finies, ainsi qu'une preuve que la conjonction des "réflexions" de ces parties finies entraîne sa survie. Dans ce cas il utilise le télétransport. Il reste imprudent car il ne sait toujours pas si le niveau du mécanisme choisi est adéquat, et en particulier si G* s'applique à lui-même "vu comme formule arithmétique" relativement à ce niveau. Mais, si le mécanisme est correct et si son choix est adéquat, il aura prouvé le plus qu'il est possible, par les deux théorèmes de Solovay, de prouver, pour un mécaniste.

G* \ G constitue un espace des solutions à notre équation $x \rightarrow \neg\Box x$. Justifier sa survie au télétransport est analogue à la preuve de la consistance d'une de ses extensions. Notons (avec la sémantique de Kripke) que l'axiome modale 4 est un théorème de G, ce qui pose des problèmes pour l'utilisation directe de G pour la mesure de l'indéterminisme (le "calcul des probabilités"). En effet la formule 4 entraîne la transitivité pour la relation d'accessibilité, et la sélection concerne les extensions *immédiates*. Je suggèrerai dans la section suivante une variante d'une idée de Théétète qui entraîne la disparition de 4 et la disparition de la fermeture pour la nécessitation.

G* \ G est-il une logique? Tenter de définir ce qu'est une *logique* nous entraînerait trop loin. Je me contente de deux remarques:

1. Aucune tautologie classique (ni intuitioniste[1] donc) n'est "démontrable" par (c'est-à-dire n'appartient à) G* \ G, puisque G démontre les tautologies classiques.
2. G* \ G est, quand même, fermé pour le modus ponens (MP). En effet si $A \in \mathrm{G}^* \setminus \mathrm{G}$, et $A \rightarrow B \in \mathrm{G}^* \setminus \mathrm{G}$, alors A et $A \rightarrow B \in \mathrm{G}^*$, et donc $B \in \mathrm{G}^*$, puisque G* est fermé pour MP. D'autre part, $G \nvdash B$, sinon G prouverait $A \rightarrow B$. En effet avec MP et le schéma tautologique (classique et intuitioniste) $A \rightarrow (B \rightarrow A)$, on dérive la règle $\frac{A}{B \rightarrow A}$. Donc $B \in \mathrm{G}^* \setminus \mathrm{G}$.

5.3 La phénoménologie du sujet

Admettons que le sujet est celui qui par nature est sujet de connaissance: il est celui qui peut savoir, ou qui peut connaître à la première personne. Que peut dire la machine *à son sujet*?

[1] Pour le mot "intuitionisme" je me conforme à l'orthographe proposée par M. Largeault (Largeault, 1992)

5.3.1 L'idée de Théétète

La justifiabilité formelle, qu'elle soit justifiable (axiomatisée par G) ou incommunicable (axiomatisée par G*), ne distingue pas la première et la troisième personne. Elle ne distingue pas "moi" de mon doppelgänger. De plus ni G, ni G* ne prouvent simultanément la réflexion $\Box A \rightarrow A$ et la nécessitation $\frac{A}{\Box A}$. Cela exclut l'usage directe de la justifiabilité formelle pour capturer la connaissabilité de la machine.

Une des plus vieilles idées de la philosophie, qui est énoncée notamment par Platon (voir le Ménon, et surtout le Théétète, voir (Platon, 1950), voir aussi (Burnyeat, 1991)) est de définir la connaissance par la justification *vraie*, de définir la connaissance de A par la justification de A accompagné —par définition— de la vérité de A. Cela revient à définir un nouveau connecteur modal correspondant à

$$\mathrm{Bew}(\ulcorner A \urcorner) \;\&\; A$$

En vertu du théorème de Tarski, il n'est pas possible d'interpréter arithmétiquement cette modalité par une formule du genre

$$\mathrm{Bew}(\ulcorner A \urcorner) \;\&\; \mathrm{Vrai}(\ulcorner A \urcorner)$$

puisqu'il n'existe pas de prédicat arithmétisable "Vrai" représentant la vérité arithmétique. Et de fait, cette forme de connaissance n'est pas arithmétisable. Si on admet que cette connaissance décrit bien une connaissance de la première personne, cela garantit que la première personne ne pourra se reconnaître en aucune présentation à la troisième personne d'elle-même. Ce qui est effectivement le cas avec le computationnalisme. La machine peut cependant communiquer (à la troisième personne) des propositions sur cette connaissabilité. Il est évident qu'à présent, la formule de réflexion $\Box A \rightarrow A$ est un schéma de formules communicables, puisque la machine prouve trivialement :

$$(\mathrm{Bew}(\ulcorner A \urcorner) \,\&\, A) \rightarrow A$$

On obtient ainsi une logique à la troisième personne du discours de la première personne. La première personne, sous la forme d'une logique intuitioniste IL, peut être retrouvée en inversant une transformation proposée indépendamment par Kolmogorov en 1932 et par Gödel en 1933 pour interpréter modalement la logique intuitioniste (Kolmogorov, 1932; Gödel, 1933). Une preuve précise de l'idée de Gödel apparaît chez McKinsey et Tarski (McKinsey and Tarski, 1948). Grâce au travail de Grzegorczyk de 1967, de façon indépendante Kuznetsov & Muravitsky 1977, Goldblatt 1978, et Boolos 1980 ont pu démontrer que la logique S4Grz, disposant de la formule de Kripke, et

des règles du modus ponens et nécessitation, ainsi que de la curieuse formule de Grzegorczyk 1967 :

$$\Box(\Box(A \to \Box A) \to A) \to A$$

axiomatise complètement la prouvabilité sur la connaissabilité (Grzegorczyk, 1967; Kuznetsov and Muravitsky, 1977; Goldblatt, 1978; Boolos, 1980c; Boolos, 1980a; Boolos, 1980b). De même Goldblatt a pu démontrer la complétude de la logique IL pour son interprétation arithmétique (Goldblatt 1978).

Boolos et Goldblatt ont démontré que notre "sujet", qu'il soit décrit à la troisième personne (S4Grz) ou par son discours à la première personne (IL) ne distingue pas la connaissabilité de la vérité. En effet, les "starifications" de ces logiques n'apportent pas de propositions nouvelles: S4Grz = S4Grz*, IL = IL*. *Du point de vue du sujet*, il n'y a pas de vérités qui ne soient pas connaissables (voir plus loin pour les références).

G* permet de démontrer que la première et troisième personne sont extensionnellement identiques (démontre les mêmes propositions de l'arithmétique). Il s'agit bien de la même personne vue selon des points de vue différents.

5.3.2 Réfutation de Lucas et Penrose

On peut, *sans user de l'hypothèse du computationnalisme*, utiliser les logiques G, G*, S4Grz, pour *invalider* de façon précise l'usage des phénomènes d'incomplétude visant à réfuter le mécanisme comme celle de Lucas que Penrose a remis récemment sur le tapis (Lucas, 1961; Lucas, 1968; Penrose, 1989). La part correcte de ces tentatives de réfutation montre seulement que *si* nous sommes des machines alors nous ne pouvons pas savoir quelle machine nous sommes, ce qui rejoint les conclusions de (Benacerraf, 1967), mais aussi les conclusions naturelles du computationnalisme obtenues avec les expériences par la pensée.

Regardons cela de plus près. Ce sera l'occasion de préciser rigoureusement la version arithmétique de l'idée de Théétète. En 1959, Lucas a tenté de réfuter le computationnalisme, sous la forme béhavioriste d'une sorte de test de Turing limité aux propositions de l'arithmétique informelle (voir Lucas 1961, voir aussi Penrose 1989, qui utilise cet argument et défend alors la possibilité d'une théorie quantique de la conscience : notons que je propose exactement l'inverse, accepter le computationnalisme et dériver de l'incomplétude une phénoménologie quantique de la matière).

L'argumentation de Lucas illustre à la fois la difficulté d'identifier le connaisseur sujet "$\boxdot$" avec la machine objet "$\Box$" et les difficultés conceptuelles qui se présentent lors des expériences par la pensée de multiplication de soi-même. Je pense avec Post (1922!), et Webb (1980) qui a consacré un ouvrage sur cet argument, que le raisonnement de Lucas est invalide. Je partage néanmoins avec Lucas et Penrose l'idée que le théorème de Gödel s'applique aux machines digitales. Cette dernière remarque est bien sûr une digression puisqu'avec notre stratégie pour isoler les phénoménologies nous nous sommes limités aux machines Löbiennes, qui sont a priori sujettes aux phénomènes d'incomplétudes.

Je pense, à présent, comme Benacerraf (1967), qu'une bonne partie du raisonnement de Lucas peut être rendu valide : la conclusion antimécaniste doit être affaiblie. Ce n'est pas que je ne suis pas une machine, mais seulement que si je suis une machine, alors je ne peux pas me reconnaître, de façon prouvable ou communicable, dans cette machine; ce que nous avons déjà illustré avec le computationnalisme au moyen de l'expérience par la pensée de la duplication de soi.

Je propose une reconstruction de cet argument, due en partie à (Benacerraf, 1967; Chihara, 1972; Reinhardt, 1985; Reinhardt, 1986), voir aussi (Wang, 1974). Cette reconstruction de la réfutation de Lucas, est formellement similaire aux critiques du mécanisme basées sur les expériences par la pensée de l'autoduplication (si je suis duplicable, je ne peux pas me reconnaître dans le dupliqué).

Lucas prend comme hypothèse qu'il est sain ($\boxdot A \rightarrow A$) et il décide de se comparer seulement aux machines saines ($\Box A \rightarrow A$). Lucas commet une identification entre machine et système formel (ce qui, on l'a vu, est une conséquence plausible du computationnalisme, au niveau adéquat avec la thèse de Church, mais nous n'avons pas besoin de cette identification: il est évident que si "je" suis une machine saine alors "je" suis une machine Löbienne).

La communicabilité $\Box$ de la machine Löbienne est arithmétisable et diagonalisable, c'est-à-dire qu'il existe un énoncé arithmétique p tel que $p \leftrightarrow \neg\Box p$. La machine saine lui étant présentée, Lucas peut trouver cet énoncé p, et démontrer $p \leftrightarrow \neg\Box p$.

L'argumentation de Lucas, ou plutôt sa reconstruction, peut alors être résumée dans la dérivation suivante :

1) $\boxdot (p \leftrightarrow \neg\Box p)$	machine löbienne;
2) $\boxdot (\Box p \rightarrow \neg p)$	calcul propositionnel;
3) $\Box p \rightarrow p$	la machine est saine ... (par hypothèse!)
4) $\boxdot (\Box p \rightarrow p)$	... sait Lucas;
5) $\boxdot (\Box p \rightarrow (p \wedge \neg p))$	calc. prop. + 2) et 4)
6) $\boxdot (\Box p \rightarrow \bot)$	calcul propositionnel;
7) $\boxdot \neg\Box p$	calcul propositionnel;
8) $\boxdot p$	par 7) et 1)
9) $\neg\Box p$	par 7) et $\boxdot p \rightarrow p$

et donc l'ensemble des propositions arithmétiques p que je (Lucas) peux savoir, c'est-à-dire telles que $\boxdot p$, est différent de l'ensemble des propositions arithmétiques p que la machine peut savoir (= démontrer, selon Lucas), c'est-à-dire telles que $\Box p$. Lucas semble être à même de se distinguer de toutes les machines (saines) sur un test de Turing limité à l'arithmétique. Où est l'erreur ?

En faveur de Lucas, une chose est claire : si $\boxdot$ obéit à S4, il n'est pas arithmétiquement définissable (il n'est pas finitairement définissable par lui-même). En effet, dans ce cas, il serait diagonalisable et il existerait un énoncé p tel que $p \leftrightarrow \neg\boxdot p$, donc $\boxdot p \rightarrow \neg p$, or $\boxdot p \rightarrow p$ (par la réflexion T), donc $\neg \boxdot p$, donc p (puisque $p \leftrightarrow \neg \boxdot p$), donc $\boxdot p$ (par nécessitation), donc $\bot$, puisqu'on a à la fois $\neg \boxdot p$ et $\boxdot p$. Remarquons la similarité de cette preuve avec la réfutation de Lucas.

Un raisonnement similaire montre directement que G* ne peut pas être fermé pour la nécessitation, puisqu'il a la réflexion. On voit en fait qu'aucune forme de communicabilité ne peut être à la fois diagonalisable, obéir à l'axiome de réflexion et être fermée pour la nécessitation.

En passant, on obtient une démonstration du théorème de Tarski comme quoi la vérité arithmétique n'est pas arithmétiquement définissable. En effet si tel était le cas, on disposerait d'un prédicat "Vrai" de vérité, et l'idée de Théétète pourrait produire une connais-

sance "$\boxdot p$" qui serait intensionnellement équivalente à $\mathrm{Bew}(\ulcorner p \urcorner) \wedge \mathrm{Vrai}(\ulcorner p \urcorner)$, qui obéirait à la réflexion, serait fermé pour la nécessitation et serait en même temps diagonalisable.

Pour trouver de façon précise l'erreur de Lucas dans notre contexte, c'est-à-dire avec le computationnalisme, il reste à appliquer l'idée de Théétète à la logique G. En réalité ceci n'est pas nécessaire, on peut se contenter de travailler dans une arithmétique étendue avec S4 (comme l'arithmétique épistémique de (Reinhardt, 1985; Reinhardt, 1986), ou de (Shapiro, 1985)).

S4 désigne le système suivant :

AXIOMES :	$\Box(A \to B) \to (\Box A \to \Box B)$	K
	$\Box A \to A$	T
	$\Box A \to \Box\Box A$	4
RÈGLES :	$\frac{A \ , A \to B}{B}$	MP
	$\frac{A}{\Box A}$	NEC

Dans ce cas, cependant, la relation entre la prouvabilité dans la théorie et la prouvabilité informelle n'est pas claire.

L'avantage de l'idée de Théétète appliquée à une machine auto-référentiellement correcte est de garantir au départ l'égalité *extensionnelle* de la prouvabilité formelle et de la prouvabilité intuitive.

L'idée de Théétète est capturée de façon précise par la transformation modale BGKM (pour Boolos 1980, Goldblatt 1978, Kusnetzov et Muravitsky 1977) de l'ensemble des formules modales dans l'ensemble des formules modales : MPL $\longrightarrow$ MPL :

Les variables propositionnelles sont supposées avoir été ordonnées p_i.

$$\begin{array}{l}
\mathrm{BGKM}(p_i) = p_i, \\
\mathrm{BGKM}(A \vee B) = \mathrm{BGKM}(A) \vee \mathrm{BGKM}(B), \\
\mathrm{BGKM}(A \wedge B) = \mathrm{BGKM}(A) \wedge \mathrm{BGKM}(B), \\
\mathrm{BGKM}(\neg A) = \neg \mathrm{BGKM}(A), \\
\mathrm{BGKM}(\Box A) = \Box(\mathrm{BGKM}(A)) \wedge \mathrm{BGKM}(A).
\end{array}$$

On peut alors démontrer:

Proposition 10

$$\mathrm{S4Grz} \vdash A \text{ ssi } \mathrm{G} \vdash \mathrm{BGKM}(A)$$

où S4Grz est naturellement le système S4 + Grz:

AXIOMES :	$\Box(A \to B) \to (\Box A \to \Box B)$	K
	$\Box A \to A$	T
	$\Box A \to \Box\Box A$	4
	$\Box(\Box(A \to \Box A) \to A) \to A$	Grz
RÈGLES :	$\frac{A \ , A \to B}{B}$	MP
	$\frac{A}{\Box A}$	NEC

On regarde alors le raisonnement de Lucas au niveau *de la vérité* G^*, et on constate que l'erreur se situe dans le passage de la ligne 3) à la ligne 4) puisque $G^* \vdash 3)$, mais $G^* \nvdash 4)$.

En composant BGKM avec l'interprétation arithmétique (dans le langage de la machine löbienne) de G, on obtient une interprétation arithmétique de $\boxdot$, c'est-à-dire, on obtient les transformations (paramétrées par F): $MB_F \circ$ BGKM : LPM $\longrightarrow$ L(M), et

Proposition 11

$$\text{S4Grz} \vdash A \text{ ssi pour toute réalisation } F \text{ M} \vdash MB_F(\text{BGKM}(A))$$

Ainsi $\boxdot p$ est interprété par $\text{Bew}(\ulcorner p \urcorner) \wedge p$. La théorie S4Grz, constitue ainsi une axiomatisation naturelle de la connaissabilité (intuitive et non diagonalisable) de la machine. Le bord flou du sujet —y compris l'incapacité qu'il a de se reconnaître objectivement— est capturé par le fait que $\boxdot$ n'est pas effectivement définissable par la machine, ni arithmétisable, ni diagonalisable. Comme il fallait s'y attendre, l'idée de Théétète empêche le sujet de se définir effectivement lui-même. De même, le sujet machine peut se dupliquer, mais il ne peut pas effectivement (de façon effective, constructive, communicable, ou connaissable) se reconnaître dans le dupliqué. On rejoint d'une certaine façon l'idée de Brouwer (le fondateur de la philosophie intuitioniste, cf (van Stigt, 1990)) selon laquelle le sujet (et son oeuvre) n'est pas (prouvablement) axiomatisable.

La formule de Grzegorczyk entraîne l'antisymétrie de la relation d'accessibilité pour les modèles finis. Cela suggère une interprétation temporelle (au sens subjectif) du développement local de la connaissance du sujet, ce qui permet une interprétation arithmétique du temps subjectif à-la-Bergson 1939 (voir aussi Dogen 1232-1253), ce qui encore est proche de la philosophie de la conscience et du développement temporel du soi de Brouwer (voir (Brouwer, 1905; Brouwer, 1983), voir aussi (Grzegorczyk, 1964)).

Et de fait, une logique intuitioniste représentable arithmétiquement, IL, émerge à ce stade. En effet, Gödel (1933) a suggéré et McKinsey & Tarski (1948) ont démontré qu'on peut interpréter la logique intuitioniste dans le système modal S4. Grzegorczyk a étendu le résultat pour S4Grz.

Voilà la transformation de 1933 de Gödel. Attention il s'agit d'une transformation du langage propositionnel (non modal) dans le langage propositionnel modal LPM:

$$\begin{aligned}
&\text{G33}(p_i) = \Box p_i \\
&\text{G33}(A \wedge B) = \text{G33}(A) \wedge \text{G33}(B) \\
&\text{G33}(A \vee B) = \text{G33}(A) \vee \text{G33}(B) \\
&\text{G33}(A \rightarrow B) = \Box\text{G33}(A) \rightarrow \Box\text{G33}(B) \\
&\text{G33}(\neg A) = \Box\neg\text{G33}(A)
\end{aligned}$$

On a, avec IL pour Logique Intuitioniste (Grzegorczyk 1967):

Proposition 12

$$\text{IL} \vdash A \text{ ssi S4Grz} \vdash \text{G33}(A)$$

Il suffit de composer les différentes transformations pour extraire l'interprétation arithmétique de l'intuitionisme (Goldblatt 1978, voir aussi Artemov 1990):

Proposition 13

$$\text{IL} \vdash A \text{ ssi pour toute réalisation } F \text{ M} \vdash MB_F(\text{BGKM}(\text{G33}A)))$$

J'argumente, avec le computationnalisme, en faveur de l'idée qu'il s'agit de la part solipsiste du sujet. Un résultat formel, qui confirme ce point de vue, est que le passage à G^* ne rajoute aucune formule : ni pour S4Grz (Boolos 1980a, 1980b), ni pour IL (Goldblatt 1978). Les propositions absolument indécidables mais vraies, c'est-à-dire telles que $\text{G}^* \vdash p \wedge \neg\boxdot p$, ne sont pas des images de propositions intuitionistes par la transformation BGKM. Avec des notations évidentes, Boolos et Goldblatt ont démontré que S4Grz = S4Grz* et IL = IL* : le solipsiste identifie correctement (de son point de vue) la prouvabilité et la vérité.

Notons qu'Artemov justifie (proprement) que l'idée de Théétète, précisément l'idée de définir la prouvabilité intuitive $\boxdot$ par la prouvabilité formelle accompagnée de la vérité, $\Box p \wedge p$, peut être érigée en une thèse de philosophie des mathématiques (comme l'est la thèse classique de Church, voir Artemov 1990, Marchal 1995).

Conclusion. Lucas a essayé de tirer une contradiction de l'ensemble des propositions $\{\Box p \rightarrow p, \boxdot p \rightarrow p, \Box p \leftrightarrow \boxdot p\}$, où p est une proposition arithmétique, $\Box$ représente la prouvabilité par une machine auto-référentiellement correcte et $\boxdot$ représente la prouvabilité intuitive de *l'incorrigible* Lucas.

La part correcte du raisonnement de Lucas montre seulement le caractère contradictoire de l'ensemble $\{\Box p \rightarrow p, \boxdot p \rightarrow p, \boxdot(\Box p \leftrightarrow \boxdot p)\}$. On a bien $G^* \vdash \Box p \rightarrow p, G^* \vdash \boxdot p \rightarrow p, G^* \vdash (\Box p \leftrightarrow \boxdot p)$, mais $G^* \nvdash \boxdot(\Box p \leftrightarrow \boxdot p)$. On retrouve bien la solution de Benacerraf.

Je mentionne le fait que Penrose commet clairement cette erreur dans "The Emperor's New Mind" (Penrose, 1989), mais ne la commet plus dans "The Shadow of the Mind". Il ne semble toutefois pas tenir compte de cette correction (Penrose, 1994). Penrose semble persuadé que la proposition "je suis une machine" n'a de sens *intéressant* que si elle entraîne la proposition "je sais quelle machine je suis". Mais, avec les expériences par la pensée, j'ai démontré que *si* le computationnalisme est correct, alors je ne peux jamais démontrer (ni savoir, avec l'idée de Théétète) quelle machine je suis. Et encore avec l'idée de Théétète, on vient de démontrer que cette impossibilité est l'apanage de toutes les machines löbiennes. Et on voit ici que les machines löbiennes elles-mêmes peuvent démontrer que si elles sont consistantes, alors, elles ne peuvent pas démontrer qu'elles sont telles ou telles autres machines. Notons que la question de savoir si une machine (löbienne) peut démontrer qu'elle est une machine *non spécifiée* est ouverte. Une telle question peut être formalisée dans l'extension du premier ordre de G^* (qui n'est pas complètement axiomatisable), et sort du cadre de ce travail (voir (Marchal, 1995; Reinhardt, 1985; Reinhardt, 1986)).

Pour paraphraser Post 1922, l'argument "Gödélien" ne peut pas montrer que l'homme n'est pas une machine. Post ajoute que l'argument montre seulement que l'homme ne peut pas construire une machine prouvant les mêmes théorèmes (de l'arithmétique) que lui. Ceci est encore trop dire puisqu'avec le second théorème de récursion de Kleene (Kleene, 1952), on peut rendre une machine quelconque (extensionnellement parlant) autoreproductible.Ce que je montre (avec le mécanisme) c'est qu'un sujet ne peut pas à la fois construire une machine capable de prouver les mêmes théorèmes (de l'arithmétique) que lui et *en même temps prouver* qu'il en est bien ainsi, c'est-à-dire prouver que cette machine est capable de prouver les mêmes théorèmes que lui. En interprétant $\boxdot$ par "je sais" et $\Box$ par "il croit", où "il" joue le rôle d'une duplication de "moi" (le rôle du doppelgänger), les situations paradoxales des expériences par la pensée sont suffisamment clarifiées pour confirmer (cela ne veut pas dire prouver) la thèse digitale et empêcher qu'on ne prenne des expériences par la pensée reposant sur la duplication, pour des réfutations du mécanisme.

En résumé, les erreurs dans l'usage du théorème de Gödel, ou des paradoxes de la duplication, pour réfuter le mécanisme reviennent en général (et à une reconstruction logique de l'argument près) à un usage simultané de T, Nec et de la diagonalisation (ou de la représentabilité arithmétique). Le tableau suivant récapitule la situation et peut servir de garde-fou contre ce genre de confusion intensionnelle en philosophie (mécaniste) de l'esprit:

	G	G^*	S4Grz
T	−	+	+
Nec	+	−	+
Diag	+	+	−

D'autres applications de ces logiques à la philosophie de l'esprit (aux rêves, aux réalités artificielles, aux malins génies et au cogito de Descartes, à la conscience-durée selon Brouwer-Bergson, etc.) sont proposées dans le rapport technique Marchal 1995. On peut consulter aussi (Slezak, 1983) pour une analyse assez semblable du cogito cartésien. Cette dernière résulte aussi d'une réfutation approfondie de Lucas (Slezak, 1982).

Notons encore que cette phénoménologie de la connaissance fournit une solution au problème du connaisseur (the Knower Paradox, (Kaplan and Montague, 1961), voir aussi (Grim, 1991)).

5.4 Phénoménologies de l'objet

La phénoménologie de l'objet, comme celle du sujet, est produite par une variante modale de l'autoréférence. Plus précisément elle va résulter des opérations successives suivantes.

- Une variante affaiblie de l'idée de Théétète: la restriction à la vérité va être remplacée par la restriction à la possibilité, c'est-à-dire ici la consistance. Cela va donner deux logiques: Z et sa "starification" Z^*. La motivation de base est le "darwinisme arithmétique" tel qu'il est imposé par les arguments du déployeur universel et du graphe filmé.

- La restriction des réalisations arithmétiques des formules atomiques aux énoncés Σ_1. Avec la thèse de Church de tels énoncés correspondent, du point de vue d'une machine löbienne, aux "feuilles" de l'arbre du déploiement universel. Cela va donner à nouveau un couple de logiques Z_1 et sa "starification" Z_1^*. Cette dernière *devrait*, si le computationnalisme est correct *et* si notre prospection philosophique est pertinente, correspondre à la phénoménologie de la matière.

5.4.1 De la vérité à la possibilité

L'idée de Théétète qui consiste à définir la connaissance par la justification vraie est la meilleure façon de garantir le lien entre le sujet et la vérité, quitte à faire de ce sujet un être essentiellemment *solipsiste*, incapable de se reconnaître en aucun autre. La connaissance privée et subjective, intuitive, est obtenue en liant à la base, la machine démonstratrice et la vérité. L'argument de l'auto-multiplication, ainsi que l'espoir de capturer, au moins, une notion de probabilité 1 pour la téléportation, nous invite à attacher la machine démonstratrice non plus à la vérité, mais à la possibilité ou à la consistance. On s'intéresse alors aux logiques modales dont le carré $\Box A$ admet une interprétation arithmétique du style:

$$\mathrm{Bew}(\ulcorner A \urcorner) \;\&\; \neg\mathrm{Bew}(\ulcorner \neg A \urcorner)$$

La raison principale est que la "probabilité" (la mesure associée à la sélection), dans les expériences d'automutiplication est "définie" sur le domaine de reconstitution. Le philosophe mécaniste fait abstraction (à la grande frayeur de P_2) de ses annihilations possibles. Il s'agit de la forme très abstraite de darwinisme arithmétique : les sondages sont restreints aux populations de "machines (sur)vivantes". Ceci est justifié par le fait que la sélection ne dépend pas des délais, et que la mesure dépend donc de l'entiereté du déploiement: la conscience supervient sur tous les états computationnels relatifs atteint par le déployeur. Pour la phénoménologie de l'esprit, l'idée de Théétète reposait sur une prouvabilité (justification formelle) restreinte par la vérité. Pour la phénoménologie de la matière, l'argument du déployeur universelle nécessite d'affaiblir cette restriction en remplaçant la *vérité* par la *possibilité*, c'est-à-dire ici, la consistance arithmétique.

Cela revient à définir un nouveau connecteur modal ■, et une nouvelle transformation DEON (pour déontique):

$$\begin{array}{l}\mathrm{DEON}(p) = p \text{ avec } p \text{ variable propositionnelle}\\ \mathrm{DEON}(A \vee B) = \mathrm{DEON}(A) \vee \mathrm{DEON}(B)\\ \mathrm{DEON}(A \wedge B) = \mathrm{DEON}(A) \wedge \mathrm{DEON}(B)\\ \mathrm{DEON}(\neg A) = \neg \mathrm{DEON}(A)\\ \mathrm{DEON}(\Box A) = \Box \mathrm{DEON}(A) \wedge \Diamond \mathrm{DEON}(A))\end{array}$$

On obtient une logique que j'appelle Z, partiellement axiomatisée par KDX_1, avec la formule K de Kripke et la formule "déontique" D:

$$\Box A \rightarrow \Diamond A$$

D est un affaiblissement de la formule de réflexion $\Box A \rightarrow A$ (souvent appelé T).

De telles nuances, entre prouvabilité, vérité et possibilité (consistance, sont rendues possibles par les nuances intrinsèques de la prouvabilité formelle "Bew" de Gödel. Ces formules, K et D, ainsi que leurs conséquences par déductions propositionnelles (modus ponens) ainsi que celles déduites avec l'usage de la règle de monotonie $\frac{p \rightarrow q}{\Box p \rightarrow \Box q}$ sont arithmétiquement saines, mais elles n'axiomatisent pas complètement Z. X_1 désigne alors zéro ou plusieurs axiomes qu'il reste à isoler.

Toutefois, grâce au premier théorème de complétude de Solovay, c'est-à-dire grâce à G, on dispose facilement d'un démonstrateur de théorèmes pour cette logique. Cela suffit à définir de façon précise Z comme ensemble de formules, à défaut de présentation axiomatique:

$$\mathrm{Z} = \mathrm{KD}X_1 = \{A \ \mid \mathrm{G} \vdash \mathrm{DEON}(A)\}$$

Et il en est de même pour la part incommunicable de ces logiques pour laquelle G^* permet la construction aisée d'un démonstrateur de théorèmes.

On obtient une logique $Z^* = KTX_2$, c'est-à-dire partiellement axiomatisée par K, T et où X_2 désigne encore zéro ou plusieurs axiomes qu'il reste à isoler. Comme ensemble de formules, Z^* est parfaitement bien défini:

$$Z^* = KTX_2 = \{A \mid G^* \vdash DEON(A)\}$$

Aucune de ces logiques ne prouvent la "transitivité": $\Box A \rightarrow \Box\Box A$, ni ne sont fermées pour la règle de nécessitation. Ces logiques n'admettent pas de sémantiques de Kripke. Pour $Z_1 = KDX_1$ on peut utiliser une sémantique plus large due à Scott et Montague et qui sied particulièrement bien à la notion de probabilité "immédiate" recherchée ici. On peut montrer que Z^*, comme G^*, n'admet pas de sémantique de Kripke, ni de sémantique de Scott-Montague (voir plus bas, avec l'annexe A). Pour Z^*, on peut espérer, en s'inspirant de la sémantique de Boolos pour G^*, parvenir à isoler une sémantique en terme de suites de modèles de Scott-Montague (Boolos, 1980c), cf aussi (Solovay, 1976).

5.4.2 La Σ_1-restriction

L'affaiblissement de l'idée de Théétète, qui affaiblit la vérité par la possibilité, ne suffit pas pour la phénoménologie de la physique. Les états d'esprit "atomiques" sont les états des machines löbiennes atteints par le déployeur universel. Il faut en tenir compte. On peut argumenter que ces états correspondent aux arrêts possibles des machines possibles, et, en terme de propositions, sont analogues aux énoncés Σ_1 (de la forme $\exists n P(n)$ avec P récursif). On peut trouver dans (Vickers, 1989) une motivation implicite pour un concept similaire d'observation (voir aussi (Abramsky, 1987)). Avec la thèse de Church l'ensemble des propositions Σ_1, qui existe indépendamment de nous avec le réalisme arithmétique, constitue un déployeur universel abstrait.

On peut donc raisonnablement espérer isoler une logique de "l'observable" en restreignant la logique des énoncés *prouvable et consistant* sur les réalisations arithmétiques Σ_1. Cela revient à restreindre l'interprétation arithmétique des variables propositionnelles des formules modales aux énoncés arithmétiques Σ_1.

Les machines löbiennes prouvent leur propre Σ_1-complétude. Plus précisément elle prouve les formules:

$$p \rightarrow Bew(\ulcorner p \urcorner)$$

où p désigne un énoncé Σ_1.

Visser a démontré que le système, que j'appelle V, comprenant les théorèmes de G accompagné de la formule $p \rightarrow \Box p$ et fermé pour le modus ponens et la nécessitation, axiomatise complètement la logique de la prouvabilité où les formules atomiques sont restreintes aux énoncés Σ_1 (Visser, 1985). Il convient toutefois d'affaiblir la règle de substitution aux énoncés atomiques. Par

exemple, on ne peut pas remplacer p par $\Diamond\top$ dans $p \rightarrow \Box p$. En fait, l'interprétation arithmétique de $\Diamond\top$ est $\Pi_1 \setminus \Sigma_1$. Si p représente un énoncé Σ_1, alors l'interprétation arithmétique de $\Box p$ est aussi Σ_1. On peut donc substituer p par $\Box p$, $\Box\Box p$, etc.

On obtient à nouveau une théorie V* complète *pour la vérité* de ces propositions en abandonnant la nécessitation et en complétant la théorie avec la formule de réflexion T ($\Box p \rightarrow p$). Notons que V* prouve $p \leftrightarrow \Box p$, mais (heureusement) V* ne prouve pas $\Box(p \leftrightarrow \Box p)$. Pourquoi heureusement ? Car si V* prouvait les nécessitations de $p \rightarrow \Box p$, la logique V* collapserait avec le calcul propositionnel. On peut montrer que V et V* sont décidables.

Pour isoler la part prouvable et la part vraie correspondant à la phénoménologie de la matière il suffit d'appliquer à la fois la restriction Σ_1 de Visser et l'affaiblissement de l'idée de Théétète sur la prouvabilité "Bew" gödélienne. Cela revient à étudier *le couple* de logiques $\mathrm{Z}_1 = \mathrm{KDX}_3$ et $\mathrm{Z}_1^* = \mathrm{KBX}_4$ qui correspondent aux logiques du $\Box$ où $\Box p$ est interprété par

$$Bew(\ulcorner p \urcorner) \;\&\; \neg Bew(\ulcorner \neg p \urcorner)$$

Avec p Σ_1. B est la formule $p \rightarrow \Box\Diamond p$. *Dans les modèles de Kripke*, elle correspond à la symétrie de la relation d'accessibilité du référentiel qui la respecte.

V et V* permettent de construire un démonstrateur de théorèmes pour les parts communicables et non-communicables respectivement. Ceci est rendu possible grâce aux théorèmes de complétude pour V et V* de Visser (Visser 1985). X_3 et X_4 désigne à nouveau les formules (peut-être en nombre infini) nécessaires pour obtenir un résultat de complétude.

La partie prouvable est donc obtenue ainsi:

$$\mathrm{Z}_1 = \mathrm{KD}X_3 = \{A \;\mid \mathrm{V} \vdash \mathrm{DEON}(A)\}$$

La partie vraie (prouvable ou non prouvable) est obtenue ainsi:

$$\mathrm{Z}_1^* = \mathrm{KB}X_4 = \{A \;\mid \mathrm{V}^* \vdash \mathrm{DEON}(A)\}$$

Comme pour G et G*, ainsi que Z et Z*, et à la différence de S4Grz ou de IL, la différence est non vide. Cette différence, $\mathrm{Z}_1^* \setminus \mathrm{Z}_1$, suggère une interprétation arithmétique des *qualia* "arithmétiques". Cela donne en effet une logique de propositions non-communicables, concernant des énoncés "immédiatement observables".

Remarque. Je précise que tous les théorèmes de complétude arithmétique cités jusqu'ici concernent les logiques modales propositionnelles. On peut

démontrer qu'au niveau des logiques du premier ordre aucune des théories modales n'admettent une extension décidable complète.

Précisons quelques points. On consultera l'annexe A pour une brève introduction à la sémantique de Scott-Montague.

On pourrait se demander ce que donnerait la Σ_1-restriction sur le sujet, vu à la troisième personne (S4Grz). L'idée est de capturer une notion de sensation vraie. On obtient une logique $\mathrm{S4Grz}_{\Sigma_1}$ qui possède tous les axiomes et qui est fermée pour toutes les règles d'inférence. Cette logique collapse. Elle prouve l'équivalence de toutes formules avec leur nécessitation et leur possibilisation. Autrement dit $\mathrm{S4Grz}_{\Sigma_1}$ est égale au calcul propositionnel. Son carré est strictement équivalent au non-arithmétisable, et donc non diagonalisable, prédicat de vérité. Je montre plus loin que tout espoir n'est pas perdu pour une notion de sensation vraie.

Z_1^*, comme V^*, frôle le collapse, en ce sens que, pour les formules atomiques p, Z_1^* prouve $p \leftrightarrow \Box p$ (V^* prouve $p \leftrightarrow \blacksquare p$), mais même pour les formules atomiques le collapse est évité car Z_1^* ne prouve pas $\Box(p \leftrightarrow \Box p)$. Notons que Z_1 ne prouve ni $\Box p \rightarrow p$, ni $p \rightarrow \Box p$. Ceci montre en particulier que $Z_1^* \setminus Z_1$ est non vide.

Ce qui évite le collapse à Z_1^*, est l'absence de fermeture pour la nécessitation. Et cette absence résulte de l'exigence de la consistance de l'extension et de la non-prouvabilité gödélienne de la consistance. Ni G, ni V ne prouvent $\Box\top \wedge \Diamond\top$. Aucune des logiques Z ne sont fermées pour la nécessitation. Les logiques Z et Z_1 admettent cependant une sémantique de Scott-Montague (voir plus bas).

On utilisera le fait que $\blacklozenge p$ est équivalent à $\neg\blacksquare\neg p$ qui est équivalent à $\neg(\Box\neg p \wedge \Diamond\neg p)$ qui est équivalent à $(\neg\Box\neg p \vee \neg\Diamond\neg p)$ qui est équivalent à $\Diamond p \vee \Box p$ qui est équivalent à $\Box p \vee \Diamond p$.

Dans le tableau suivant je donne quelques résultats qui étendent les nuances intensionnelles décrites plus haut. La formule "Dual T" est la duale modale de T: $p \rightarrow \Diamond p$. Elle est introduite pour distinguer Z de Z_1. G ne prouve pas $p \rightarrow \blacklozenge p$, même pour p atomique (par exemple $p = \top$) et donc Z ne prouve pas $p \rightarrow \Diamond p$. V prouve $p \rightarrow \blacklozenge p$, pour p atomique puisque, pour p atomique, V prouve $p \rightarrow \Box p$, et donc V prouve $p \rightarrow \Box p \vee \Diamond p$.

	G	G^*	S4Grz(*)	Z	Z^*	Z_1	Z_1^*	$\mathrm{S4Grz}_{\Sigma_1}$
T	−	+	+	−	+	−	+	+
Nec	+	−	+	−	−	−	−	+
Diag	+	+	−	+	+	+	+	−
N	+	+	+	−	+	−	+	+
RM	+	−	+	+	−	+	−	+
B	−	−	−	−	−	−	+	+
4	+	+	+	−	−	−	−	+
D	−	+	+	+	+	+	+	+
5	−	−	−	−	−	−	−	+
dual T	−	+	+	−	+	+	+	+

Z est fermé pour la règle de monotonie RM: $\frac{A \rightarrow B}{\Box A \rightarrow \Box B}$
Preuve. D'abord G est fermé pour RM. En effet, par une simple application de la nécessitation et de K, on a la fermeture de G pour $\frac{p \rightarrow q}{\Box p \rightarrow \Box q}$. A présent pour Z, cela découle du fait que $(\Box q \wedge \Diamond p) \wedge (p \rightarrow q)$ entraîne $\Diamond q$.

Et cela vaut a fortiori pour Z_1, où les p et q sont limités aux formules atomiques. Cela entraîne encore que Z et Z_1 sont fermés pour la règle RE $\frac{p \leftrightarrow q}{\Box p \leftrightarrow \Box q}$. Grâce à quoi, ces logiques bénéficient d'une sémantique de Scott-Montague.

Z^* n'est pas fermé pour RM, ni pour RE. On vérifie en effet facilement que $Z^* \vdash \top \leftrightarrow \Box\top$, mais $Z^* \nvdash \Box\top \leftrightarrow \Box\Box\top$.

Ceci n'est pas étonnant. G^* lui-même n'est pas fermé pour RE et RM, puisque $G^* \vdash \top \leftrightarrow \Diamond\top$, mais $G^* \nvdash \Box\top \leftrightarrow \Box\Diamond\top$.

Z^* démontre N: $Z^* \vdash \Box\top$, et T: $Z^* \vdash \Box p \rightarrow p$.

Je suis la nomenclature de Chellas où N désigne la règle d'adjonction de la formule $\Box\top$. $Z^* \vdash \Box\top$ est une conséquence directe de $G^* \vdash \Diamond\top$. Et $Z^* \vdash \Box\top$ est une conséquence (classique) directe de $G^* \vdash \Box p \rightarrow p$.

Z^* est donc une extension de KT_-, KT sans Nécessitation, et pourrait être proposé pour une forme de connaissance immédiate et incorrigible de la part d'une machine. Ce résultat montre que Z est strictement inclus dans Z^*.

Question. Est-il possible d'obtenir une axiomatisation de Z^* à partir de Z + T + la suppression de RM?

On montre facilement que $Z^* \nvdash \Box p \rightarrow \Box\Box p$. Donc Z non plus puisque $Z \subset Z^*$.

Z Z_1 perdent la fermeture pour la nécessitation, elles admettent cependant une sémantique de Scott-Montague.
Il est facile de se convaincre qu'elles prouvent en outre les formules K, M, R, où:

K est l'axiome de Kripke, $\Box(p \rightarrow q) \rightarrow (\Box p \rightarrow \Box q)$
M est la formule $\Box(p \wedge q) \rightarrow (\Box p \wedge \Box q)$,
R est la formule $(\Box p \wedge \Box q) \rightarrow \Box(p \wedge q)$.

Cela est aisé à vérifier. Notons que les logiques Z, sont décidables, puisque G et G^* le sont.

Le fait que les formules M et R sont des théorèmes permet de montrer que pour les mondes α, les systèmes de voisinages de Scott-Montague $\mathcal{N}(\alpha)$ sont des quasi-filtres:

- si x et y appartiennent à $\mathcal{N}(\alpha)$, alors $x \cap y$ appartient à $\mathcal{N}(\alpha)$,
- si $x \cap y$ appartient à $\mathcal{N}(\alpha)$, alors x et y appartiennent à $\mathcal{N}(\alpha)$.

Les systèmes de voisinages ne sont pas des filtres parce qu'ils n'ont pas d'éléments maximals. Ceci est dû au fait que KD$X \nvdash \Box\top$.

Après avoir appliqué la version faible de l'idée de Théétète à la prouvabilité gödélienne, on peut, curieusement, réitéré l'application de l'idée proprement dite de Théétète. La subtilité de l'incomplétude permet en effet de s'intéresser à la logique du carré $\Box$ où $\Box$ est interprété arithmétiquement par:

$$Bew(\ulcorner p \urcorner) \;\&\; \neg Bew(\ulcorner \neg p \urcorner) \;\&\; p$$

avec p Σ_1. Il n'y a pas de collapse! La différence entre le couple de logiques obtenues X \ X^* suggèrent l'existence d'une notion (inattendue) de *qualia vrai*. Voilà (peut-être) notre sensation physique vraie recherchée.

Je démontre plus loin que Z_1^* et X^* prouvent $p \rightarrow \Box\Diamond p$.

5.5 Comparaison avec la physique actuelle

La matière est un concept aussi nébuleux que la conscience. D'où vient-elle ? Quelle est sa nature ? Pourquoi semble-t-elle obéir à des lois mathématiques ? Comment la définir non circulairement ? etc. Les progrès de la physique contemporaine ont rendu le concept de matière encore plus nébuleux. Les travaux d'Einstein (notamment avec Podolski et Rosen, voir aussi de Broglie 1957) ainsi que le travail théorique de Bell 1964, ainsi que les travaux expérimentaux d'Aspect (pour citer les travaux les plus significatifs et les plus connus du domaine) ont permis de constater que la matière est même, en quelque sorte, expérimentalement nébuleuse (voir annexe C, voir (Bell, 1964; Clauser et al., 1969; Einstein et al., 1935; de Broglie, 1957; Aspect, 1976; Aspect et al., 1982)).

Il est temps de soumettre notre approche purement introspective de la physique (quoique partiellement communicable avec le computationnalisme) avec la physique moderne, en particulier avec la mécanique quantique. En réalité on aurait pu aborder cette comparaison tout le long de notre chemin. Ce que j'illustre à présent.

5.5.1 Indéterminisme et non-localité

Comme en mécanique quantique, le computationnalisme met en évidence un indéterminisme fort, ainsi qu'une forme de non-localité (éventuellement communicable à la troisième personne). C'était l'objet des chapitres précédents.

5.5.2 Etats multiples et relatifs

Avec le déployeur universel (concret ou abstrait grâce au graphe filmé) le computationnalisme entraîne l'existence d'une phénoménologie de mondes multiples ou d'états parallèles. Les états relatifs de la formulation d'Everett (voir annexe C) de la mécanique quantique sont des cas particuliers d'états computationnelles relatifs. (Le déployeur zigzague aussi sur les calculs de l'ordinateur quantique de (Deutsch, 1985)). Les probabilités sont aussi avantageusement définies sur des sortes d'histoires. Ce qui donne d'intéressants modèles de logiques quantiques (cf (Isham, 1994), voir aussi le récent ouvrage (Bub, 1997) pour une approche critique). Le computationnalisme force une interprétation "Many Minds, No World" (beaucoup d'esprits pas de mondes) de la mécanique quantique (Je dois cette expression à M. Bas C. van Fraassen).

Cette interprétation "Many Minds, No World" de la mécanique quantique résulte ici d'une sorte d'interprétation d'un genre plus général de type "many types no token" de l'arithmétique (et des arguments du déployeur universel et du graphe filmé). Elle permet de répondre à deux objections souvent rencontrée contre l'interprétation d'Everett. Par exemple, nous n'avons pas le

problème de l'extravagance ontologique des univers multiples puisque nous n'avons plus besoin d'aucun univers. Nous n'avons pas non plus le problème de justifier l'apparence correcte des probabilités alors que celle-ci ne sont justifiables qu'en terme de limite (Albert, 1992; Bohm and Hiley, 1993). Cela résulte directement du fait que la relation de sélection ne dépend pas des délais, et donc que le point de vue de toute (première) personne dépend pour tous les "ici et maintenant" possibles, de l'entièreté du déploiement.

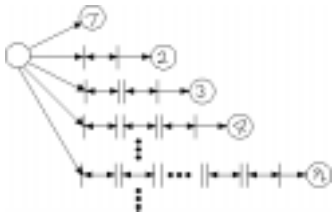


Figure 5.1: La conscience supervient sur une collection limite d'états

La phénoménologie de la matière rendue nécessaire par l'hypothèse du mécanisme étend de façon conservative la phénoménologie de la réduction du paquet d'onde des formulations *sans réduction* de la mécanique quantique (voir annexe C).

Je ne prétends pas ici que le mécanisme justifie les probabilités quantiques (ceci reste à montrer et je décris une tentative plus bas). Je dis seulement que si on trouve les bonnes probabilités comme limite, elles seront d'office justifiées pour chaque expérience que peut faire un observateur. En particulier l'invariance de la sélection pour les délais et lieux de reconstitutions (virtuelles ou, avec l'argument du graphe filmé, purement arithmétique) justifie le caractère non-observable des histoires computationnelles appartenant à des collection d'histoires de mesure nulle.

Notons encore la ressemblance déjà mentionnée entre la *normalité profonde* de la phénoménologie de l'univers, et la notion de profondeur quantique (Q-depth) de (Deutsch, 1985).

5.5.3 Logique quantique : contrefactualité

Le graphe filmé montre que l'essence de la notion de "causalité computationnelle" ou d'histoire computationnelle, repose sur la notion de contrefactualité. Un article d'Hardegree pourrait faire de l'argument du graphe filmé un raccourci entre le computationnalisme et la logique quantique comme squelette d'une phénoménologie de la matière. En effet Hardegree a découvert que l'"implication matérielle" (voir annexe C) de la logique quantique "orthodoxe" (orthomodulaire, voir (Birkhoff and von Neumann, 1936; Mittelstaedt, 1978; Dalla Chiara, 1986)) peut être interprétée comme une conditionnelle contrefactuelle selon l'approche modale de Stalnaker (Hardegree, 1976; Stalnaker,

1984). Ici, le concept syntactique clé est le concept de conditionnelle contrefactuelle. Le concept sémantique clé est celui de mondes suffisamment similaires.

5.5.4 Logiques quantiques : mesures et qualia "arithmétiques"

La façon la plus directe de comparer la phénoménologie de la matière dérivée du mécanisme et la physique actuelle consiste à comparer la "logique empirique" des physiciens, la logique quantique et la logique de la matière dérivée ici à partir des modalités de l'autoréférence. Le fait remarquable est le suivant. Les théories Z_1^* et X^* prouvent, avec p atomique, la formule

$$p \rightarrow \Box\Diamond p$$

Il s'agit justement de la formule qui a permis à Goldblatt de construire une logique modale (fermée pour la nécessitation) interprétant la logique quantique, de la même manière que S4, ou S4Grz, permettent d'interpréter modalement la logique intuitioniste (Goldblatt 1974). Cela invite naturellement à rechercher les logiques faibles (internes) correspondantes x, x^*, z, z^*:

$$\frac{\mathrm{IL}}{\mathrm{S4Grz}} :: \frac{z}{\mathrm{Z}_1} :: \frac{z^*}{\mathrm{Z^*}_1} :: \frac{x}{\mathrm{X}} :: \frac{x^*}{\mathrm{X}^*}$$

Cela devrait donner des interprétations arithmétiques de la logique quantique. Il faut prendre cette remarque avec précaution car nos logiques modales ne sont ni fermées pour la nécessitation, ni ne permettent un usage inconsidéré de la substitution. Ces systèmes peuvent être utilisés pour comparer la phénoménologie de la non-localité computationnelle avec la non-localité quantique. Le présent travail constitue un pont entre Maudlin 1989 et Maudlin 1994 (voir plus loin). On peut aussi utiliser les travaux de Dalla Chiara et de J. L. Bell pour extraire une notion de "probabilité 1" à partir de ce genre de logique modale (Dalla Chiara, 1977b; Bell, 1986), voir aussi (Fattorosi-Barnaba and Amati, 1987; Alechina, 1994).

Démontrons que la phénoménologie de la matière isolée ici prouve B :

Théorème 14 *Z_1^* prouve $p \rightarrow \Box\Diamond p$*

Preuve. Il suffit de montrer que $\mathrm{G}^* \vdash p \rightarrow \blacksquare\blacklozenge p$ pour les formules atomiques. Comme $\blacksquare p = \Box p \wedge \Diamond p$ et $\blacklozenge p = \Box p \vee \Diamond p$, il suffit de montrer que, si $\mathrm{V}^* \vdash p$, alors

$$\mathrm{V}^* \vdash \Box(\Box p \vee \Diamond p) \wedge \Diamond(\Box p \vee \Diamond p)$$

Cela ne présente pas de difficultés particulières. La démonstration qui suit utilise le théorème de la déduction. Cela est permis pour les dérivations qui n'utilisent pas la nécessitation (voir annexe A). V* prouve p entraîne V* prouve $\Box p$ (car V* prouve $p \rightarrow \Box p$, la nécessitation n'est pas utilisée ici). V* prouve donc $\Box p \vee \Box\Diamond p$, donc (V* prouvant 4), V* prouve $\Box\Box p \vee \Box\Diamond p$, donc V* prouve $\Box(\Box p \vee \Diamond p)$. On a utilisé ici le fait que G* prouve $\Box A \vee \Box B \rightarrow \Box(A \vee B)$. A présent, vu l'absence de nécessitation dans la dérivation, on peut utiliser le théorème de déduction, et on obtient V* prouve $\Box p \rightarrow \Box(\Box p \vee \Diamond p)$. De même, la fermeture pour la possibilisation de G* entraîne, lorsque V* prouve p, que V* prouve $\Diamond(\Box p \vee \Diamond p)$.

En 1936, Birkhoff et von Neumann ont en effet considéré la logique (algébrique) QL que l'on peut tirer des propositions quantiques, celles-ci correspondant à des sous-espaces d'un espace de Hilbert (voir annexe C). Cette logique est parfois appelée *logique quantique minimale* pour la distinguer de celle auquel on rajoute un axiome d'orthomodularité (voir plus bas).

Le statut de la logique quantique concernant les fondements de la mécanique quantique, est, en fait, souvent tenu pour nébuleux. (van Fraassen, 1974) parle franchement du *labyrinthe des* logiques quantiques, au pluriel.

La même année, Goldblatt a montré

$$\mathrm{B} \vdash \mathrm{GOLDB}(p) \Leftrightarrow \mathrm{QL} \vdash p$$

où $\mathrm{GOLDB}(p)$ constitue, comme G33, une transformation du langage propositionnel en logique modale:

$$\begin{aligned}
&\mathrm{GOLDB}(p) = \Box\Diamond p \text{ pour } p \text{ atomique}\\
&\mathrm{GOLDB}(\neg A) = \Box\neg\mathrm{GOLDB}(A)\\
&\mathrm{GOLDB}(A \wedge B) = \mathrm{GOLDB}(A) \wedge \mathrm{GOLDB}(B)
\end{aligned}$$

(Goldblatt, 1974), voir aussi (Dalla Chiara, 1977b; Dalla Chiara, 1986). Goldblatt utilise une présentation *à-la-Gentzen* de la logique quantique (voir annexe C). La raison est qu'il n'existe pas d'implication *convenable* en logique quantique. En affaiblissant la tautologie classique $p \rightarrow (q \rightarrow p)$ dans une présentation *à-la-Hilbert* du calcul propositionnel classique, (Dalla Chiara, 1976) donne une telle présentation de la logique quantique, avec une implication "matérielle" (dont l'interprétation intuitive est plus proche de la déduction cependant). Cette implication ne permet pas de théorème de la déduction.

Toujours est-il que si on veut mesurer la distance entre la phénoménologie de la matière telle que celle-ci est décrite par les physiciens (actuels) et la phénoménologie de la matière que le mécanisme oblige de dériver, on peut comparer la logique quantique et une logique, que j'appelle QuelQL* (Quel Quantum Logic?). Celle-ci est définie par

$$\mathrm{QUELQL}^* = \{p \mid \forall F_{\Sigma_1}\, \mathrm{MB}_{F_{\Sigma_1}} \circ \mathrm{DEON} \circ \mathrm{GOLDB}(p)\}$$

Autrement dit QUELQL* = $\{p \mid \mathrm{Z}_1^* \vdash \mathrm{GOLDB}(p)\}$. La part communicable est naturellement définie par

$$\mathrm{QUELQL} = \{p \mid \forall F_{\Sigma_1}\, \mathrm{M} \vdash \mathrm{MB}_{F_{\Sigma_1}} \circ \mathrm{DEON} \circ \mathrm{GOLDB}(p)\}$$

où M désigne une machine löbienne. Autrement dit QUELQL = $\{p \mid \mathrm{Z}_1 \vdash \mathrm{GOLDB}(p)\}$ Si on préfère :

$$\mathrm{QUELQL}^* = \{p \mid \mathrm{DEON}(\mathrm{GOLDB}(p)\}$$

Il est plus simple, mais moins informatif, de comparer directement B et Z_1^*. Ils ont en commun tout ce que B est capable de prouver uniquement avec le modus ponens MP, mais leurs différences symétriques respectives sont chacune non vides: B et Z_1^* prouvent chacun $\Box\top$, B prouve $\Box\Box\top$ par nécessitation NEC, Z_1^* prouve $\neg\Box\Box\top$. À présent $\Box\Box\top$ n'est pas la traduction par GOLDB d'une proposition quantique, et il reste à mesurer l'étendue des conséquences de l'absence de la nécessitation de Z_1^*. Malgré l'absence de nécessitation (NEC) et de monotonie (RM), il se pourrait que QL et QUELQL* soit identique. Par exemple B $\vdash p$ entraîne B $\vdash \Diamond p$ par réflexion (contraposée) et par MP, et ensuite, par NEC, on a B $\vdash \Box\Diamond p$. $\mathrm{Z}_1^* \vdash p$ n'entraîne pas $\mathrm{Z}_1^* \vdash \Diamond p$, car le schéma $\Box p \to p$ n'entraîne pas a priori le schéma $p \to \Diamond p$ (si p est Σ_1, $\Diamond p$ est Π_1). De même $\mathrm{Z}_1^* \vdash \Diamond p$ n'entraîne pas a priori $\mathrm{Z}_1^* \vdash \Box\Diamond p$ puisqu'on ne dispose pas de la nécessitation. Cependant $\mathrm{Z}_1^* \vdash p$ entraîne effectivement bien $\mathrm{Z}_1^* \vdash \Box\Diamond p$, puisqu'on a le modus ponens (MP) et $\mathrm{Z}_1^* \vdash p \to \Box\Diamond p$.

Il y a cependant peu de chance que QuelQL* soit identique à QL, la question est de savoir où se situe QuelQL* et QuelQL dans le "labyrinthe des logiques quantiques". Par exemple, le "Σ_1-bord computationnel" vérifie-t-il des formules de modularité, comme l'orthomodularité? Dalla Chiara 1977 donne une version modale de l'orthomodularité :

$$(p \wedge \neg q) \to \Diamond(p \wedge \Box\neg(p \wedge q))$$

Cette formule ne doit pas être un théorème de Z_1^* ou de B, elle doit cependant être vérifiée pour les *orthopropositions* (c'est-à-dire les images de p par GOLDB). De même on peut se formuler des questions du genre : le bord computationnel viole-t-il les inégalités de Bell ? Il suffit d'appliquer GOLDB à (une forme) d'inégalité de Bell (voir annexe C): $A \cap B \subseteq (A \cap C) \cup (B \cap \overline{C})$. Cela donne :

$$\Box\Diamond A \wedge \Box\Diamond B \to \Box\Diamond((\Box\Diamond A \wedge \Box\Diamond C) \vee (\Box\Diamond B \wedge \Box\neg\Box\Diamond C))$$

Avec l'absence de nécessitation et de monotonie pour Z_1^*, il s'agit plus de "crédibilités quantiques relatives" que de probabilités.

Un démonstration similaire montre que X* prouve B de la même façon que Z_1^*. Il n'y a donc pas d'analogue de la thèse d'Artemov pour la relation entre

Z_1^* et B. Cela donne plus de possibilités pour le *fil d'Ariane* des interprétations arithmétiques des logiques quantiques, tout en illustrant à nouveau les subtiles nuances des modalités de l'autoréférence gödélienne.

5.5.5 Comparaison avec Maudlin et Penrose

Dans son article de 1989 sur la conscience et le computationnalisme, Maudlin use d'une notion de computationnalisme plus restreinte que la nôtre. Il estime en effet que si le niveau de computationnalisme est quelconque le mécanisme est d'office trivialement vrai. J'ai justifié ici que le computationnalisme au contraire, quel que soit le niveau *fixé*, entraîne l'inconnaissabilité du niveau (ce qui n'entraîne pas la non-pariabilité sur *un* niveau). En fixant le niveau au départ (comme s'il était jugé connaissable), Maudlin n'a pas pu voir que son argument ne dépend pas du niveau de computationnalisme, et cela explique peut-être pourquoi il semble ne pas avoir réalisé le renversement que son argument impliquait. Son argument (équivalent au "graphe filmé", Olympia y joue le rôle du graphe filmé, (Marchal, 1988; Maudlin, 1989; Marchal, 1995)) montre, qu'aussi fin soit le niveau, le computationnalisme rend la supervénience physique impossible à ce niveau (ce qui oblige, avec l'argument du déployeur universel, de rendre l'apparence de la matière émergeante sur les histoires computationnelles possibles définies sur tous les niveaux digitalisables possibles). Autrement dit Maudlin n'a pas vu que si le computationnalisme est "(d'office) vrai quand on ne fixe pas précisément le niveau", alors son propre argument le force d'abandonner la supervénience physique, et d'accepter l'interprétation "Many Minds, No World" de l'arithmétique. En 1994 Maudlin publie un ouvrage où il aborde le problème de la compatibilité entre la théorie de la relativité et la non-localité quantique. Il semble, selon Maudlin, que si on veut conserver la théorie quantique dans un environnement minkowskien (fut-il *tiltant*, c'est-à-dire aussi bien en relativité restreinte qu'en relativité générale), il faille s'accomoder des "poisons" (écrit-il) suivants :

- abandonner l'invariance de Lorentz (principe clé pourtant de la relativité),
- admettre une causalité du futur sur le passé,
- accepter une interprétation "Everett-like" de la mécanique quantique, sous la forme "Many Minds (One World)" d'Albert et Loewer (Albert and Loewer, 1989).

Maudlin maugrée: il n'apprécie aucun des "poisons" proposés.

Nous avons cependant montré que le computationnalisme, où le niveau est un paramètre, permet d'extraire, en un seul coup, une nécessaire interprétation "Many Minds, No World" de la réalité arithmétique. Olympia (le "graphe filmé" de Maudlin, la femme-machine du conte d'Hoffmann qui inspire Maudlin) n'a pas dit son dernier mot! Si on veut conserver le matérialisme et un univers, on est obligé, comme Penrose, d'abandonner le computationnalisme forcément *à tout niveau*.

Penrose est un physicien qui aborde le problème du corps et de l'esprit (PCE) dans plusieurs ouvrages qui sont devenus des best-sellers et qui sont très controversés voir (Penrose, 1989; Penrose, 1990; Penrose, 1994). Il est opportun de comparer la présente approche avec celle de Penrose car elles abordent le même puzzle (PCE), avec essentiellement les mêmes pièces du puzzle (l'incomplétude Gödélienne et la mécanique quantique). Les pièces sont disposées cependant de façon quasi-diamétralement opposées. D'abord, outre le platonisme (chez moi limité à l'arithmétique) il y a deux importants points communs :

- Les phénomènes d'incomplétude s'appliquent aux machines.

- Le computationnalisme est faux dès lors qu'on tient à l'idée d'un univers matériel ou substantiel qui aurait un rôle dans l'existence de la conscience. (Ce que démontre le graphe filmé).

Les argumentations divergent cependant. Comme le computationnalisme est notre hypothèse de travail, on a argumenté qu'on pouvait admettre l'idée que les phénomènes d'incomplétude s'appliquent à nous, et on a démontré (avec l'argument du graphe filmé et l'argument du déployeur universel) que le concept d'univers substantiel est nécessairement redondant, c'est-à-dire incapable d'expliquer aussi bien l'origine de la conscience que l'origine de nos observations.

Penrose utilise l'incomplétude Gödelienne pour "réfuter" (incorrectement : voir plus haut) le mécanisme, et propose de modifier la mécanique quantique et la relativité (ce qui est cohérent aussi bien avec Maudlin qu'avec le présent travail) pour justifier un substantialisme non computationnel. Au contraire, j'utilise l'incomplétude *et* le mécanisme pour dériver une phénoménologie de la réalité qui étend celle des mécaniques quantiques sans réduction de l'onde. Ontologiquement et argumentativement, Penrose est au pôle opposé de ce travail. Paradoxalement peut-être, ses *propositions* sont cohérentes relativement au computationnalisme, à la différence des propositions de ceux qui veulent marier le computationnalisme avec le matérialisme, comme celles de nombreux fonctionnalistes.

5.5.6 L'arithmétique comme *"Théorie de Tout"*

Il a lieu de comparer brièvement ici notre approche avec les TOE (Theory Of Everything, théorie de tout) proposées par certains physiciens.

Les TOE reposent sur l'idée, forcément spéculative, de l'existence d'une unification des lois de la physique, ainsi que sur l'idée philosophique selon laquelle il existe un univers concret dont le statut ontologique est indépendant de nos observations (HU).

Nous avons démontré qu'un tel univers, avec le computationnalisme, ne peut expliquer ni l'origine de nos sensations à la première personne, ni surtout l'origine de nos observations et de nos mesures communicables à la troisième personne.

Avec le mécanisme, une unification des lois de la physique devrait être justifiée sur base de la mesure définie sur les histoires computationnelles.

Même sans le mécanisme, on serait en droit de poser aux physiciens, en supposant qu'ils aient unifiés les lois de la physique: "pourquoi ces équations-là? Pourquoi de telles conditions initiales? etc." Et on ne voit pas comment ils pourraient justifier l'origine de ces équations sur base d'une loi physique sans tomber dans une régression infinie (voir (Wheeler, 1994), et aussi (Deutsch, 1986a) pour une réflexion similaire, voir aussi (Gardner, 1996)).

On voit donc mal comment la physique pourrait résoudre le "problème dur de la matière": pourquoi un univers semble-t-il exister et obéir à des lois?

En ce qui concerne le "problème dur de la conscience", la situation serait d'autant plus catastrophique que les physiciens parviendraient à unifier des lois exhaustives de la nature, car alors elles rendraient superflue la conscience et la première personne. C'est le bien connu "paradoxe du fonctionnalisme" (voir annexe D, voir aussi (Tye, 1995)).

Voilà peut-être pourquoi de nombreux "scientifiques" matérialistes estiment que les deux problèmes "durs" ne sont pas scientifiques. L'attitude proprement "scientifique", me semble-t-il devrait plutôt consister, lorsqu'on a le sentiment qu'une science A ne peut pas résoudre un problème, de voir si une autre science B ne peut pas résoudre le problème, quitte à remettre en cause les a priori philosophiques qui faisaient de la science A une science fondamentale. J'ai démontré ici que c'est exactement ce qui doit se passer si on prend l'hypothèse du computationnalisme au sérieux. Une psychologie générale, reposant exclusivement sur l'informatique théorique et/ou la théorie des nombres, et ça de façon non réductionniste, *doit* devenir fondamentale.

Le miracle est que la thèse de Church et la non-trivialité de l'autoréférence des machines "abstraites" rend cette approche, non réductionniste par nécessaire incomplétude, *possible*. L'autre miracle est la ressemblance entre les phénoménologies du mécanisme et les phénoménologies de la mécanique quantique *sans réduction de l'onde* (voir annexe C).

Martin Gardner exprime son scepticisme à l'égard de l'idée qu'il puisse y avoir une *théorie de tout*, dans son article sur la théorie des supercordes. Cette théorie est considérée comme une candidate à une *théorie de tout* par des physiciens. Il dit :

> There is, of course, no way a scientist can answer the superultimate question of why, as Stephen Hawking recently put it, the universe bothers to exist. (Gardner, 1996)

Mais il n'y a aucun moyen pour un scientifique de *prouver* l'existence de l'univers, et avec l'hypothèse du computationnalisme, un scientifique peut répondre *Sir(e), je n'ai pas besoin de cette hypothèse*. On ne prétend pas ici avoir une explication de l'existence de l'apparence de l'univers, on prétend avoir seulement démontré, avec les arguments du déployeur universel (et donc avec la thèse de Church) et du graphe filmé, que si on accepte l'hypothèse du mécanisme on doit rendre compte de cette apparence de toute façon. Quant aux similarités entre la phénoménologie mécaniste de la matière et le monde quantique, elles constituent (seulement) un début de confirmation inductive du computationnalisme, et de l'importance des modalités de l'autoréférence (cf. aussi (Dalla Chiara, 1977a)).

En fait, indépendamment de toute considération sur la nature de la science fondamentale, il est souvent tenu pour "évident" qu'il est impossible de savoir pourquoi nous existons. Cette évidence repose sur le sentiment qu'une explication doit reposer sur des prémisses reposant elles-mêmes sur une explication, et ainsi de suite. On réalise cependant que pour concevoir et juger définitive cette impossibilité, il est nécessaire de concevoir la litanie des nombres naturels: le "ainsi de suite".

Cependant, avec le computationnalisme, ce qui est nécessaire pour concevoir cette litanie, est suffisant pour comprendre pourquoi il est nécessaire pour les machines abstraites et consistantes, autoréférentiellement correctes relativement à leur type d'environnement computationnel "le plus proche" d'en venir à se poser elles-mêmes des questions sur leur nature et la nature de leurs environnements, et d'en venir, parfois, à produire des inférences correctes.

La portée du computationnalisme est d'autant plus grande qu'on sait depuis les résultats d'incomplétudes, et l'écroulement qui s'en suivi du logicisme, qu'il n'est pas possible d'axiomatiser l'arithmétique ou l'informatique de façon finie ou réductionniste. Le computationnalisme permet et oblige d'extraire une phénoménologie de l'esprit et une phénoménologie de la matière de l'arithmétique, mais laisse nécessairement intact le mystère de nos croyances en la vérité arithmétique, justifiant partiellement, de "l'intérieur", le caractère injustifiable de son ontologie.

Voilà pourquoi, avec la thèse de Church, et la confirmation quantique du mécanisme, l'arithmétique intuitive, alias la théorie des nombres et ses variantes intensionnelles, pourrait bien être la plus simple et la plus riche "théorie de tout" qu'on puisse avoir à notre disposition.

Il ne s'agit pas ici de proposer une panacée universelle. Il s'agit au contraire de comprendre que le computationnalisme force un renversement de point de vue, qui rend plus large encore notre ignorance relative, et qui rend plus vaste l'espace de nos horizons accessibles. En toute matière, sans jeu de mots, l'acte de foi du *philosophe mécaniste* lui donne plus de raisons d'espérer et plus de raisons de craindre.

Annexe A

Logique modale

Les logiques modales considérées sont toutes des extensions de la logique propositionnelle classique. On dispose d'un alphabet

$$\top, \bot, \wedge, \vee, \neg, \rightarrow, p, q, r, ..., \Box, \Diamond$$

$p, q, r...$ sont les variables propositionnelles. On dispose d'un ensemble de formules. C'est l'ensemble des formules générées par un nombre fini d'applications exclusives des règles suivantes :

1. $\top, \bot, p, q, r, ...$ sont des formules (appelées formules atomiques)
2. si A est une formule, $(\neg A), (\Box A), (\Diamond A)$ sont des formules (dites composées)
3. si A et B sont des formules, $(A \wedge B), (A \rightarrow B), (A \vee B)$ sont des formules (dites composées)

Exemples: "$(\neg p)$" est une formule, "$((\Box p) \rightarrow p)$", "$(\neg((\Box p) \rightarrow p))$" sont des formules, "$(\neg \wedge p \, \Box$" n'est pas une formule. Abréviation : Nous supprimerons les parenthèses formelles pour alléger l'écriture. Par exemple l'expression "$\neg(\Box p \rightarrow p)$" est considérée comme une abréviation de la formule "$(\neg((\Box p) \rightarrow p))$".

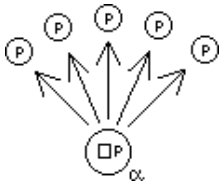
Remarque. "*Partout* p" est intuitivement équivalent à "*nulle part* $\neg p$", c'est-à-dire "$\neg$ *quelque part* $\neg p$", de même avec les couples (*nécessaire*, *possible*), ($toujours$, $quelquefois$) etc. On s'intéressera exclusivement aux logiques pour lesquelles $\Box = \neg \Diamond \neg$, et réciproquement $\Diamond = \neg \Box \neg$. Remarquons qu'il en est de même pour les quantificateurs de la logique classique de prédicats $\forall$ et $\exists$, ($\forall = \neg \exists \neg$, $\exists = \neg \forall \neg$).

A.1 La sémantique de Kripke

L'idée intuitive de $\Box p$, est que p est vrai dans tous les mondes possibles, ou dans tous les états possibles, j'utilise "monde" et "état" comme des termes informels ou primitifs). L'idée intuitive de $\Diamond p$ est qu'il existe au moins un monde dans lequel p est vrai. Kripke relativise cette idée *à chaque monde*. Je désigne les mondes par des lettres grecques α, β, γ....

ξ désignera une (méta)-variable parcourant les mondes. L'idée de Kripke est d'introduire une relation d'accessibilité entre les mondes et d'interpréter $\Box p$ dans un monde α par le fait que p est vrai dans tous les mondes accessibles à partir de α.

Dans les schémas, les ronds représentent les mondes (les états). Les formules écrites à l'intérieur des mondes sont vraies dans ces mondes. La relation d'accessibilité est représentée par une flèche.



Remarquons que $\Box p$ est équivalent à $\neg \Diamond \neg p$, en particulier la proposition $\Box p$ est toujours vraie (et $\Diamond p$ est toujours fausse) dans un monde duquel ne part aucune flèche. Un tel monde ou état est appelé un *dernier monde* ou un *dernier état*.

La figure A.1 illustre, par exemple, avec la sémantique de Kripke, le carré Aristotélicien sur lequel Aristote distinguait le contraire de la négation avec les modalités (ontiques) nécessaire et possible:

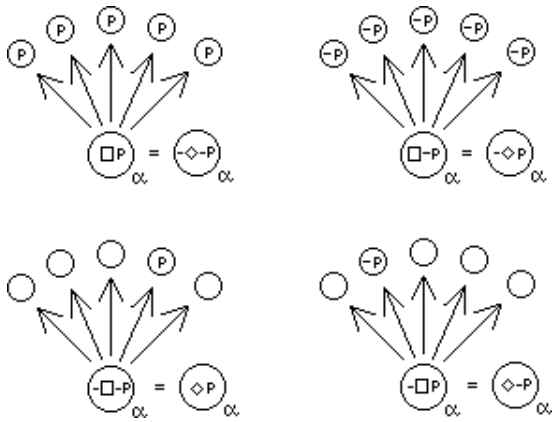


Figure A.1: Le carré Aristotélicien

Un référentiel (W, R) est un ensemble W dont les éléments sont appelés mondes ou états, muni d'une relation binaire R, appelée relation d'accessibilité.

Un modèle (W, R, V) est obtenu lorsqu'est assignée dans chaque monde une valeur, vrai ou faux, pour les variables propositionnelles $p, q, r, ...$ Si L désigne le sous-ensemble $p, q, r, ...$ de l'alphabet, l'assignation est capturée par une fonction V de $L \times W$ dans $\{vrai, faux\}$.

Chaque monde est supposé obéir à la logique classique, si bien que V définit une valuation booléenne pour chaque monde. Cela signifie que si la proposition p est vraie dans un monde α, et si q est vraie dans α, alors $p \wedge q$ est vraie dans α, etc. Je rappelle que $p \rightarrow q$ est classiquement vraie si p est fausse ou si q est vraie (ou encore si $(p \wedge \neg q)$ est fausse). $\bot$ est une constante propositionnelle désignant, dans chaque monde le faux, et $\top$ est une constante propositionnelle désignant dans chaque monde le vrai.

Résumons l'idée de Kripke on a:

$\Box A$ est vrai dans α ssi pour tout monde ξ tel que $\alpha R \xi$, A est vrai dans ξ.

De même:

$\Diamond A$ est vrai dans α ssi il existe un monde ξ tel que $\alpha R \xi$ et A est vrai dans ξ.

$\alpha R\xi$ est lu α accède à ξ, ou encore ξ est accessible à partir de α.

Remarque: En logique classique non modale, la valeur de vérité d'une formule est univoquement déterminée par la valeur des sous-formules, et donc par la valeur des variables propositionnelles. Ce n'est plus le cas en logique modale. La valeur de vérité de $\Box p$, dans un monde α, ne dépend pas, a priori, de la valeur de vérité de p dans α, comme on le voit dans la figure A.2.

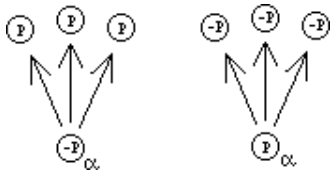


Figure A.2: La logique modale n'est pas vérifonctionnelle

Il n'y a donc pas moyen d'utiliser une table de vérité pour évaluer une formule modale à partir des valeurs de ses variables propositionnelles.

Définition fondamentale. Nous savons qu'en logique classique une tautologie est une formule qui est vraie quelle que soit la valuation de ses variables propositionnelles. Ainsi $p \to p,\ p \vee \neg p,\ p \to \top,\ \bot \to p$, sont des tautologies classiques. On dira qu'un référentiel (W, R) **respecte** une formule A si, quelle que soit la valuation V que l'on pourrait choisir, A est vrai dans tous les mondes de W. Dit autrement : un référentiel (W, R) respecte une formule A si et seulement si A est vraie dans tous les mondes dans tous les modèles que l'on peut construire sur le référentiel.

Conséquences:

1. Tous les référentiels respectent les tautologies classiques non modales puisque la logique classique est valable dans tous les mondes.
2. Pour la même raison tous les référentiels respectent les tautologies classiques dans lesquelles on a substitué les variables propositionnelles par des formules quelconques, comme $\Box p \to \Box p, \Diamond p \vee \neg \Diamond p$, etc.
3. $(\Box(p \to q) \wedge \Box p) \to \Box q$ est vrai dans tous les mondes de tous les modèles.

Preuve. Supposons que $(\Box(p \to q) \wedge \Box p)$ soit vraie dans un monde α. Alors $\Box(p \to q)$ et $\Box p$ sont chacune vraie dans α (par la sémantique classique (*tarskienne*) du $\wedge$).

Par Kripke: $\Box p$ est vraie dans α, si p est vraie dans tous les mondes accessibles à partir de a (et donc a fortiori s'il n'en existe pas), de même $\Box(p \to q)$ est vraie dans α si $(p \to q)$ est vraie dans tous les mondes accessibles à partir de α.

Mais chaque monde respecte la logique classique, et donc q est vraie dans tous les mondes auxquels accède α.

Mais si q est vraie dans tous les mondes accessibles à partir de α, alors, par Kripke, $\Box q$ est vraie dans α. Donc, dans un monde α quelconque, $\Box q$ ne peut pas être fausse en même temps que $(\Box(p \to q) \wedge \Box p)$ soit vraie. Donc $(\Box(p \to q) \wedge \Box p) \to \Box q$ est vraie dans tous les mondes, quelles que soient les valeurs de p et q.

Conclusion. La formule $(\Box(p \to q) \wedge \Box p) \to \Box q$, ou plutôt la formule (tautologiquement) équivalente $\Box(p \to q) \to (\Box p \to \Box q)$ est respectée par tous les référentiels. On la désigne par K (pour Kripke).

$$\Box(p \to q) \to (\Box p \to \Box q) \quad \text{K}$$

Toutes les formules ne sont pas respectées par tous les référentiels. Beaucoup de formules sont cependant respectées par une classe de référentiels caractérisée par la relation binaire. C'est là que réside l'intérêt de la sémantique de Kripke: associer un type de référentiel (c'est-à-dire une relation binaire) à une formule modale. Cela permet, par exemple, de rapidement réaliser l'indépendance sémantique de nombreuses formules modales. En particulier il n'est pas difficile de démontrer:

Proposition 15 (W, R) *respecte* $\Box p \to \Box\Box p$ *ssi* R *est transitive.*

Je rappelle qu'une relation binaire définie sur un ensemble E est transitive si xRy et yRz entraîne xRz, pour x, y, z appartenant à E. Les référentiels transitifs caractérisent ainsi la formule modale $\Box p \to \Box\Box p$. Un relation sur E est réflexive si xRx, et symétrique si xRy entraîne yRx, avec toujours x, y quelconques appartenant à E.
On a:

Proposition 16 (W, R) *respecte* $\Box p \to p$ *ssi* R *est réflexive.*

Proposition 17 (W, R) *respecte* $p \to \Box\Diamond p$ *ssi* R *est symétrique.*

Les référentiels réflexifs caractérisent ainsi la formule modale $\Box p \to p$, et les référentiels symétriques caractérisent la formule modale $p \to \Box\Diamond p$. Et, on est assuré de l'indépendance logique des formules T (c'est le nom "officiel de $\Box p \to p$), 4 (c'est le nom "officiel" de $\Box p \to \Box\Box p$) et B (c'est le nom "officiel" de $p \to \Box\Diamond p$).

Définition. Un référentiel (W, R) est *idéal* s'il ne possède pas de dernier monde. Je dirai simplement que R est idéale sur W, ou encore que R est idéale.

Définition. Un monde est transitoire ssi il n'est pas un dernier monde.

Évidemment, un référentiel est idéal si et seulement si tous ses mondes sont transitoires.

Exemples:

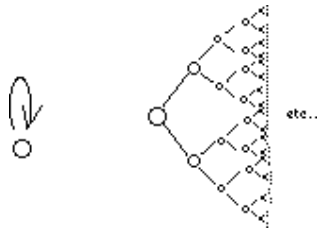


Figure A.3: la boucle et l'éventail sont idéales

On a:

Proposition 18 (W, R) *respecte* $\Box p \to \Diamond p$ *ssi* R *est idéale.*

La formule "$\Box p \rightarrow \Diamond p$ est appelée D. J'appelle C la formule "$\Diamond p \rightarrow \neg\Box\Diamond p$"

Définition. Un référentiel (W, R) est réaliste si pour tout monde transitoire α appartenant à W il existe un dernier monde δ accessible à partir de α.

Si on interprète un dernier monde (état) comme un décès, on voit qu'à la différence d'un référentiel idéal où "l'immortalité" est en quelque sorte garantie, dans un référentiel réaliste, bien que l'immortalité est possible (si le référentiel est infini ou possède une boucle) elle n'est jamais garantie. Partout on peut emprunter un chemin (une flèche, ou une suite de flèches si la relation est transitive) qui aboutit à un dernier monde.

Théorème 19 *(W, R) respecte $\Diamond p \rightarrow \neg\Box\Diamond p$ ssi R est réaliste.*

On peut trouver la preuve de cette proposition (et des propositions du même genre, voir aussi (Boolos, 1979; Boolos, 1993; Chellas, 1980)) dans le rapport technique (Marchal, 1995).

A.2 Théories et démonstrations

Jusqu'à présent, nous avons un langage permettant de construire des formules modales et nous avons une structure géométrique (le référentiel) permettant la caractérisation de certaines d'entre elles. Nous savons par exemple construire des modèles validant $\Box p \rightarrow p$, $\Box p \rightarrow \Box\Box p$, $(\Box p \rightarrow p) \wedge (p \rightarrow \Box\Diamond p)$, ainsi que des contre-exemples justifiant l'indépendance sémantique de formules. Par exemple encore, on peut trouver une valuation sur un référentiel transitif et non réflexif validant $\Box p \rightarrow \Box\Box p$ et possédant un monde où $\Box p \rightarrow p$ est fausse. Je définis à présent une classe de *théorie* modale, et une notion de démonstration qui se comporte convenablement (de façon correcte et complète) relativement à la sémantique de Kripke. Grâce au résultat de complétude et de correction (soundness) les indépendances sémantiques (faciles à visualiser) démontrent automatiquement les indépendances syntactiques. Plus loin j'exhibe une sémantique plus générale attribuée à Scott et Montague (Chellas, 1980).

Définition Une théorie (formelle) est présentée par la donnée d'un ensemble de formules accompagnée de règles dite règles d'inférence, permettant la déduction de nouvelles formules. Une théorie est un ensemble de formules fermé pour l'application des règles d'inférence.

Définition La présentation d'une théorie est l'ensemble des axiomes choisis et des règles d'inférence. Une théorie peut avoir de nombreuses présentations.

Comme la logique modale est une extension de la logique propositionnelle classique, je présente d'abord une formalisation de celle-ci. Je choisis celle de (Kleene, 1952). L'avantage de la formalisation de Kleene est que deux parmi ses "affaiblissements" possibles donnent une formalisation pour la logique quantique et la logique intuitioniste respectivement. A, B, C sont des métavariables représentant des formules quelconques. Il s'agit donc de schémas d'axiomes. J'utilise des schémas d'axiomes pour ne pas avoir de règles de substitution difficiles et longues à énoncer:

AXIOMES :	$A \to (B \to A)$	principe de l'a posteriori
	$(A \to B) \to ((A \to (B \to C)) \to (A \to C))$	
	$A \to (B \to A \wedge B)$	
	$(A \to C) \to ((B \to C) \to (A \vee B \to C))$	
	$A \to (A \vee B)$	
	$B \to A \vee B$	
	$(A \wedge B) \to A$	
	$(A \wedge B) \to B$	
	$(A \to B) \to ((A \to \neg B) \to \neg A)$	
	$A \vee \neg A$	principe du tiers-exclu
RÈGLES :	$\frac{A\ ,\ A \to B}{B}$	Modus Ponens

Logiques propositionnelles faibles

On aura l'occasion de s'intéresser à quelques logiques plus faibles, du point de vue déductif (et donc plus puissantes du point de vue sémantique), que la logique classique. Il s'agit essentiellement de la logique intuitioniste et de la logique quantique. La logique intuitioniste (formalisée par Heyting, voir Troelstra & Van Dalen 1988) peut être obtenue en remplaçant le principe du tiers-exclu par un principe plus faible: $\neg p \to (p \to q)$. La logique quantique peut de même être obtenue en affaiblissant le principe de l'a posteriori (voir annexe C). Gödel (1933) et Goldblatt (1974) ont respectivement montré comment traduire la logique intuitioniste avec S4 (KT4), et la logique quantique avec B (KTB).

Les logiques modales "Kripke-convenables"

Les logiques modales qui ont une sémantique de Kripke, possèdent comme axiomes et règles, outre les axiomes propositionnels de Kleene et la règle du modus ponens MP, l'axiome K et la règle de nécessitation $\frac{p}{\Box p}$. Un théorème est soit un axiome, soit une formule obtenue par un nombre fini d'applications de la règle du modus ponens, ou de la règle de nécessitation, à partir d'axiomes ou de théorèmes déjà démontrés. Cela va résulter des propositions qui vont suivre et du fait (qu'on a déjà vu) que K est respecté par tous les référentiels.

Les principaux systèmes considérés sont K, KT, KT4 (appelé S4), KB, KD, KC, etc. où XYZ représente ici des logiques modales avec X, Y, Z comme axiomes, et sont fermées pour la modus ponens et la nécessitation.

Notons que la "fermeture pour le $\Box$", la nécessitation, est une caractéristique de la sémantique de Kripke. On rencontrera des situations où la nécessitation, n'est pas vérifiée, et où, donc, d'autres sémantiques, comme celle plus générale des modèles minimaux (de Scott et Montague selon Chellas 1980) doivent être utilisées. On rencontrera des systèmes qui n'ont ni sémantique de Kripke, ni Scott-Montague (G^* par exemple).

Proposition 20 *Si un référentiel respecte un (ou plusieur axiomes) alors il respecte les théorèmes que l'on obtient avec ce (ou ces) axiomes par application des règles d'inférence du modus ponens et de la nécessitation.*

Preuve. Le modus ponens est vérifié dans chaque monde. En effet, chaque monde vérifie la logique classique. Donc si A est vraie dans tous les mondes, et si $A \to B$ est vraie dans tous les mondes, alors B est vraie dans tous les mondes.

Les référentiels respectent les formules obtenues par nécessitation sur des formules déjà respectées. En effet si A est vraie dans tous les mondes d'un référentiel W, alors en particulier, pour chaque monde ξ, A est vraie dans tous les mondes accessibles à partir de ξ, donc $\Box A$ est vraie dans tous les mondes ξ du référentiel, et donc $\Box A$ est respectée aussi. En

particulier si une proposition A est respectée par une classe de référentiel, $\Box A$ est respectée par cette classe de référentiel.

Donc, le respect dans un référentiel est fermé pour la nécessitation et le modus ponens, et cela entraîne que tous les théorèmes de K (KT, K4, etc.) sont respectés dans un référentiel quelconque (réflexif. transitif, etc.) L'inverse est aussi vrai.

Proposition 21 *Toutes les formules respectées par les référentiels quelconques (réflexifs, transitifs, idéaux, symétriques, réalistes, etc.) sont démontrables dans* K (KT, K4, KD, KB, KC *etc.).*

Idée de la preuve. La technique classique consiste à construire un modèle canonique qui satisfait tous les théorèmes de K (KT, K4, etc.), et seulement les théorèmes de K (KT, K4, etc.), et à montrer que ce modèle appartient à la classe de référentiels appropriée, c'est-à-dire quelconque (réflexif, transitif, etc.). Dans ce cas, si une formule est respectée dans tous les référentiels de la classe appropriée, elle est d'office satisfaite dans le modèle canonique et est donc un théorème de la théorie correspondante. Les logiciens depuis longtemps travaillent avec des modèles dont les mondes ou les univers sont des ensembles de formules, comme les modèles de Herbrand où les structures libres des algébristes. Un monde est défini par un ensemble consistant maximal de formules. Consistant est défini de façon purement syntaxique. Pour K (KT, K4, etc.), un ensemble de formules est consistant s'il n'existe pas de formule $F_1, ..., F_n$ tel que K (KT, K4, etc.) démontre $\neg(F_1 \wedge ... \wedge F_n)$. Si E est un ensemble consistant de formules, on peut montrer que:

1. $E \cup \{F\}$ est consistant ou $E \cup \{\neg F\}$ est consistant,
2. E est contenu dans un ensemble consistant maximal, c'est-à-dire un ensemble consistant qui contient toute formule F ou sa négation $\neg F$.

Pour définir le modèle canonique, on doit définir d'une part le référentiel, c'est-à-dire l'ensemble W des mondes et la relation R d'accessibilité, et d'autre part la valuation V. On prend pour W, l'ensemble des mondes du modèle canonique, l'ensemble des ensembles maximaux consistants. La relation d'accessibilité R du modèle canonique est définie par

$$aRb \text{ ssi } \{F \mid \Box F \in \alpha\} \subseteq \beta$$

On peut alors montrer, c'est le point délicat, qu'on a bien pour chaque monde $\alpha \in W$ que si F appartient à tous les mondes β accessibles à partir de α, alors $\Box F$ appartient au monde α. Reste à définir la valuation V: une formule atomique p est *vraie* dans un monde si elle *appartient* à ce monde. Il n'est plus difficile alors de prouver, par induction sur la complexité des formules F, que F est vraie dans le monde α ssi F appartient à α.

Il ne reste plus alors qu'a montrer que les modèles canoniques de KT, KT4, ainsi de suite sont réflexifs, transitifs etc. Cela ne pose aucune difficulté (voir par exemple Boolos 79).

Remarque Avec la logique modale, le métathéorème de déduction n'est pas valide dans les dérivations qui utilise la nécessitation. On ne peut pas, comme on le fait usuellement, décharger p dans $\frac{p}{\Box p}$, pour prétendre, ayant obtenu $\frac{}{p \rightarrow \Box p}$, avoir démontrer $p \rightarrow \Box p$. On a $\frac{p}{\Box p}$ (nécessitation), mais $p \rightarrow \Box p$ est déjà contredit par le monde α dans le modèle de la figure A.4.

Tant qu'on utilise pas la nécessitation, la procédure de déduction habituelle est bénigne.

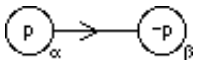


Figure A.4: Contre-exemple au métathéorème de déduction

A.3 La sémantique de Scott et Montague

Définition(voir Chellas 1980)

- La règle d'inférence RE est la suivante: $\frac{A \leftrightarrow B}{\Box A \leftrightarrow \Box B}$.
- Une logique modale est dite classique minimale si
  1. ses théorèmes sont fermés pour la règle RE
  2. le schéma $\Box A \leftrightarrow \neg\Diamond\neg A$ est vérifié.

Les logiques classiques minimales admettent une sémantique dite des voisinages ou encore sémantique de Scott-Montague (Chellas 1980).

Un modèle de Scott-Montague, appelé aussi *modèle minimal*, $(W, \mathcal{N}, V)$, est la donnée d'un ensemble de mondes W, d'une fonction $\mathcal{N}$ de W dans 2^{2^W}, qui associe à chaque monde α un ensemble d'ensembles de mondes, $\mathcal{N}(\alpha)$ appelé système de voisinage de α. De nouveau, chaque monde satisfait la logique propositionnelle classique.

Définition (voir Chellas 1980) $\Box A$ est vrai dans un monde α si l'ensemble des mondes (de W) où A est vrai, noté $[A]$, appartient à $\mathcal{N}(\alpha)$:

- $\alpha \models \Box A \Leftrightarrow [A] \in \mathcal{N}(\alpha)$,
  de même on exige, pour avoir $\neg\Box\neg = \Diamond$
- $\alpha \models \Diamond A \Leftrightarrow W \setminus [A] \notin \mathcal{N}(\alpha)$.

A est satisfaite par un modèle si A est vraie dans tous les mondes du modèle.

Comme $\top$ est satisfaite dans tous les mondes, il suffit pour illustrer la perte de nécessitation d'un modèle minimal où $\Box\top$ n'est pas satisfait:

$$\alpha \not\models \Box\top \text{ ssi } [\top] \notin \mathcal{N}(\alpha).$$

Comme $[\top] = W$, il suffit que W n'appartienne pas à $\mathcal{N}(\alpha)$. Par exemple, $W = \{1, 2\}$ et $N(1) = N(2) = \{\}$.

Définition. A est C-valide si A est vrai dans tous les mondes d'une classe C de modèles minimaux.

Cela donne une sémantique pour la règle d'inférence. On a $\frac{A}{B}$ si la C-validité de A entraîne la C-validité de B. Je renvoie à Chellas 1980 pour plus d'information. Je vais me contenter de montrer que la règle RE est toujours vérifiée avec les modèles de Scott-Montague.

Proposition 22 *La règle* $\frac{A \leftrightarrow B}{\Box A \leftrightarrow \Box B}$ *est valide pour les modèles minimaux.*

Preuve. Si $A \leftrightarrow B$ est C-valide, $A \leftrightarrow B$ est vraie dans tous les mondes des modèles de la classe C. Du coup, dans tous ces modèles $[A] = [B]$, mais alors quel que soit α, monde d'un de ces modèles, $[A] \in \mathcal{N}(\alpha)$ ssi $[B] \in \mathcal{N}(\alpha)$, et donc pour tout $\alpha, \alpha \models \Box A$ ssi $\alpha \models \Box B$ et, par calcul propositionnel classique dans α, on a $\alpha \models \Box A \leftrightarrow \Box B$.

Scott-Montague généralise Kripke

J'ai dit plus haut que la sémantique de Scott-Montague était plus générale que la sémantique de Kripke. On peut préciser cette assertion au moyen d'un résultat remarquable (voir Chellas 1980, Segeberg 1973) qui permet d'associer à tout modèle de Kripke un modèle minimal, et inversément à tout modèle minimal "augmenté", un modèle de Kripke. Un modèle minimal est augmenté si on a, pour tout monde α lui appartenant :

- $X \subseteq Y$ et $X \in \mathcal{N}(\alpha)$ entraîne $Y \in \mathcal{N}(\alpha)$,
- $\cap\mathcal{N}(\alpha) \in \mathcal{N}(\alpha)$.

Etant donné un modèle de Kripke, on lui associe un modèle standard sémantiquement équivalent (les mêmes mondes vérifient les mêmes formules) en définissant la fonction $\mathcal{N}$ par

$$X \in \mathcal{N}(\alpha) \text{ ssi } \{\beta \in W \mid \alpha R\beta\} \subseteq X$$

Et inversément, étant donné un modèle minimal augmenté, on lui associe un modèle de Kripke sémantiquement équivalent en définissant la relation d'accessibilité, pour tous les mondes α et β, ainsi :

$$\alpha R\beta \text{ ssi } \beta \in \cap\mathcal{N}(\alpha)$$

Pour la démonstration on consultera Chellas 1980.

Annexe B

La thèse de Church

B.1 Généralité et histoire

Church voulait définir formellement le concept (pré-théorique, philosophique, intuitif) de fonctions calculables. Une fonction définie sur les nombres naturels et à valeurs dans les nombres naturels, est considérée comme étant calculable, si on peut *en principe* la calculer, c'est-à-dire s'il existe une règle permettant de la calculer. Cette règle doit être publiquement communicable en un temps fini. Il est important de comprendre que cette "définition" informelle est non-constructive. On demande que la règle existe en principe. On ne demande pas que la règle soit explicitement donnée. En particulier l'existence de la règle ne doit pas être explicite. Les fonctions sont considérées comme étant des objets mathématiques définis extensionnellement (par leur ensemble de couples). Par exemple la fonction constante F_0 qui envoie tout nombre sur 0 :

$$\{(0,0)(1,0)(2,0)(3,0)(4,0),(5,0),...\},$$

est trivialement calculable. Et, de même, la fonction F_1 qui envoie tout sur 1 :

$$\{(0,1)(1,1)(2,1)(3,1)(4,1),(5,1),...\},$$

est calculable. Donc la fonction F suivante, présentée intensionnellement de la façon suivante :

$$\begin{aligned} F(x) &= 1 \text{ si la conjecture de Goldbach est vraie} \\ &= 0 \text{ sinon.} \end{aligned}$$

est *aussi* calculable, malgré qu'à partir de cette présentation (ou "intension" comme on dit dans le domaine) personne ne sait la calculer. F est calculable, car, extensionnellement parlant F est égale à F_0, ou F est égale à F_1. La règle pour la calculer existe, mais à l'heure actuelle personne ne la connaît.

On a déjà rencontré une situation similaire avec la définition du computationnalisme où on exige l'existence du niveau de substitution sans exiger qu'on puisse définir explicitement ce niveau.

Pour définir mathématiquement la notion de fonctions calculables, il suffit donc de définir de façon précise ce qu'on entend par *règle publiquement communicable en un temps fini*.

Church parvint à une définition de fonction calculable avec la notion formelle de fonctions lambda-définissables. A présent, on peut démontrer qu'une fonction est lambda-définissable

si et seulement si elle est programmable (calculable par un ordinateur). Une version moderne de la thèse de Church est alors donnée par :

Une fonction de N dans N est calculable ssi elle est programmable.

Conceptuellement cette thèse est très forte: elle entraîne le théorème d'incomplétude de Gödel, ce que j'illustre plus bas, et de nombreux autres phénomènes d'incomplétude. Autrement dit : si le théorème de Gödel, avait été faux, la thèse de Church serait elle-même fausse. Comme le théorème de Gödel étaient tout à fait inattendu dans la communauté des mathématiciens (à l'exception des précurseurs comme Post et Turing, mais aussi Markov), on peut dire que la thèse de Chuch fut tout autant inattendue.

Historiquement, c'est Post, en 1922, qui est le premier à proposer la "thèse de Church". Pour lui elle constituait une loi naturelle de l'esprit, et elle devait être justifiée par des considérations psychologiques. Il est aussi le premier à dériver de cette thèse le phénomène d'incomplétude. Il est encore le premier à "réfuter" l'hypothèse du mécanisme avec l'incomplétude (comme Lucas et Penrose par exemple), et il est encore le premier à réaliser que cette réfutation est incorrecte (à la différence de Lucas et Penrose). Le travail de Post date de 1922, sera proposé et refusé à la publication en 1943. Il sera finalement publié en 1965 après sa mort (en 1957), par Davis (Post, 1922; Davis, 1965).

Avant que Gödel ne démontre en 1931 son résultat d'incomplétude, Hilbert espérait qu'un formalisme assez puissant allait pouvoir sécuriser formellement les fondements des mathématiques. Russell et Whitehead s'était attaqué à cette tâche avec Principia Mathematica. Le théorème de Gödel a ruiné cet espoir.

La thèse est généralement attribuée à Church, mais pour Church, c'était une définition. Kleene a estimé un temps cette "définition" inadéquate. C'est en échouant dans la critique (voir plus loin) de cette "définition" que Kleene a créé le vocable "thèse de Church" et qu'il en est devenu un des grands partisans. En fait Kleene est le premier à réaliser la nature essentiellement hypothétique de cette "définition". Il devint aussi le premier chaleureux partisan de cette thèse. On devrait donc parler plutôt de *thèse* de Kleene. (voir (Kleene, 1952)).

B.2 La thèse de Church entraîne l'incomplétude de Gödel

Une théorie formelle est la donnée d'un langage formel, d'un ensemble d'axiomes et d'un ensemble de règles d'inférence. On exige d'une théorie formelle les choses suivantes. On peut reconnaître mécaniquement si une formule du langage est un axiome. Et on peut vérifier mécaniquement si une formule est un théorème, c'est-à-dire est une formule du langage dérivable en un nombre fini d'applications des règles d'inférence à partir des axiomes. Si en plus on peut vérifier qu'une formule fermée n'est pas un théorème on dit que la théorie est complète. Une théorie est saine si elle ne prouve que des formules (de l'arithmétique par exemple) qui sont vraies. Le résultat de Gödel est qu'une telle théorie, saine et complète pour l'arithmétique, n'existe pas. Ceci constitue une généralisation du théorème de Gödel à partir de la thèse de Church:

Proposition 23 *$TC \Rightarrow$ incomplétude des systèmes formels riches et consistants*

Preuve. En effet, avec la thèse de Church, l'ensemble des fonctions (partielles et totales) calculables est identique à l'ensemble des fonctions programmables, par exemple en Fortran (pour fixer les idées). Considérons une énumération des fonctions de N dans N, programmables (en Fortran ou tout autre langage assez riche ce qui revient au même avec

la thèse de Church). Une telle énumération est obtenue en placant les programmes Fortran disposant d'une seule entrée par ordre de longueur. Si plusieurs programmes ont la même longueur, on range ceux-ci par ordre "alphabétique" (en décidant préalablement de placer un ordre arbitraire sur les touches du clavier).

Voilà l'énumération :

$$\varphi_1(x), \varphi_2(x), \varphi_3(x), \varphi_4(x), \varphi_5(x), \varphi_6(x), \varphi_7(x), ...$$

Etant donné le caractère numériquement codables des programmes, il est possible de traduire les propositions du genre $\exists y\, \varphi(x) = y, \neg\, \exists y\, \varphi(x) = y, ...$ en proposition purement arithmétique.

Etant donné le caractère mécanique de la génération des programmes, l'existence d'une théorie formelle saine et complète TFSC, rend la fonction (de 2 entrées) suivante F_{TFSC} intuitivement totale calculable :

$$F_{TFSC}(z,x) = \begin{cases} y & \text{si } \varphi_z(x) = y \\ 0 & \text{si TFSC prouve } \neg\, \exists\, y\ \varphi_z(x) = y \end{cases}$$

Comme $F_{TFSC}(z,x)$ est intuitivement calculable, on peut construire une énumération de toutes les fonctions *totales* calculables suivantes. Avec la thèse de Church, cette énumération contient toutes les fonctions totales calculables.

$$\varphi_1^*(x), \varphi_2^*(x), \varphi_3^*(x), \varphi_4^*(x), \varphi_5^*(x), \varphi_6^*(x), \varphi_7^*(x), ...$$

Mais dans ce cas la fonction *diagonale*, à une entrée, $G_{TFSC}(x) = \varphi_x^*(x) + 1$, est totale calculable, et elle est différente de toutes les fonctions φ_i. G_{TFSC} est totale et intuitivement calculable, mais est différente de toutes les fonctions fortran programmables, en contradiction avec la thèse de Church. La négation du théorème de Gödel réfute la thèse de Church. Donc la thèse de Church entraîne l'incomplétude de Gödel. Ce qu'il fallait démontrer.

Bien sûr, dans *un certain sens* G_{TFSC} *est* fortran programmable : il est possible de décrire l'algorithme présenté plus haut en fortran. Mais, et c'est ce que nous avons montré, la fonction décrite par cette procédure est nécessairement *partielle*. Pour certaine valeur de x $G_{TFSC}(x)$ diverge, et cela de façon non prouvable par la théorie TFSC. On pourrait trouver une théorie plus puissante TFSC′, capable de décider si G_{TFSC} diverge, mais elle sera incapable de décider, pour certaine valeur de x, si la fonction correspondante $G_{TFSC'}(x)$ diverge. La thèse de Church rend absolue la notion de calculabilité, et relative la notion de prouvabilité.

Il est raisonnable de penser que si cette preuve avait été présentée à Russell et Whitehead, ils n'auraient pas été convaincus. Ils auraient commencé par farouchement critiquer la thèse de Church, et se seraient sans aucun doute attelé à réfuter cette thèse avec Principia Mathematica. Ce faisant, ils seraient vraisemblablement arrivés au théorème d'incomplétude de Gödel.

De cette façon, le théorème de Gödel (1931) *confirme* la thèse de Church, alias la loi de Post (1922).

La thèse de Church confirme à son tour, et rend non triviale, l'hypothèse du mécanisme digitale.

En effet, si la thèse de Church était fausse il pourrait exister des machines digitales capables de surpasser (en quantité de fonctions calculables) les ordinateurs. Le théorème de Gödel ne s'appliquerait pas a priori à de telles machines. Avec TC, le théorème de Gödel s'applique à toutes les machines digitales (suffisamment riches, c'est-à-dire capable de démontrer les théorèmes de l'arithmétique élémentaire). La thèse de Church donne donc apparemment un espoir aux non-mécanistes qui voudraient réfuter le mécanisme digital: il suffit de parvenir à montrer que le théorème de Gödel ne s'applique pas à nous, en exhibant

par exemple une preuve informelle d'une vérité non machine-accessible. Mais avec TC, une telle preuve ne peut pas être effective, sinon elle est machine accessible. Avec TC et le mécanisme, le théorème de Gödel s'applique aux machines *et* s'applique à nous. La thèse de Church alliée au mécanisme fait des résultats d'incomplétude les prémisses d'une psychologie exacte (voir aussi le rapport technique Marchal 1995, et Myhill 1952).

Judson Webb (1980) résume cette situation en affirmant que la thèse de Church est l'ange gardien du mécanisme. Il montre que la thèse de Church protège le mécanisme de toute vision réductionniste du monde des machines et qu'elle rend les réfutations du mécanisme (ou de la thèse de Church) presqu'automatiquement non-communicable à une troisième personne.

Gödel, malgré son théorème de 1931, ne croira pas à d'emblée à la thèse de Church. Après la publication de l'article de Turing (voir (Davis, 1965)), il admettra la thèse de Church, et il estimera que la thèse de Church est une sorte de miracle épistémologique. Pour la première fois une notion métamathématique semble ne pas dépendre du système formel choisi. On dit, en informatique théorique, qu'une telle notion est "machine indépendante". Cela introduit en quelque sorte une notion objective d'objectivité : une notion est objective si elle est valable pour toutes les machines universelles. C'est cette notion de machine-indépendance qui nous permet d'espérer isoler une mesure objective, machine universelle-indépendante, sur la collection entière des états computationnels. En absence du "miracle de Gödel" le présent travail n'aurait aucun sens.

Indépendamment de Post et de Church, Turing propose une thèse équivalente et l'appuie sur une analyse abstraite de la notion de calculateur humain. Gödel ne verra pas le bénéfice que la philosophie mécaniste peut tirer de la thèse de Church. En fait Gödel défendra une philosophie plutôt non-mécaniste (Wang, 1974; Marchal, 1990; Marchal, 1995). Il montrera par ailleurs que la thèse de Church entraîne l'existence de propositions *absolument* indécidables pour l'esprit humain, ce qui ne sembla pas lui plaire . En fait Gödel est piégé : il se rend compte que la thèse de Church augmente considérablement la portée de son théorème d'incomplétude, mais il n'apprécie pas l'idée que la portée du théorème soit agrandie au point de s'appliquer à nous.

Aujourd'hui la thèse de Church est acceptée de façon quasi-unanime par les (méta)mathématiciens et les philosophes concernés, mais cela ne doit en aucune façon nous faire oublier le caractère révolutionnaire, miraculeux de cette thèse.

On définit habituellement la machine universelle de Turing comme étant une machine de Turing capable d'émuler (simuler parfaitement, aux temps d'éxécutions près) n'importe quelle machine *de Turing*. On peut cependant démontrer que la machine de Turing est à même d'émuler n'importe quel programme fortran ou n'importe quel ordinateur quantique, etc. Avec la thèse de Church, la machine universelle de Turing peut émuler n'importe quelle machine digitale. C'est la thèse de Church qui permet de supprimer le qualificatif "de Turing" pour la machine universelle. La thèse de Church rend ainsi la machine universelle *vraiment* universelle, et elle rend de la même façon le déployeur universel vraiment universel. En fait, c'est la notion même d'universalité qui est rendue machine-indépendante, et donc épistémologiquement absolue, au sens de Gödel (Gödel, 1946).

La thèse de Church met en outre en évidence une nécessaire activité capitale de la machine consistante universelle : rester silencieuse lorqu'on lui pose certaines questions, ou devenir inconsistante, si elle veut avoir réponse à tout! Cela joue un rôle récurrant pour la dérivation des phénoménologies de l'esprit et de la matière.

Remarque La thèse de Church n'est pas une proposition mathématique. Elle appartient au domaine de la philosophie des mathématiques. Elle est scientifique cependant, dans le sens qu'elle est en principe réfutable et confirmable. Par exemple, il suffirait de trouver une fonction "clairement" calculable qui ne soit pas programmable pour la réfuter. Le fait que la thèse de Church entraîne le théorème de Gödel illustre qu'une thèse philosophique peut avoir des conséquences purement mathématiques (ce qui par ailleurs illustre encore son

caractère réfutable et confirmable). Dans le présent travail l'hypothèse du computationnalisme joue un rôle analogue. Il s'agit d'une hypothèse à la fois philosophique et scientifique (en principe réfutable) qui permet de transformer le problème du corps et de l'esprit (réputé être un problème de philosophie) en un problème de pure mathématique (isoler une mesure unique satisfaisant à certaines contraintes sur l'ensemble des états computationnels).

B.3 La thèse de Church intensionnelle

Afin d'éviter une confusion, je précise tout de suite que la thèse de Church intensionnelle que je considère concerne toujours les fonctions pensées extensionnellement.

La thèse de Church extensionnelle habituelle TC peut être reformulée de la façon suivante :

> Toute machine universelle est capable de calculer toute fonction calculable par n'importe quelle machine universelle.

La thèse de Church intensionnelle TCI est apparemment plus forte. Elle énonce que

> Toute machine universelle est capable de calculer toute fonction calculable par n'importe quelle machine universelle, *et cela de la même façon*, c'est-à-dire avec le même algorithme (on fait abstraction du temps d'exécution).

La thèse de Church extensionnelle entraîne la thèse de Church intensionnelle.

Preuve (informelle). Cela découle directement de la discrétisation de la procédure de calcul. Une machine universelle travaille en effet par étapes successives. Le passage d'une étape à une autre est récursif, de même que la reconnaissance d'une condition d'arrêt, et donc ce passage et cette reconnaissance d'arrêt peuvent être codés "extensionnellement" par des fonctions récursives et sont donc calculables par n'importe quelle machine universelle.

Une autre façon de se convaincre de ce résultat est de réaliser que non seulement on peut écrire dans un langage de programmation L_A (c'est-à-dire le langage d'une machine universelle A) le code d'une machine universelle quelconque B, mais on peut écrire dans tout langage de programmation L_A un débogueur ou un traceur capable d'évaluer par étapes successives les L_B-programmes, en simulant les étapes successives de la machine universelle B. En particulier si une fonction F est calculée par un certain algorithme décrit dans L_B, alors un débogueur de L_B, écrit en L_A, permet à la machine A de calculer F de la même façon que le fait la machine B avec l'algorithme décrit par L_B.

B.4 La thèse de Church permet de réhabiliter une "philosophie de Pythagore"

Par "philosophie de Pythagore", ce que je désignerai par PP_0 ou simplement PP, j'entends l'idée selon laquelle *tout est nombre (naturel) ou rapport de nombres (naturels).* Après avoir découvert les rapports harmoniques, Pythagore aurait, en effet, postulé que, en particulier, toutes les grandeurs mesurables devraient pouvoir s'écrire sous forme de quotient de nombres naturels. Cette philosophie a été jugée réfutée après la découverte, par les Pythagoriciens eux-mêmes, que la longueur d'un coté d'un carré de surface 2 ne peut pas s'écrire sous forme d'un tel rapport.

Il n'empêche que cette longueur, de la diagonale de ce carré de surface 2, peut s'écrire, et s'écrit traditionnellement sous forme de racine : $\sqrt{2}$. Et l'extraction de la racine est une

opération louable du genre de celle qu'on apprend sur les bancs de l'école. On pourrait donc imaginer une nouvelle école pythagoricienne PP_1 selon laquelle tout est nombre ou rapport de nombres ou racine de nombre. Les mathématiciens savent cependant qu'une telle philosophie aurait été jugée réfutée au 19ième siècle lorsqu'Abel, en 1824, démontre que d'une façon générale les solutions d'une équation polynomiale de degré supérieur à 4 ne peuvent pas s'écrire sous forme de radicaux. De même il est bien connu qu'Hermite, en 1873, et Lindemann, en 1882 (voir Simmons 1992) démontreront que le nombre d'Euler e et le nombre π, dont les décimales sont pourtant calculables, ne peuvent pas être solution d'une équation polynomiale, et donc en particulier, n'admettent pas de descriptions sous forme radicales (en terme de racine, somme, soustraction et rapport). Mais pourquoi ne pas imaginer à nouveau une école pythagoricienne selon laquelle tout est nombre ou combinaisons de nombre, où les combinaisons comprennent les rapports, les racines, et le minimum indispensable pour calculer les solutions des équations polynomiales, ainsi que e et π ?

Notons qu'un nombre réel (positif) peut être codé par une fonction des naturels dans les naturels. Par exemple, ayant choisi une base, en envoyant 0 sur la partie entière, 1 sur la première décimale, 2 sur la 2ème, ainsi de suite. La question, alors, est de savoir, si la suite des écoles pythagoriciennes va converger, dans le sens qu'il existerait une école universelle PP_u capable de calculer toutes les fonctions calculables. Il existerait alors une collection finie d'opérations calculables, dont les combinaisons finies et descriptibles (communicables) permettraient de calculer toutes les fonctions calculables, et en particulier tous les réels calculables. Si on avait demandé s'il existe une école capable de calculer *tous* les réels, la réponse aurait été non. On connait l'argument par la diagonale de Cantor montrant que les réels sont indénombrables, alors que les fonctions de l'école PP_u sont clairement dénombrables puisqu'on peut énumérer les descriptions des combinaisons finiment communicables, combinaisons des opérations calculables admises comme primitive par cette école.

Curieusement l'argument de la diagonale semble pouvoir réfuter, non seulement l'existence d'une école capable de calculer tous les réels, mais semble pouvoir réfuter aussi l'existence d'une école capable de calculer toutes les fonctions calculables, ou tous les réels calculables. C'est avec l'argument de la diagonale que Kleene a en effet estimé, pendant un temps, pouvoir *réfuter la "définition" de Church*. L'échec de cette réfutation, autrement dit la fermeture de l'ensemble des fonctions calculables pour l'opération de diagonalisation, constitue la motivation conceptuelle[1] la plus profonde pour la thèse de Church. C'est en découvrant cette fermeture que Kleene est devenu un partisan zélé de la thèse de Church. Cela vaut la peine de regarder pourquoi la diagonalisation *semble* pouvoir réfuter la thèse de Church. On retrouvera par ailleurs le fait que la thèse entraîne l'incomplétude.

En effet, j'ai déjà mentionné le fait que les fonctions calculables par l'école PP_u sont énumérables. Enumérons-les donc:

$$f_0, f_1, f_2, f_3, f_4, f_5, f_6, f_7, ...$$

mais alors la fonction g définie au moyen de la *première diagonalisation*

$$g(n) = f_n(n) + 1$$

n'est pas calculable par l'école PP_u. En effet, si elle était calculable par l'école PP_u, elle appartiendrait à l'énumération. Il existerait alors un nombre k tel que $g = f_k$.

On pourrait procéder alors à une *deuxième diagonalisation*, en appliquant g sur l'indice k correspondant à sa description. On obtient

[1]La motivation est *conceptuelle* par opposition à la motivation *empirique* de base qui est que toutes les définitions de fonctions calculables inventées jusqu'à présent sont prouvablement équivalentes en ce qui concerne la classe des fonctions partielles calculables définies.

$$g(k) = f_k(k)$$

puisque $g = f_k$, et on obtient aussi

$$g(k) = f_k(k) + 1$$

par définition de k, et donc on a montré:

$$g(k) = g(k) + 1$$

Absurde? Aurions-nous réfuté l'universalité de PP_u, aurions-nous réfuté la thèse de Church?

En réalité, ce couple de diagonalisations ne réfute pas la thèse de Church, mais entraîne, avec et sans la thèse de Church, de nombreuses conséquences intéressantes.

Sans la thèse de Church Si les pythagoriciens de l'école PP_u prétendent que leurs procédures ne définit que des fonctions au sens habituel du terme, c'est-à-dire des fonctions totales, c'est-à-dire des fonctions *définies* sur *chaque* nombre naturel, alors chaque f_k est totale, et alors g_k est totale. Dans ce cas $g(k)$ ne peut pas être égale à $g(k) + 1$, g ne peut pas être égale à une f_i, et nous avons, effectivement, réfuté l'universalité de PP_u. On a démontré l'impossibilité d'énumérer la collection des fonctions totales calculables. La démonstration est constructive: étant donné une énumération de fonctions totales calculables, la diagonalisation permet de générer une nouvelle fonction totale calculable. La diagonalisation fait ici office de limite constructive permettant de générer une hiérarchie transfinie d'écoles pythagoriciennes de plus en plus larges.

Avec la thèse de Church Dans ce cas, une école universelle du style PP_u existe, et il en existe même beaucoup (FORTRAN, LISP, le jeu de la vie de Conway (Wainwright, 1974), le problème de n corps (Moore, 1990), l'ordinateur de von Neumann, l'ordinateur quantique de Deutsch (Deutsch, 1985), etc.). Que se passe-t-il alors avec l'énumération des fonctions f_i de PP_u? Comme on a démontré que $g(k) = g(k) + 1$, on a démontré que g n'est pas définie en k. La machine universelle qui calcule g sur k ne s'arrête pas. On a démontré que toute énumération de fonctions qui contient toutes les fonctions totales calculables contient nécessairement des fonctions partielles calculables, c'est-à-dire non partout définies. Nous sommes aussi arrivés à deux pas d'un phénomène d'incomplétude. En effet, aucune machine universelle, ni aucune théorie complète et axiomatisable ne peut nous permettre de distinguer uniformément les indices des fonctions totales calculables des indices des fonctions partielles calculables, puisque cela permettrait en parcourant les f_i d'extraire une sous-énumération complète des fonctions totales calculables. Mais alors la double diagonalisation s'ébranlerait à nouveau.

Notons *qu'avec la thèse de Church*, il est possible de préciser les conséquences, *sans la thèse de Church*, de cette double diagonalisation, en construisant, par exemple, des hiérarchies ω_1^{CK}-transfinies d'écoles pythagoriciennes, où ω_1^{CK} représente l'ordinal de Church et Kleene. Il est le plus petit ordinal non constructif, voir (Church and Kleene, 1937).

Avec la thèse de Church, la vieille philosophie de Pythagore est réhabilitable sous la forme "tout est nombre ou transformation universelle de nombre". Par *transformation universelle* d'un nombre, j'entends simplement le travail d'une machine universelle appliquée à ce nombre.

Le prix de la “réhabilitation” est qu'on ne peut pas prédire, d'une façon générale, le produit d'une transformation universelle, ni même s'il y en a un. On doit alors reconnaître le caractère universel et objectif (machine-indépendant), de l'incomplétude.

Annexe C

Mécanique quantique

Cette annexe, un peu vieillie, est tirée sans gros changement du rapport technique (Marchal 1995). Elle ne remplace pas la consultation d'ouvrages classiques, comme (d'Espagnat, 1971; Jammer, 1974). Trois ouvrages récents de mécanique quantique sont particulièrement pertinents pour notre approche : (Albert, 1992; Maudlin, 1994; Bub, 1997).

C.1 Le doute qui vient de la chimie

Comme dit Watson "la cellule obéit aux lois de la chimie", et les motivations mécanistes qui proviennent de la biologie moléculaire justifient peut-être une croyance dans un mécanisme indexical relatif aux lois de la chimie (Watson, 1968; Pauling, 1966). Si celles-ci se révélaient non-mécanisables, le mécanisme se verrait affaibli, peut-être réfuté, certainement relativisé. Cette suggestion est d'autant plus fondée que les lois de la chimie sont capturées par la mécanique quantique. Celle-ci, en dépit de son nom (mécanique est utilisé dans le sens Newtonien), attire les philosophes qui verraient dans les faits décrits (et jusqu'à présent confirmés) de cette théorie une justification empirique de la nature non-mécaniste du monde et/ou de la conscience.

Les arguments anti-mécanistes fondés sur la mécanique quantique sont variés. examinons-en brièvement quelques-uns :

a) Le plus ancien argument: la MQ met en évidence un indéterminisme intrinsèque dans le monde (ou plus précisément concernant les relations entre l'observateur et le monde). Le mécanisme est déterministe. Donc notre relation au monde n'est pas mécaniste. Ceux qui usent de cet argument sont tentés d'"expliquer" le libre-arbitre au moyen de cet indéterminisme. L'argument a déjà été réfuté par Carnap ou McKay ou Schrödinger. De plus, on a montré que le mécanisme n'est pas déterministe.

b) Le plus récent argument: la MQ rend possible des matériaux très particuliers, comme par exemple les quasi-cristaux de Penrose et Shechtman voir Penrose 1989). Penrose suggère, sans vraiment se convaincre lui-même semble-t-il, que le cerveau pourrait être une sorte de quasi-cristal. De même Margeneau 1984 et Squires 1990 cherchent à utiliser la MQ pour développer une théorie dualiste et non mécaniste de l'esprit (voir (Squires, 1990)).

Les arguments suivants méritent d'être examinés plus en détail parce que l'hypothèse mécaniste (indexicale) les éclaire considérablement. A cette fin j'expose le minimum qu'il faut avoir à l'esprit sur la mécanique quantique pour suivre l'argument.

Newton concevait la matière et la lumière comme constituées de particules interagissant les unes avec les autres. Huygens quant à lui réserve cette façon de voir pour la matière exclusivement. Il développe une théorie ondulatoire de la lumière qui rend compte avec succès de nombreux phénomènes lumineux. Einstein mettra en évidence, dans son travail

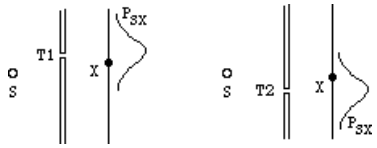
sur l'effet photo-électrique, un aspect corpusculaire de la lumière, sans détrôner pour autant la théorie ondulatoire. Il fonde ainsi la théorie quantique de la lumière. De Broglie étend cet aspect onde-corpuscule de la lumière à la matière. Cela permet de rendre compte du comportement des électrons dans les atomes décrits par Bohr et cela signifie la naissance de la théorie quantique de la matière. Le tableau suivant résume l'évolution du concept de lumière et de matière:

Newton	lumière	corpuscule
	matière	corpuscule
Huygens	lumière	onde
	matière	corpuscule
Einstein	lumière	corpuscule & onde
de Broglie	matière	onde & corpuscule

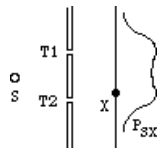
Born, à son tour, donne une interprétation probabiliste de l'onde accompagnant la particule. Il associe donc une grandeur qu'il appelle amplitude de probabilité à la particule. C'est cette amplitude qui oscille et est responsable de la présence des phénomènes quantiques d'interférence ondulatoire. A présent, lorsqu'on décide d'observer la position de la particule, le résultat sera une position bien précise, résultat prédit par l'amplitude de probabilité de la valeur de l'onde en cette position mise au carré.

Illustration : supposons qu'une particule ait le choix entre deux chemins pour aller d'une source S vers un point X d'un écran, en traversant par exemple une plaque comportant deux trous T_1 et T_2. Dans le cas classique, la probabilité P_{SX/T_1} (resp P_{SX/T_2}) d'arriver en X en passant par le trou $T1$ (resp T_2) est égale au produit des probabilités d'aller de S à T_1 (resp T_2) avec la probabilité d'aller de T_1 (resp T_2) à X:

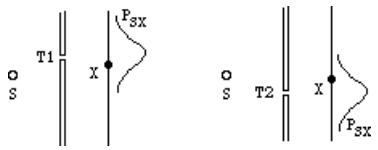
$$P_{SX/T_1} = P_{ST_1} \times P_{T_1X} \quad (resp \quad P_{SX/T_2} = P_{ST_2} \times P_{T_2X}).$$



La probabilité d'aller de S en X, quel que soit le trou emprunté est alors égale à la somme de ces deux probabilités: $P_{SX} = P_{ST_1} \times P_{T_1X} + P_{ST_2} \times P_{T_2X}$.



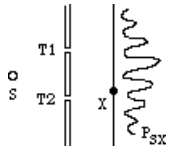
Quantiquement, le raisonnement est le même excepté que, tant qu'aucune mesure n'est effectuée, on additionne et on multiplie les amplitudes ondulatoires. Pour trouver la probabilité du résultat de la mesure finale de la position de la particule sur l'écran, on élève l'amplitude finale au carré. Si A_{XY} représente l'amplitude pour aller de X en Y, la probabilité d'aller de X en $T_1, P_{SX/T_1}$, est égale à $(A_{ST_1} \times A_{T_1X})^2$, et de même $P_{SX/T_2} = (A_{ST_2} \times A_{T_2X})^2$. Avec un seul trou ouvert on retrouve la situation classique,



mais avec les deux trous ouverts, l'amplitude finale est égale à la somme des amplitudes correspondant aux alternatives: $A_{SX} = (A_S T_1 \times A_{T_1 X}) + (A_{ST_2} \times A_{T_2 X})$, si bien que la probabilité pour aller de S en X vaut:

$$\begin{aligned} P_{SX} &= ((A_{ST1} \times A_{T1X}) + (A_{ST2} \times A_{T2X}))^2 \\ &= P_{ST1} \times P_{T1X} + P_{ST2} \times P_{T2X} + 2(A_{ST1} \times A_{T1X}) \times (A_{ST2} \times A_{T2X}) \end{aligned}$$

Dans le cas quantique, un terme supplémentaire doit donc être pris en compte. Comme l'amplitude est ondulatoire, c'est-à-dire que $AXY = Ae^{i \times quelque\ chose}$ (avec le "quelque chose" dépendant, d'une façon générale du temps et de l'espace X, Y), la valeur de la probabilité finale sera oscillante le long de l'écran:



Si on avait mesuré par quel trou passe la particule (par exemple en mesurant l'impulsion de l'écran), on aurait pu additionner directement les amplitudes au carré des chemins alternatifs, et le terme d'interférence aurait été annulé.

La suppression du terme d'interférence est-elle due à la perturbation physique de la mesure? Bohr admet déjà dans ses discussions avec Einstein que la perturbation qu'il invoque au sujet de l'observation n'est pas physique au sens usuel du terme. Einstein le précise avec son paradoxe d'Einstein (voir de Broglie 1957), et cela apparaît clairement avec le paradoxe d'Einstein Podolski Rosen (EPR) en admettant qu'une perturbation physique ordinaire (causale) doit être locale. Le fait est qu'avant la mesure l'amplitude est non nulle dans une vaste région de l'espace. Après la mesure elle est partout nulle sauf à l'endroit où la particule est détectée. Lorsque deux particules interagissent le système des deux particules est décrit en mécanique quantique par une seule fonction d'onde, si bien que malgré qu'elles puissent s'éloigner l'une de l'autre, la mesure sur une des deux particules supprime instantanément des possibles termes d'interférence concernant des mesures possibles sur l'autre particule. Cette non-localité, mise en évidence par Einstein et qui a pu être testée expérimentalement grâce au travaux de Bell (voir plus loin) ne permet pas d'envoyer de l'information à distance parce qu'elle exhibe seulement une corrélation statistique. Il n'est pas possible de transmettre de l'information avec deux dés qui seraient corrélés puisqu'il n'y a pas moyen de prédire ou de choisir le résultat du lancement du dé. Localement, le principe de localité est respecté bien qu'il faille ajouter localement un oracle (une donnée infinie) aléatoire pure. Les expériences de la pensée de l'automultiplication donne une phénoménologie mécaniste pour un tel oracle.

Axiomes de la mécanique quantique: la mécanique quantique décrit deux types d'évolution d'un système physique:

1. le développement continu et déterminé de l'onde (décrit par l'équation différentielle de Schrödinger dans le cas non relativiste par exemple). La grandeur oscillante est appelée amplitude.

2. la réduction abrupte de l'onde lors d'une observation. Le résultat de l'observation peut être prédit avec une probabilité calculable à partir de la forme de l'onde, (l'amplitude de l'onde au carré). C'est le principe de réduction.

Nous pouvons aborder l'argument suivant. Il est à la base d'une approche contra-mécaniste en philosophie de l'esprit:

c) L'argument classique : il constitue la théorie de la mesure de von Neumann (1932), reprise par (London and Bauer, 1939; Wigner, 1967a) etc. Puisque le comportement de tout ce qui est physique y compris mon corps et mon cerveau, en l'absence d'observation semble être décrit par le développement déterminé et continu de l'onde, c'est que la réduction de l'onde est opérée par ce qui est non physique en moi, c'est-à-dire, selon von Neumann, la conscience (parfois appelée ici spectateur ultime de la chaîne de von Neumann, la chaîne étant constituée de l'objet observé, l'oeil, la rétine, le nerf optique, les neurones du cortex visuel, etc.)

Cette théorie soulève des difficultés logiques et physiques considérables. Imaginez une mesure quantique faite par un aveugle qui se contente de photographier les résultats (sans les connaître, et donc sans prendre conscience de ces résultats, et donc sans réduire l'onde), et d'envoyer ces résultats par la poste à un ami physicien. Avec la théorie de von Neumann, le physicien aveugle est décrit par une superposition ondulatoire d'états incompatibles tant que son ami n'a pas pris conscience du résultat (voir Shimony 1963, ou encore 1989 pour des analyses plus détaillées de cet argument).

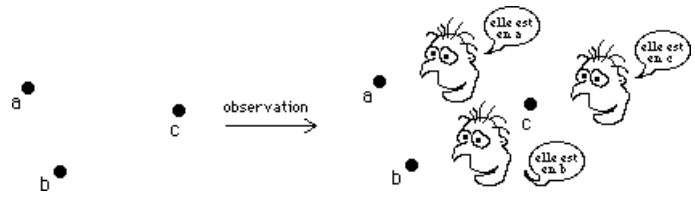
Avec l'hypothèse mécaniste, l'observateur-machine ne peut pas être privilégié par rapport à l'objet observé. L'observateur doit donc être décrit, comme l'objet, par l'équation continue et déterministe de l'onde. Il y a ici un plongement du sujet dans l'objet. Everett, en 1957, montre qu'il est encore possible d'assigner des états aux systèmes observés à condition de relativiser la notion d'état. L'observateur-machine se retrouvera lui-même dans une superposition d' états incompatibles. Mais comme le montreront indépendamment Graham et Hartle, chacun des observateurs-machines, multiplié par les interactions, placera dans sa mémoire un résultat bien précis, cohérent pour chacun des observateurs avec la statistique attachée au postulat de réduction de l'onde. Everett inaugure ainsi une des premières formulations de la mécanique quantique sans réduction de l'onde. Il construit à partir de l'équation d'onde appliquée au système complet [observateur + objet-observé] une phénoménologie de la réduction en utilisant l'hypothèse que l'observateur est une machine et qu'elle obéit dès lors aux lois de la physique.

Afin que cela soit plus claire voici une présentation plus schématique des positions respectives de von Neumann et d'Everett.

Une particule, tant qu'elle n'est pas observée, se manifeste comme si elle était en plusieurs endroits à la fois par l'intermédiaire de son amplitude ondulante. Des observations indirectes permettent de confirmer cette prédiction (ou plutôt rétrodiction) de la mécanique quantique, mais si on observe la position de la particule directement, l'observateur positionnera la particule à un endroit précis:

C'est la réduction de l'onde. Comme le cerveau de l'observateur est lui-même un système de particules, von Neumann a conclu que la réduction est opérée par la conscience

-explicitement non physique, et dissociée ainsi du cerveau- de l'observateur. Avec l'hypothèse mécaniste, on est obligé d'admettre que l'observateur a le même statut que la particule non observée. Dans ce cas il n'y a tout simplement pas de réduction:



L'observateur se multiplie comme la particule observée. Mais chacun des nouveaux observateurs pensent localiser la particule en un endroit précis : pour chaque observateur on obtient une phénoménologie de la réduction du paquet d'onde. Everett a montré que le formalisme de la mécanique quantique permet de justifier pourquoi l'observateur ne perçoit pas sa propre multiplication ni la présence de ses "sosies". De plus cette interprétation est nécessaire lorsqu'on désire appliquer la mécanique quantique en cosmologie où les systèmes sont gigantesques et incluent d'office les observateurs.

L'interprétation ou la formulation d'Everett de la mécanique quantique est pour le moins contre-intuitive puisque l'observateur est multiplié, sans qu'il ne puisse d'ailleurs le remarquer directement. Cette interprétation est controversée. Cependant, toutes les interprétations de la mécanique quantique sont contre-intuitives et controversées.

Tout argument en faveur de l'interprétation d'Everett, ou plus généralement en faveur des interprétations sans réduction du paquet d'onde, fait de la mécanique quantique une compagne pour la philosophie mécaniste de l'esprit plutôt qu'une plausible concurrente. L'argument le plus simple pour défendre Everett est peut-être donné par l'application d'une version conceptuelle du rasoir d'Occam : pour le philosophe mécaniste la théorie

équation de Schrödinger

est non seulement plus courte, mais est surtout conceptuellement plus simple que la théorie

équation de Schrödinger + principe de réduction

D'autant plus que le principe de réduction se réfère de façon arbitraire à l'observateur et à ses décisions.

Nous avons eu l'occasion de constater que la multiplication du sujet est un phénomène plus intrinsèquement mécaniste que quantique. Contentons-nous momentanément de retenir que les faits quantiques non seulement ne contredisent pas l'hypothèse mécaniste, mais semble plutôt la confirmer. L'interprétation d'Everett mue le doute qui vient de la chimie en motivation pour le mécanisme.

C.2 Quel effet cela fait-il d'être une machine dans un univers quantique?

C.2.1 L'indéterminisme quantique est un cas particulier de l'indéterminisme abrupte mécaniste

Avec l'interprétation d'Everett, l'indéterminisme quantique est un cas particulier de l'indéterminisme abrupte mécaniste. Les solutions de l'équation d'onde constituent un espace

vectoriel. La fonction d'onde d'une particule x peut avantageusement être représentée par un vecteur dans un espace vectoriel V dont la base est définie par les grandeurs que l'on peut mesurer sur la particule. Lorsque deux particules n'interagissent pas, elles sont descriptibles par la donnée indépendante de deux vecteurs x et y, et l'ensemble des deux particules est descriptible par un vecteur xy élément d'un produit dit tensoriel $V_1 \otimes V_2$ des espaces vectoriels associés à chacune des particules. Après une interaction, leurs ondes se superposent et évoluent dans l'espace produit. D'une façon générale la nouvelle fonction d'onde est une superposition linéaire des vecteurs de la base de $V_1 \otimes V_2$:

$$F = \sum_{i,j} a_{ij} u_i v_j$$

Dans le cas d'une interaction entre une particule P, décrite par (1)

$$(1)\ \ \Phi = \sum_n a_n \phi_n \ \ \text{avec}\ A(\phi_n) = A_n \phi_n$$

et un appareil de mesure, décrit par

$$\Psi = \sum_n b_m \varphi_m$$

où le carré du module de la "valeur propre" a_n donne la probabilité de trouver A_n si on mesure A, l'état général après l'interaction est donné par

$$\Gamma = \sum_{n,m} c_{nm} \phi_n \varphi_m$$

Pour que l'interaction entre l'appareil de mesure décrit par Ψ puisse servir à mesurer une grandeur (définissant la base $\{\varphi_n\}$ dans laquelle l'état de la particule est décrite), il doit exister une correspondance biunivoque entre les états de l'appareil de mesure et les états de la particule de telle façon que la connaissance de l'un entraîne la connaissance de l'autre. Von Neumann définit ainsi une mesure (idéale) par la condition

$$c_{nm} = c_n \delta_{nm}$$

L'état final est dès lors décrit par

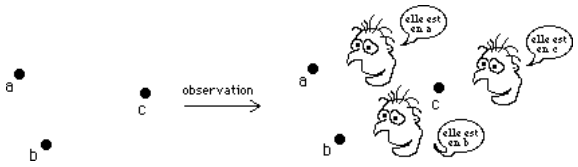
$$(2)\ \ \Gamma = \sum_n c_n \phi_n \varphi_n$$

L'observation de l'appareil de mesure donnera un état propre de l'appareil corrélé avec celui de la particule. Si la particule est dans un état propre ϕ_n de la grandeur observée, la mesure est décrite par l'évolution :

$$\phi_n \varphi_m \rightarrow \phi_n \varphi_n$$

Dans ce cas l'appareil ne perturbe pas la particule, et en reste séparé. La linéarité de l'évolution impose cependant d'admettre que si l'état initial de la particule se trouve être une superposition de vecteurs propres de A, décrit par (1), le système appareil de mesure + particule observée se trouve dans l'état de superposition général décrit par (2).

Dans l'interprétation d'Everett, l'observateur n'est pas privilégié. On le suppose à même d'autodistinguer ses propres états mentaux (du moins ceux relatifs à la mesure qu'il effectue dans un laboratoire). En ce sens il joue le rôle d'un appareil de mesure supplémentaire, et



l'interaction finale entre la particule, l'appareil de mesure et l'observateur sera décrit par une superposition (une somme) d'état du genre

$$\phi_n \varphi_n \xi_n$$

où x_n représente un état propre auto-distinctible de l'observateur. C'est, avec le collapse appareil de mesure et observateur, ce que représente justement le petit dessin :

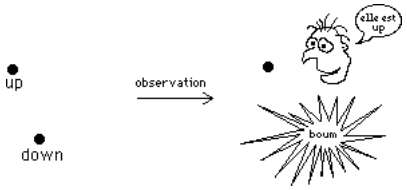
La réduction du paquet d'onde est remplacée ainsi par une suite de divisions de l'observateur. Notons avec Tipler qu'il n'est pas nécessaire de considérer que le cosmos entier se divise. Aucune partie du cosmos n'entrant pas en corrélation avec la mesure effectuée au laboratoire ne doit se diviser. Mathématiquement cela est dû au fait que le cosmos reste factorisable dans l'évolution du système particule + appareil + observateur + cosmos. (Tipler 1986, voir (Penrose and Isham, 1986)). Néanmoins l'observateur est multiplié, ou en tout cas ses états d'esprit, ou simplement ses mémoires sont multipliées. En acceptant cette division du sujet comme rendant compte de la phénoménologie de la réduction, l'interprétation d'Everett fait de l'indéterminisme quantique, mais aussi de la non-localité quantique, un cas particulier de l'indéterminisme abrupte et de la non-localité, qu'on a extrait de l'hypothèse mécaniste. En montrant comment retrouver les statistiques quantiques généralement associées aux réductions de l'onde, Everett utilise implicitement la formule "P = 1/2". Zeh compare explicitement la division quantique avec celle de l'amibe, voir (Marchal, 1988; Zeh, 1990). Ceci est encore plus clair dans les dérivations plus fines de la statistique quantique due indépendamment à Graham, Hartle, mais aussi, semble-t-il Finkelstein (selon Tipler). On peut faire les mêmes remarques concernant la non-observabilité de l'autodivision.

C.2.2 Le point de vue du chat de Schrödinger

Jusqu'à présent nous avons vu que le mécanisme et l'interprétation d'Everett se supporte l'un l'autre, mais a priori le mécanisme n'entraîne pas les faits quantiques. Everett use le caractère mécaniste de la mémoire et utilise explicitement MEC-IND sous la forme P=1/2 dans l'auto-division. Il se pourrait que MEC-IND soit vrai et que l'interprétation d'Everett, et même la mécanique quantique, soient fausses. Toutefois, dès à présent, l'hypothèse conjuguée d'Everett et du mécanisme permet de concevoir une forme de survie, et de distinguer la mort clinique de la mort absolue (Marchal, 1988; Moravec, 1988). Considérons un sujet humain réalisant une mesure sur un système quantique dont l'espace des états est de dimension 2, les vecteurs propres sont ϕ_{up} et ϕ_{down}, correspondant au valeur propre UP et DOWN (comme le spin d'un électron).

$$\Phi = \sum_n a_n \phi_n = \frac{1}{\sqrt{2}}(\phi_{up} + \phi_{down})$$

Et supposons qu'on ait branché une bombe au dispositif de mesure de telle façon que sa mise à feu soit activée par le résultat down de la mesure du spin effectuée par le sujet. Le sujet décide d'itérer cette mesure sur un lot indéfini de particules, toutes les particules étant individuellement (comme cela est possible dans l'interprétation d'Everett, voir aussi



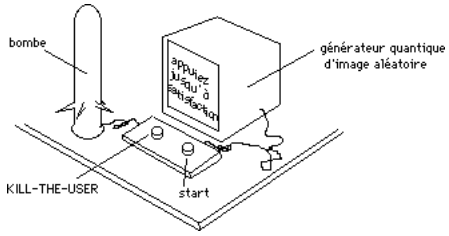
Hartle) décrites par Φ. L'expérience est strictement équivalente à un simple télétransport. L'observateur se divise et un des deux observateurs est annihilé par l'explosion de la bombe.

De l'intérieur, c'est-à-dire du point de vue de l'observateur itérant l'expérience, il sélectionne la continuation UP, UP, UP, etc. par suite de télétransports. Pour un observateur extérieur, non-annihilé par la bombe, la suite UP, UP, UP, est stochastiquement impossible et il sera témoin, conformément à la prédiction quantique, de l'explosion tuant l'observateur. Cette expérience illustre le caractère confirmable de l'interprétation d'Everett. A moins d'annihiler à chaque expérience la planète entière, ce type de confirmation ne peut être qu'une confirmation à la première personne. Cela saute aux yeux dans cette version du chat de Schrödinger, où l'on se met à la place du chat de Schrödinger, mais découle conceptuellement du mécanisme pour les expériences de survie à des duplications, ou téléportations (= duplication + annihilation).

C.2.3 L'ordinateur avec l'instruction "KILL-THE-USER"

En construisant un ordinateur muni d'un dispositif de choix quantique (un $\vee$ quantique (de type $\frac{1}{\sqrt{2}}(\phi_{up}+\phi_{down})$) qui permettra de diviser l'ordinateur en deux configurations parallèles (en une superposition de deux configurations incluant l'utilisateur), et en le munissant d'une instruction "KILL-THE-USER", c'est-à-dire en rajoutant une connexion entre le décodeur d'instruction et une bombe semblable à celle décrite plus haut, on obtient un ordinateur personnel (c'est le moins que l'on puisse dire) dont on peut prouver qu'il est capable de calculer plus rapidement la classe des problèmes NP (Non déterministe Polynomiaux). Avec cette machine on peut aborder des problèmes plus complexes débordant cette classe. Toutefois on prend le risque de sélectionner des univers (des extensions de soi) ou l'on survit à l'annihilation par un effet quantique comme celui de l'effet tunnel. L'interprétation d'Everett donne un modèle de relation sujet-objet qui rend impossible de garantir une annihilation absolue.

Une machine plus élémentaire illustre l'interprétation d'Everett.



En actionnant le bouton START, l'utilisateur déclenche le générateur d'images dont tous les choix d'allumage de pixel sont déterminés par des combinaisons de "$\vee$" quantiques. Ensuite, tant que l'utilisateur n'est pas satisfait de l'image (ou du texte !) proposé(e), il appuie sur le bouton KILL-THE-USER et se fait "annihiler" dans tous les mondes (états) où il n'est pas satisfait. Notons que du point de vue de l'utilisateur, il est le seul à pouvoir utiliser cette machine. Pour un observateur extérieur, utiliser cette machine est équivalent à

se tuer. Expérience par la pensée: que vont penser les observateurs extérieurs que l'utilisateur va rencontrer ?

C.2.4 Confirmation à la troisième personne de l'interprétation d'Everett

Deutsch montre comment construire un ordinateur quantique capable d'user d'interférences entre les calculs effectués parallèlement par superposition quantique, et capable ainsi de calculer plus rapidement une classe précise de problèmes (strictement inclue dans la classe des fonctions accélérées par l'ordinateur avec touche KILL-THE-USER). Selon Deutsch, une telle machine permettrait de confirmer, à la troisième personne, l'interprétation d'Everett (Deutsch, 1986b). Un partisan de l'interprétation de Bohm peut cependant invoquer la notion (non-mécaniste) de zombie pour invalider (logiquement) cette argumentation (Bohm and Hiley, 1993).

C.2.5 Le rôle pédagogique de l'interprétation d'Everett

L'interprétation d' Everett est une opportunité concernant la question du choix de niveau de fonctionnalisme. Elle permet de concevoir un mécanisme très faible de bas niveau. Le candidat à la duplication qui estime qu'il faut dupliquer tout l'univers, ou une grande portion d'univers pour qu'il puisse survivre peut exhiber un exemple concret de telle duplication en invoquant l'interprétation d'Everett. Le calcul de Graham et Hartle revient à justifier l'existence et l'unicité d'une mesure extraite du computationnalisme, sous la forme implicite de P=1/2. La question se repose toutefois pour l'hypothèse mécaniste générale. C'est ce que j'illustre dans le texte.

C.2.6 Peut-on croire à l'interprétation d'Everett ?

Martin Gardner écrit :

> The many-worlds interpretation has been called a beautiful theory nobody can believe. (Gardner 1988)

Il cite à ce sujet DeWitt, un enthousiaste (comme de nombreux cosmologistes tâchant de marier la gravitation universelle avec le phénomène quantique) s'exclamant :

> I still recall vividly the shock I experienced on first encountering this multiworld concept. The idea of 10^{100+} slightly imperfect copies of oneself all constantly splitting into further copies, which ultimately become unrecognizable, is not easy to reconcile with commonsense (Dewitt, 1970).

Roger Penrose témoigne d'un profond scepticisme

> I cannot believe the many-worlds interpretation make physical sens.

J'ai déjà illustré la difficulté qu'il y a pour croire que P=1/2 en MEC-IND, et le caractère ironique des propositions à ce sujet. Le théorie de la conscience proposée ici tente de capturer la difficulté de croire, d'une façon générale, à la possibilité de la vérité concernant ses propres extensions. Les expériences par la pensée proposées illustrent cette difficulté.

C.3 Inégalités de Bell et Logique Quantique

C.3.1 Violation des Inégalités de Bell

On sait depuis, Einstein & Al. 1935, que certains principes de base sont incompatibles avec les prédictions de la mécanique quantique. Depuis Bell 1964, on a pu mettre en évidence et réaliser des expériences précises dont les résultats sont incompatibles avec ces mêmes principes.

Principes de Base

1. Principe de réalité d'Einstein : si, sans perturber un système, on peut prédire avec certitude la valeur d'une quantité que l'on va mesurer, alors il y a un élément de réalité (indépendant de "moi" et de "mes" décisions) associé à cette quantité.
2. Principe de séparabilité : une influence ne peut pas se propager avec une vitesse supérieure à la vitesse de la lumière.

   Remarque : la violation du principe de séparabilité ne contredit la relativité restreinte qu'à condition d'adopter sans restriction le principe de réalité d'Einstein.
3. Principe d'induction : si on effectue un très grand nombre de fois la même expérience dans les mêmes conditions (définies macroscopiquement), et qu'on obtient toujours le même résultat, alors on obtiendra encore le même résultat si on réitère l'expérience. (De nettes présomptions sont au moins justifiées).

   Ex : celui qui pense que le soleil se lèvera demain matin utilise le principe d'induction.
4. Principe de contrafactualité : une qualité peut être attribuée contrafactuellement, c'est-à-dire, en se référant au résultat certain d'une expérience NON réalisée. Cela permet par exemple, d'affirmer que le morceau de fer que je tiens en main est magnétique simplement parce que je suis sûr qu'il *attirerait* de la limaille de fer si je lui en présentais.

   Remarque : le principe de contrafactualité permet de remplacer dans le principe de réalité, l'expression "que l'on va mesurer" par "que l'on mesurerait".

Démonstration d'une inégalité de Bell

Si A, B et C sont trois ensembles finis, alors on a :

$$A \cap B \subseteq (A \cap C) \cup (B \cap \overline{C})$$

avec $\overline{C}$ représentant le complémentaire de C.

En effet, si $x \in A \cap B$ alors :

- si $x \in C$, on a $x \in A \cap C$ et donc x appartient au deuxième membre;
- si $x \notin C$ on a $x \in \overline{C}$ et $x \in B \cap \overline{C}$ et donc x appartient au deuxième membre également.

Désignons le nombre d'éléments de $A \cap B$ par $N_{AB}(++)$. En réalisant que $A \cap C$ et $B \cap \overline{C}$ sont disjoints, on a, avec des notations évidentes:

$$N_{AB}(++) \leq N_{AC}(++) + N_{BC}(+-)$$

Imaginons à présent un ensemble dont les éléments sont susceptibles de réagir positivement, ou négativement, à trois types de tests dichotomiques donnés A, B et C. Je dirai, avec abus de langage, que l'élément x a la propriété A (resp. $\overline{A}$) si je peux prédire avec certitude que x réagirait positivement (resp. négativement) si je réalisais le test A. Prenons dans notre ensemble trois échantillons possédant le même nombre d'éléments. Alors, avec une probabilité qui se rapproche de 1 lorsque la taille des échantillons augmente, le nombre d'éléments $n_{AB}(++)$ de l'échantillon 1 est inférieur ou égal au nombre d'éléments $n_{AC}(++)$ de l'échantillon 2 ajouté au nombre d'éléments $n_{BC}(+-)$ de l'échantillon 3 :

$$n_{AB}(++) \leq n_{AC}(++) + n_{BC}(+-)$$

Cette relation est connue sous le nom d'inégalité de Bell. Cette présentation est inspirée de d'Espagnat.

Remarque : le principe de séparabilité n'a pas été utilisé jusqu'ici.

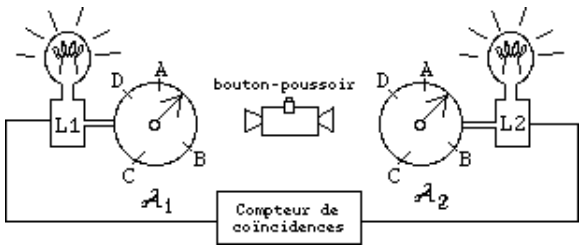
Violations expérimentales des inégalités de Bell

Il est facile de montrer qu'il existe des prédictions de la mécanique quantique violant l'inégalité de Bell. On peut même avancer que la contradiction réside dans les principes de réalité et de contrafactualité. C'est la raison pour laquelle certains physiciens se sont proposé de modifier la mécanique quantique. Les nouvelles théories ne peuvent évidemment admettre les mêmes prédictions que la mécanique quantique (puisque ce sont les prédictions qui violent l'inégalité). Des tests expérimentaux ont donc été réalisés. Deux faits sont remarquables:

- Ces expériences, indépendamment de toutes interprétations, violent l'inégalité de Bell. Un de nos principes au moins devra être abandonné.
- L'inégalité est violée dans les rapports prédits par le formalisme quantique. Ceci suggère à nouveau une faiblesse du principe de réalité. Compte tenu du succès, sur le plan prédictif, de la relativité restreinte, il est déjà difficile de nier le principe de séparabilité. Quant au principe d'induction, le plus difficile à justifier philosophiquement (Hume), il est à la base même de toute démarche scientifique sur base empirique.

Je vais décrire une expérience type du genre de celles qui violent l'inégalité en prenant soin de n'utiliser aucun terme de la physique (contemporaine). On ne sait ni ce qu'est un photon, ni ce qu'est un polariseur, encore moins ce qu'est une fonction d'onde !

Je considère le dispositif suivant:



Il est constitué d'un dispositif, pourvu d'un bouton-poussoir, situé symétriquement entre deux appareils $\mathcal{A}_1$ et $\mathcal{A}_2$ reliés chacun à une lampe L_1 et L_2. $\mathcal{A}_1$ et $\mathcal{A}_2$ sont

identiques et possèdent respectivement au moins quatre degrés de liberté correspondant à quatre états possibles : A, B, C, D. Les deux lampes sont à leur tour reliées à un compteur de coïncidences qui compte le nombre de fois où les deux lampes se sont allumées simultanément.

Voici les faits de base :

- Lorsque les appareils $\mathcal{A}_1$ et $\mathcal{A}_2$ sont chacun dans l'état D, chaque fois que j'appuie sur le bouton, les deux lampes s'allument toujours.
- Chaque fois que je presse le bouton, et que $\mathcal{A}_1$ et $\mathcal{A}_2$ sont dans le même état, ou bien $L1$ et $L2$ s'allument ensemble, ou bien ni $L1$ ni $L2$ ne s'allument. Il y a *toujours* coïncidence.

Avec les principe de séparabilité et de réalité, je dois conclure qu'à chaque pression "quelque chose" quitte la région du bouton-poussoir pour l'appareil $\mathcal{A}_1$ (de même pour l'appareil $\mathcal{A}_2$). Ce sont ces quelque choses qui jouent le rôle d'éléments de réalité de la démonstration de l'inégalité. Ce sont ces "éléments de réalité" qui sont comptés.

Si je sais avec certitude que la lampe $L1$ (resp $L2$) s'allumerait si l'appareil $\mathcal{A}_1$ (resp $\mathcal{A}_2$) était dans un état A_i (où A_i est un des quatre degrés de liberté), je dirai, avec un abus de langage évident, que les éléments correspondants ont la propriété A_i (j'utilise les principes de réalité et de contrafactualité).

En particulier, puisque je peux prédire avec certitude que si $\mathcal{A}_1$ et $\mathcal{A}_2$ sont dans l'état D, à chaque pression les deux lampes s'allument, on aura deviné que D est la simple propriété "d'existence indépendante" de l'élément de réalité. Avec le principe d'induction, sur base du premier fait, tous les éléments créés, après qu'on ait poussé sur le bouton, ont cette propriété D.

De même, en utilisant encore le principe d'induction sur base du second fait, je peux affirmer qu'après avoir poussé sur le bouton-poussoir, l'élément de gauche possède les mêmes propriétés (parmi A, B, C, D) que l'élément de droite.

Ainsi pour chaque élément, je peux mesurer deux propriétés intrinsèques: il suffit de placer $\mathcal{A}_1$ dans l'état A et $\mathcal{A}_2$ dans l'état B (par exemple), si $L1$ s'allume, $L2$ s'allumerait avec certitude si $\mathcal{A}_2$ était dans l'état A, donc l'élément qui va vers $\mathcal{A}_2$ possède la propriété A, et donc l'autre élément aussi. De même, si $L2$ s'allume, les deux éléments ont la propriété B.

Comme après chaque pression les éléments possèdent toujours la propriété D, si une lampe ne s'allume pas lorsque l'appareil est dans l'état A (resp. B ou C) l'élément correspondant possède nécessairement la propriété A (resp. B ou C).

L'expérience, où les inégalités de Bell sont violées, consiste à pousser N fois (N grand) sur le bouton, avec

- $\mathcal{A}_1$ et $\mathcal{A}_2$ dans l'état A et B respectivement, on compte le nombre de fois où les deux lampes s'allument : $n_{AB}(++)$;

- $\mathcal{A}_1$ et $\mathcal{A}_2$ dans l'état A et C respectivement, on compte le nombre de fois où les deux lampes s'allument : $n_{AC}(++)$;

- $\mathcal{A}_1$ et $\mathcal{A}_2$ dans l'état B et C respectivement, on compte le nombre de fois où $L1$ s'allume avec $L2$ qui ne s'allume pas : $n_{BC}(+-)$.

Paradoxalement on trouve : $n_{AB}(++) > n_{AC}(++) + n_{BC}(+-)$

REMARQUE Il se pourrait que les éléments de gauche et de droite n'aient les mêmes propriétés que dans le cas restreint où $\mathcal{A}_1$ et $\mathcal{A}_2$ sont dans le même état. On devrait

alors admettre que ces deux appareils influent sur les conditions initiales en informant les éléments, non encore envoyés, de leurs positions. Un principe supplémentaire peut être ajouté pour interdire ce genre d'influence. Pour éviter l'usage d'un tel principe on peut faire en sorte que les états de $\mathcal{A}_1$ et de $\mathcal{A}_2$ soient choisis un temps $\frac{L}{c-\varepsilon}$ après avoir appuyé sur le bouton-poussoir, (L représentant la distance entre le bouton et un appareil, c représentant la vitesse de la lumière). Une expérience de ce type a été proposée par Aspect en 1976 et réalisée depuis avec succès (Aspect, 1976; Aspect et al., 1982).

Comme les cônes de lumière de chaque particule intervenant dans cette expérience, y compris les atomes de mon propre cerveau, lequel a choisi les paramètres A, B, C....., ont probablement une intersection non vide, on peut donc encore imaginer une interprétation causale et réaliste. Mais une telle conspiration cosmique des éléments de la nature interdirait d'inférer quoi que ce soit d'universel à partir d'observations de la nature.

C.3.2 Logique quantique

La relation booléenne

$$A \cap B \subseteq (A \cap C) \cup (B \cap \overline{C})$$

peut être considérée comme une interprétation algébrique de la tautologie classique $A \wedge B \rightarrow (A \wedge C) \vee (B \wedge \overline{C})$, ou plus exactement de la règle classiquement correcte $A \wedge B \Rightarrow (A \wedge C) \vee (B \wedge \overline{C})$.

Existe-t-il une axiomatisation de la logique quantique QL, c'est-à-dire une théorie axiomatisant les propositions quantiques ?

Une telle logique a été proposée par Birkhoff et Von Neumann en 1936.

Comme pour la logique intuitioniste, on peut la voir comme un affaiblissement de la logique classique. Selon Dalla Chiara 1976, on obtient une axiomatisation Hilbertienne de la logique quantique à partir d'une présentation à-*la-Kleene* du calcul propositionnel classique (voir 1.2), en affaiblissant l'axiome classique (le principe de l'a fortiori) :

$$p \rightarrow (q \rightarrow p)$$

Ce schéma d'axiome n'est accepté en logique quantique que pour des propositions implicationnelles à la place de p. Les propositions implicationnelles ont la forme $A \rightarrow B$, ou une combinaison booléenne de telles formes, avec A et B formules quelconques.

Cette présentation Hilbertienne donne l'impression qu'il existe une implication matérielle en QL, mais le théorème de déduction de Herbrand n'est pas valable, et l'usage de l'implication de Dalla Chiara est assez "pathologique". Il s'agit plutôt d'une déduction. Il est nécessaire de rajouter comme règle d'inférence, en plus du modus ponens (et des éventuelles règles habituelles pour le traitement des quantificateurs), une règle de l'a posteriori:

$$p \Rightarrow (q \rightarrow p)$$

La coutume est plutôt de dire que la logique quantique n'a pas d'implication et de présenter alors la logique dans une présentation à la Gentzen (rien qu'avec des règles d'inférence) :

$A \Rightarrow A$
$A \wedge B \Rightarrow A$
$A \wedge B \Rightarrow B$
$A \Rightarrow \neg\neg A$
$\neg\neg A \Rightarrow A$
$A \wedge \neg A \Rightarrow B$
$Si\, A \Rightarrow B\, et\, B \Rightarrow C\, alors\, A \Rightarrow C$
$Si\, A \Rightarrow B\, et\, A \Rightarrow C\, alors\, A \Rightarrow B \wedge C$
$Si\, A \Rightarrow B\, alors\, \neg B \Rightarrow \neg A$

Il s'agit ici de version faible de la logique quantique. En général, on rajoute, pour avoir la logique quantique proprement dite, un axiome d'orthomodularité, par exemple le schéma d'axiomes

$$(p \rightarrow q) \rightarrow (q \rightarrow (p \vee (\neg p \wedge q))$$

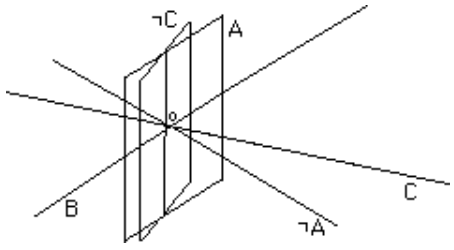
pour le système Hilbertien de Dalla Chiara, ou la règle

$$A \wedge (A \vee (\neg A \wedge B)) \Rightarrow B$$

dans la présentation à la Gentzen (voir Goldblatt 1974). $A \vee B$ est, par définition, une abréviation de $\neg(\neg A \wedge B)$.

La logique quantique est née des considérations algébriques et est présentée souvent exclusivement par sa sémantique algébrique. Je renvoie à la littérature pour les théorèmes de complétude correspondant.

En gros, la valuation d'une proposition quantique est donnée par un sous-espace d'un espace vectoriel muni d'une notion d'orthogonalité. Le & est capturé par l'intersection (l'intersection de sous-espaces est un sous-espace); le $\vee$ est capturé par la somme des sous-espaces (= le plus petit sous-espace engendré); la négation $\neg$ est capturée par la relation d'orthogonalité vectorielle (étendue aux sous-espaces : un espace est orthogonal à un autre si tous les vecteurs de cet espace sont orthogonaux à tous les vecteurs de l'autre espace). De plus, comme il est usuel dans les sémantiques algébriques, la déduction $\Rightarrow$ est capturée par la relation "être sous-espace". L'espace euclidien R^3 suffit ici pour réfuter la relation "booléenne" correspondant à l'inégalité de Bell. B est un sous-espace de A, et $A \cap B = B$. Il suffit alors de prendre C de telle façon qu'il soit unidimensionnel et non orthogonal à A. Dans ce cas $A \cap C$ et $B \cap C$ se réduisent au sous-espace vectoriel nul comme on le voit sur la figure suivante :



La somme des deux vectoriels est le vectoriel nul, et $A \cap B = B$ n'est pas un sous-espace du vectoriel nul.

De la même façon on peut vérifier que : $p \wedge (q \vee r) \not\Leftrightarrow (p \wedge q) \vee (p \wedge r)$, c'est-à-dire la logique quantique est non distributive.

Annexe D

Zombies et compagnie

Dans cette annexe sont regroupés

- Une introduction brève au problème du corps et de l'esprit.
- Une explication comme quoi le computationnalisme est aussi, selon d'autre critère, une hypothèse forte.
- Un approfondissement qu'il s'agit d'une hypothèse faible ainsi que des précisions sur les notions de niveaux et de cerveau généralisé.

A nouveau cette annexe ne remplace pas la lecture de bons livres, par exemple le récent (Tye, 1995) ou, pour un recueil d'articles : (Warner and Szubka, 1994).

D.1 Le problème du corps et de l'esprit

Il n'est pas facile d'énoncer de but en blanc le problème du corps et de l'esprit[1]. Il y a autant d'énonciations possibles qu'il y a de systèmes philosophiques. Il n'est pas non plus nécessaire d'étudier en détail ces énonciations puisque le présent travail consiste en grande partie en la dérivation d'une nouvelle formulation de ce problème dans le cadre computationnaliste.

Essentiellement, résoudre le problème du corps et de l'esprit (PCE) consiste à rendre compte de l'écart entre ce qui est désigné par le corps, ou la matière, ou l'objet ou d'une façon générale tout ce qui est communicable à la troisième personne, d'une part et, d'autre part l'esprit, l'âme, le soi, le vécu, la subjectivité ou d'une façon générale ce que la première personne, appelée aussi le sujet, est capable de ressentir de façon privée.

Est-il seulement possible d'aborder objectivement la nature de la subjectivité ? Cet espoir n'est-il pas à la base contradictoire ?

On doit reconnaître, avec (Nagel, 1994), que la science a d'autant plus progressé qu'elle a mis le sujet à l'écart. On peut dire sans crainte d'exagérer, que la science, ou les scientifiques, ne s'intéressent qu'aux propositions communicables à la troisième personne, et donc vérifiables par la troisième personne.

[1] souvent appelé le problème des rapports de l'âme et du corps. Pour une introduction historique à la problématique, voir (Baertschi, 1992)

On ne s'étonne pas alors que certains "scientifiques" sont tentés d'écarter le problème du corps et de l'esprit en le jugeant comme étant a priori non scientifique.

Cette attitude repose cependant sur une confusion de niveaux ou de catégories. En effet, rien *a priori* n'interdit, moyennant définitions et hypothèses *comme toujours en sciences*, d'isoler un discours communicable à (et donc vérifiable par) la troisième personne, discours portant *précisément* sur les discours de la première personne.

Et c'est bien ce que tente de faire la psychologie cognitive, notamment les psychologues fonctionnalistes, mais aussi à leur façon les neurophysiologues. Le succès de cette entreprise reviendrait à "naturaliser l'intention" (Pacherie, 1993) ou encore, à résoudre ce qu'il est convenu d'appeler, en philosophie de l'esprit, *le problème restreint du corps et de l'esprit* (the weak mind-body problem). On disposerait d'une théorie, c'est-à-dire un discours communicable à la troisième personne (je dirai simplement: un discours 3-communicable), discours permettant d'expliquer à partir de l'état d'un cerveau (généralisé) les discours et les comportements de la première personne.

Malheureusement un tel succès évacuerait *apparemment* ce qu'on appelle "the strong mind-body problem". Je dirai : le problème *profond* du corps et de l'esprit. Le problème profond concerne la relation entre le cerveau (généralisé) et l'expérience phénoménale de la conscience elle-même, et non pas la relation entre le cerveau et les *discours*, portant éventuellement en apparence sur cette expérience phénoménale). Regardons cette évacuation de plus près :

D.1.1 Le paradoxe du fonctionnalisme

Proposition 24 *Une 3-justification complète évacue la première personne, et donc le sens, proprement dit, des 1-justifications.*

Preuve. Considérons la petite histoire suivante

> Jeanne retira rapidement sa main du réchaud. Il était froid pourtant.

L'explication typique, de la psychologie populaire ou "de grand-mère" comme on l'appelle parfois, consiste à attribuer des croyances à Jeanne, croyances en général liées à des souvenirs, comme celui de s'être brûlée à un réchaud, ou avoir vu quelqu'un se brûler à un réchaud, etc.

S'il existait une 3-justification mécaniste, le fonctionnalisme devrait interpréter une "croyance de Jeanne" par un état "computationnel du corps (cerveau) de Jeanne". "A croit quelque chose" signifie que A est dans un état particulier que l'on pourrait caractériser par une collection de comportements qui définiraient cette croyance, un peu comme on peut définir une fonction par ses couples entrées-sorties, ou un état fonctionnel d'une machine par les possibles entrée-sortie. Dans ce cas on peut attribuer un rôle causal à la croyance et justifier des propositions psychologiques. Mais si cette 3-justification est complète, l'expérience de la personne, en l'occurrence Jeanne, n'a aucun rôle, ni dans son comportement ni pour ces 1-discours, lesquels deviennent alors vide de sens, puisque ces 1-discours évoquent justement l'expérience phénoménale ainsi que le rôle de cette expérience. Par exemple Jeanne nous dit : "C'est parce que je me souviens de la douleur ressentie en touchant un réchaud que j'ai crains un moment de risquer de me brûler à nouveau".

Cela semble presque aller dans le sens des scientifiques qui pensent que le problème du corps et de l'esprit n'est pas scientifique. En particulier cette évacuation serait sans doute jugée bienvenue par les épistémologues strictement positivistes, ainsi que

par ceux qui ont une conception purement instrumentaliste de la science. Cette évacuation pourrait peut-être satisfaire également les philosophes analytiques parmi ceux qui espèrent réduire complètement les questions métaphysiques à des questions de linguistique. Mais cette évacuation serait jugée prématurée par tout ceux qui reconnaissent en eux l'expérience, ou une expérience, phénoménale (consciente donc!). Bref, elle choquerait tout ceux qui pensent qu'ils ne sont pas des zombies. C'est le paradoxe du fonctionnalisme matérialiste : son succès (apparemment complet) enlèverait tout rôle à la conscience. Celle-ci deviendrait, au mieux un épiphénomène, au pire un mot sans référent : nous serions des zombies et le prix à payer pour une 3-justification des discours de la première personne serait l'absence de la première personne. Elle évacuerait la psychologie de grand-mère qui non seulement attribue la conscience à la première personne, mais attribue un rôle à cette conscience. Mais le plus grave, concernant le présent travail, est le fait que cette évacuation rendrait inintelligible l'hypothèse du mécanisme *indexical*. En effet, l'hypothèse du mécanisme indexical a été présenté comme une réponse (en l'occurrence la réponse hypothétique "oui") à une question posée *à une première personne*. Le mécanisme a été illustré par une procédure de survie personnelle. L'aspect indexical du mécanisme présuppose un minimum de psychologie populaire.

On peut cependant retenir la leçon :

Corollaire 25 *COMP* $\Rightarrow$ *il n'y a pas de 3-justifications* complètes *des 1-discours*

Avec la thèse de Church nous avons été prévénu: certaines machines seront plus éloquentes par leur silence que par leur bavardage. Remarquons aussi que s'il existait des 3-justifications complètes il deviendrait possible de déterminer explicitement le niveau de substitution auquel on survit.

Il n'est donc pas interdit de penser qu'une 3-justification (et donc une justification de nature scientifique) puisse néanmoins expliquer complètement, quoiqu'indirectement, le caractère vrai mais incommunicable d'une partie des 1-discours, et par là rendrait compte de la première personne. Ce que nous savons à présent c'est qu'elle serait nécessairement incomplète. Mais ceci est justement rendu possible (et nécessaire!) par la thèse de Church. Une telle 3-justification pourra donc servir une phénoménologie de l'esprit, et le computationnalisme conduit naturellement à une telle phénoménologie de l'esprit.

Beaucoup de philosophes ont l'air de penser qu'une phénoménologie de l'esprit constitue par elle-même une explication (mécaniste) de la relation de l'esprit au corps. Une telle phénoménologie donnerait la solution au problème du corps et de l'esprit. L'essentiel du présent travail a consisté à montrer que tel n'est pas le cas : les arguments du DU et du GF ont en effet éliminé, avec le computationnalisme, la matière ou tout univers *substantiel*. Cette élimination a transformé le problème du corps et de l'esprit en la dérivation d'une phénoménologie de la matière. On a donc dû non seulement justifier les 1-discours et le rôle de la conscience (comme vérité accessible mais non communicable), mais on a dû aussi justifier nos croyances en une apparence mesurable (physique) et notre connaissance (vraie) des sensations physiques.

D.1.2 Une formulation générale du problème du corps et de l'esprit

Ontologie, rasoirs d' Occam et phénoménologie

L'ontologie, par définition, c'est ce qui est sensé décrire l'être, des êtres, ce qui existe, la collection de ce qui existe. Elle recouvre l'objet de l'éventuel engagement ontologique du philosophe. Elle constitue la réalité fondamentale que sa doctrine admet comme primitive. A cette occasion, le rasoir d'Occam (ou Ockham) est fréquemment invoqué pour justifier un engagement ontologique minimal. On utilise en effet le rasoir d'Occam pour satisfaire un principe de parcimonie : les entités existantes, selon ce principe, ne doivent pas être multipliées plus qu'il n'est nécessaire. Dans la littérature, on rencontre différentes sortes d'interprétations du rasoir d'Occam, pris dans un sens élargi:

Les interprétations syntactiques On veut la plus petite (dans différents sens) théorie possible. C'est, en général, l'interprétation la moins intéressante. Une théorie minimale (incompressible) est toujours difficile à comprendre. Dans la pratique on utilise toujours des présentations syntaxiquement redondantes des théories.

L'interprétation sémantique Une première idée consisterait à avoir une théorie qui admet le moins de modèles possibles. Cette idée est directement contrecarrée par les résultats de limitation des formalismes dès qu'on a des théories assez riches. Aujourd'hui la mode est plutôt aux théories ubiquistes, qui admettent de nombreux modèles permettant l'interprétation des théories dans des contextes très variés (groupes, catégories, foncteurs, etc.).

Une meilleure idée pour l'interprétation sémantique du rasoir d'Occam consiste plutôt à avoir le moins d'objets possibles dans le modèle attendu d'une théorie, du moins lorsque cette théorie est destinée à décrire une ontologie.

L'interprétation conceptuelle On veut le moins d'hypothèses *ad hoc* possibles, ou de principes philosophiques, dans la théorie ou dans la philosophie. L'interprétation sémantique peut être considérée comme une interprétation conceptuelle particulière.

définition Lorsque l'on rend compte, dans un cadre philosophique donné à partir d'une certaine ontologie O, d'un phénomène P, *sans engagement ontologique supplémentaire*, je dirai qu'on a construit une *phénoménologie* de P à partir de O. Cette construction utilise implicitement l'interprétation conceptuelle ou sémantique du rasoir d'Occam.

En philosophie de l'esprit, on distingue les monistes qui commettent un engagement ontologique unique (portant sur l'esprit ou sur la matière) des dualistes qui commettent un double engagement ontologique (l'esprit et la matière).

L'expression "problème du corps et de l'esprit" (The mind-body problem) est elle-même une formulation dualiste du problème du corps et de l'esprit. En fait, il y a autant de formulations du problème du corps et de l'esprit qu'il y a de types de systèmes philosophiques.

Exemples:

- Pour le monisme matérialiste, le problème consiste à construire une phénoménologie de l'esprit, ou d'une phénoménologie de la relation esprit/matière, à partir d'une ontologie matérielle ;

- Pour le monisme idéaliste, le problème consiste à construire une phénoménologie de la matière, ou d'une phénoménologie de la relation esprit/matière, à partir d'une ontologie ou d'une "vérité" spirituelle ;
- Pour le dualiste, le problème consiste à construire une "phénoménologie" de l'interaction entre l'esprit et la matière. Le mot phénoménologie doit être pris ici dans un sens nécessairement différent et a priori difficile à préciser : c'est la faiblesse de base du dualisme. En effet, si l'on pense qu'il existe une interaction entre l'esprit et la matière, on voit mal comment échapper à la disjonction suivante : ou bien l'esprit est une forme ou une modalité de matière, ou bien la matière est une forme ou une modalité de l'esprit. En essayant de rendre le dualisme intelligible on tombe à coup sûr dans l'un ou l'autre des monismes décrits plus haut, ou l'on tombe dans l'épiphénoménalisme. Celui-ci est juste un nom pour cesser de chercher à comprendre aussi bien l'esprit que la matière.

Au lieu de construire une phénoménologie, certains philosophes préfèrent argumenter contre l'idée que cette phénoménologie est nécessaire. Le matérialisme éliminativiste, par exemple (Churchland 1986) considère le discours du psychologue traditionnel, appelé aussi *la psychologie de grand-mère*, comme vide de sens. Je tiendrai les tentatives explicites d'élimination (phénoménologique) de ce genre comme des élaborations phénoménologiques particulières. Si l'élimination est implicite et non accompagnée de justifications, alors on est, à nouveau, en présence des positions béhavioristes ou positivistes, positions qui sont stériles (par construction) en ce qui concernent les problèmes durs de la philosophie de l'esprit.

De nombreuses confusions en philosophie des sciences proviennent du fait qu'on s'attache à préserver simultanément le mécanisme et le matérialisme, par crainte de l'idéalisme, lui-même confondu avec le solipsisme (l'idéalisme subjectif).

(Dans le rapport technique (Marchal, 1995) je traite deux exemples de confusions de ce genre, chez Bell 1987, et dans le dialogue de Changeux et Connes 1989).

Avec ces termes on peut exprimer le résultat principal de notre travail par

$$\text{COMP} \Rightarrow \text{monisme idéaliste}$$

où l'ontologie idéaliste est la vérité arithmétique, et où le problème du corps et de l'esprit est transformé en la recherche d'une phénoménologie de l'esprit *et* de la matière.

D.2 Le computationnalisme est une hypothèse forte

Pour distinguer les mécanismes possibles, une façon de procéder est de regarder des situations où ces différentes versions du mécanisme est faux.

Considérons en effet l'expérience de la duplication de soi, et analysons les résultats possibles suivants :

Absence de double Le cas le plus simple où le mécanisme, défini plus haut, est faux est l'échec complet des duplications possibles. Par exemple, le double obtenu est inanimé. Il n'a pas d'activité cérébrale, ou il n'a rien de commun avec moi. Et ça quel que soit le niveau de duplication. Dans ce cas le mécanisme est faux.

Le double est un zombie Ensuite vient le cas où le double est un zombie. Il me ressemble, il a les mêmes attitudes que moi, et aucun observateur extérieur n'est à même de le distinguer de moi. Mais par définition, il n'a pas de conscience, pas d'expérience privée, etc. Le "zombie" désigne en philosophie de l'esprit contemporaine un être qui a toute l'apparence d'un être humain, mais qui n'a pas de conscience au sens phénoménal du terme. On peut sérieusement douter de l'existence du zombie, mais le zombie est, d'un point de vue strictement logique, concevable[2]. Et si le double est toujours un zombie, le mécanisme, tel que je le considère dans ce travail, est certainement faux, indépendamment du fait que cette fausseté n'est pas vérifiable. Par contre une forme affaiblie du mécanisme demeure correcte. Il s'agit du mécanisme béhavioriste (Mec-Beh).

Définition Mec-Beh : Une machine peut posséder toutes les apparences extérieures (descriptible à la troisième personne) d'un être humain (en l'occurrence de moi).

Sous l'influence du mouvement positiviste, certains auteurs (dont Turing avec son célèbre test de l'imitation) ont défini le mécanisme de façon exclusivement béhavioriste. Du point de vue positiviste, cette définition est très ouverte : le terme "zombie" n'a pas de sens, et Turing attribue par défaut une expérience privée à toute entité capable de passer un test d'imitation *suffisamment sophistiqué*.

Cependant, cette approche du mécanisme est un peu stérile en philosophie de l'esprit où l'on tente d'attribuer des références explicites à des termes comme "privé", "conscience", "opinion personnelle" etc.

Le double est un autre Ensuite vient naturellement le cas où le double n'est pas un zombie, on lui attribue explicitement une expérience intérieure. Il est humain à part entière, mais *il n'est pas moi*. C'est ce que pense par exemple Monsieur D2 dans l'histoire de Monsieur D. Le double apparaît comme un "frère jumeau" tombant du ciel, une sorte d'imposteur. Dans ce cas, notre hypothèse computationnaliste est fausse, mais une part plus restreinte reste correcte : il s'agit du mécanisme fort (Mec-Fort), ou, si le double est digital, il s'agit de la thèse forte de l'intelligence artificielle (strong AI thesis), selon laquelle une machine (digitale) peut avoir une expérience privée. Cela n'implique pas logiquement que je (ou l'homme) soit une machine.

Définition Méc-Fort : Une machine peut être le véhicule d'une expérience privée. Une machine peut permettre la manifestation d'une expérience authentiquement subjective, intime, phénoménale, privée.

Le double est moi Ou plus exactement, le double peut être moi, même si je ne peux en aucune façon le croire lorsqu'on me le présente. C'est ce que j'appelle la version *indexicale* du mécanisme. Elle est équivalente à croire que *je* peux survivre avec un cerveau artificiel.

Définition Méc-Ind : Une machine peut être moi. Je peux survivre avec un cerveau artificiel.

[2]Un zombie est un corps sans sujet. Si on veut marier le computationnalisme avec l'interprétation ontologique de Bohm (voir (Bohm and Hiley, 1993)) de la mécanique quantique, on doit reconnaître l'existence de l'inverse de zombie : des sujets sans corps, qui ont pourtant les mêmes raisons (des histoires computationnelles de "mesure" semblable) que *nous* de croire en leurs corps. Un mariage difficile donc. L' article de (Albert, 1986), (avec un zeste de (Albert, 1990)), montre qu'on peut "photographier" en principe (et partiellement) un de ses doppelgängers quantiques. Sans le computationnalisme on peut les prendre pour des zombies.

On obtient les versions digitales correspondantes de ces définitions en remplaçant *machine* par *machine digitale*.

D'un point de vue strictement logique on peut se convaincre que :

$$\text{Mec-Ind} \Rightarrow \text{Mec-Fort} \Rightarrow \text{Mec-Beh}$$

Preuve. En effet, si vous êtes une machine (Mec-Ind), alors il existe une machine qui peut penser (Mec-Fort), et il y a une machine qui peut avoir votre comportement (Mec-Beh).

De même on peut se convaincre aisément qu'aucune parmi ces implications n'est réversible:

$$\text{Mec-Beh} \not\Rightarrow \text{Mec-Fort} \not\Rightarrow \text{Mec-Ind}$$

Preuve. En effet, la *notion* de zombie illustre le fait que le mécanisme béhavioriste n'entraîne pas le mécanisme fort. Et la perplexité non-mécaniste du philosophe P_2 au chapitre 3 illustre que le mécanisme fort n'entraîne pas le mécanisme indexical.

Dans ce sens, notre hypothèse computationnelle est une version très forte du mécanisme. Il s'agit d'une version digitale du mécanisme indexical. Mais notre hypothèse computationnelle peut aussi paraître logiquement très faible.

D.3 Le computationnalisme est une hypothèse faible

Considérons l'histoire suivante:

> Le foie de Madame A. ne fonctionne plus. Madame B. vient de décéder dans un accident de la circulation. Par chance le foie de Madame B. est en bon état et son système immunitaire est compatible avec celui de madame A. Avec l'accord de la famille B., les médecins décident de greffer le foie de madame B. dans le corps de madame A. Après trois semaines de convalescence, madame A. peut rentrer chez elle, tout le monde s'accorde pour dire qu'elle a survécu à l'opération.

Cette histoire ne pose aucun problème philosophique majeur. En irait-il de même si on remplace dans cette histoire l'organe *cerveau* à la place de l'organe *foie*? Aujourd'hui les physiologistes estimeront que si Madame A. reçoit le cerveau de Madame B., il faudra plutôt convenir que c'est Madame B. qui survit à l'opération, malgré qu'elle hérite du corps de Madame A. Mais bien sûr ceci n'est pas évident a priori. Après tout il fut un temps où des hommes (parmi les anciens grecs) pensaient que le foie constituait le siège des sentiments, et le cerveau ne servait, à leur avis, qu'à refroidir le sang. Ils auraient sans doute estimé que Madame B. survit dans la première histoire, et que Madame A. survit dans la deuxième histoire. Les physiologistes ont cependant de sérieux indices au sujet du rôle du cerveau dans le psychisme, incluant les sentiments, les souvenirs, l'identité psychologique, etc. Le cerveau semble donc être privilégié par rapport aux autres organes. On ne peut pas survivre avec une greffe naturelle de cerveau. C'est plutôt le propriétaire du cerveau qui survit, avec une greffe de corps. A cause de cette dissymétrie entre le cerveau et les autres organes, certains philosophes concluent qu'on ne peut survivre avec une greffe —naturelle ou artificielle— du cerveau. Cette conclusion n'est pas valide. Considérons en effet l'histoire suivante.

> Je travaille à ma thèse sur mon ordinateur. Celui-ci tombe en panne. Après inspection on découvre que la panne est dans le clavier. On remplace mon clavier. Dans ce cas, ma thèse "a survécu" : pas de problème.
>
> Supposons cependant qu'après inspection on découvre que la panne se situe dans le disque dur. On remplace mon disque dur avec le disque dur du premier ordinateur disponible, ma thèse "survivra-t-elle" ?

Bien sûr que non, à moins qu'on ne parvienne à extraire la configuration de mon disque dur, et à réaménager le nouveau disque dur à partir de la configuration de l'ancien. Et il en est de même avec l'histoire de Madame A. Si l'hypothèse du mécanisme est correcte, Madame A. peut survivre avec le cerveau de Madame B. pour autant que :

- On possède une description suffisamment précise (= *au bon niveau*) du cerveau de Madame A. avant l'opération.
- On parvienne à "reconfigurer" le cerveau de Madame B. de telle façon qu'il porte la configuration correspondant à la description préalable du cerveau de Madame A.

Evidemment, une *telle* opération n'est pas éthiquement faisable (et encore moins techniquement, du moins aujourd'hui). Mais pour notre réflexion théorique, elle illustre que si on procède à une greffe complète du corps, celle-ci ne réussira qu'à condition d'opérer les substitutions à un *niveau* adéquat.

L'hypothèse computationnelle peut alors s'écrire de la façon suivante :

$$\text{il existe } n \text{ BEH}(n)$$

BEH est mis pour "behaviorisme", BEH(n) signifie que je survis lorsque les substitutions ont été faites à partir d'une description effectuée à un certain niveau n. On admet une description exclusivement en termes d'entrées et de sorties pour des composants jugés élémentaires *à un certain niveau*. La nature interne des composants élémentaires n'importe pas.

L'hypothèse mécaniste de ce travail peut être considérée comme *faible* en ce sens qu'aucune restriction n'est imposée au niveau, ni même à la *notion* de niveau utilisée. En fait la notion de niveau est vague, et nous aurons l'occasion de justifier pourquoi elle doit être nécessairement vague. Tout ce qui est demandé, c'est qu'une fois le niveau de description choisi, la description puisse être digitalisée, c'est-à-dire codée numériquement.

Exemples :

(a) Certains neurophysiologistes estiment que les neurones constituent les composants élémentaires du cerveau. "BEH(*niveau des neurones*)" signifie que je survis à la substitution de mes neurones par des dispositifs fonctionnellement équivalents aux neurones, c'est-à-dire ayant des sorties identiques pour des entrées identiques.

(b) Certains neurochimistes estiment que les molécules constituent les composants élémentaires. "BEH(*niveau des molécules*)" signifie que je survis à la substitution de mes molécules par des dispositifs fonctionnellement équivalents aux molécules, pour autant, évidemment que l'on parvienne à donner un sens à cette expression. Il n'est pas exclu qu'on puisse survivre avec des neurones

électroniques au cas où, par exemple, ces neurones simulent en détails les variations des concentrations des diverses molécules au sein du neurone biologique correspondant.

(c) Certains physiciens estiment qu'il faut tenir compte de l'état quantique de la matière constituant le cerveau. A cause de la nature particulière de la mécanique quantique ce cas se scinde en deux.

- Soit, en invoquant des arguments thermodynamiques, on se contente de ce que les physiciens appellent la trace partielle des observables nécessaires pour les mesures locales de la matière du cerveau. Dans ce cas BEH("niveau quantique 1") signifie que je survis si on substitue mon cerveau par un dispositif computationnel simulant un état quantique ayant la même trace partielle.
- Soit, pour ceux qui rejetteraient l'argument thermodynamique on exige une description quantique complètement fidèle du cerveau. Dans ce cas il faut tenir compte des phénomènes de non séparabilité quantique, et il faut tenir compte de toutes les particules ou les champs qui ont interagi directement ou indirectement, à un endroit ou à un autre, à un moment ou à un autre, avec les particules de mon cerveau. On peut montrer qu'alors la description quantique de mon cerveau inclut la description quantique de tout l'univers. BEH("niveau quantique 2") signifie que je survis à une duplication de "tout l'univers". Cela ne réfute pas a priori notre hypothèse computationnaliste sauf s'il devait apparaître que l'univers est décrit explicitement par une fonction non calculable. Ceci illustre que le computationnalisme peut être considéré comme une hypothèse très faible.

Appelons *cerveau généralisé* la partie de l'univers qu'il est nécessaire de simuler parfaitement pour que mon expérience subjective puisse apparaître. Une telle partie existe avec le computationnalisme. Le niveau n du mécanisme peut être scindé en un couple d'au moins deux paramètres (g, d).

Le paramètre g désigne un coefficient de granularité. Il décrit le niveau de résolution des détails dont il faut tenir compte pour la reconstitution.

Le paramètre d désigne le diamètre maximal de la zone qu'il faut dupliquer ou simuler parfaitement pour survivre à l'expérience de télétransport.

Il est possible qu'il faille tenir compte d'autres paramètres. Le computationnalisme exige uniquement l'existence du niveau de simulation digitale. Les paramètres g, d, et d'éventuels autres paramètres doivent être finiment descriptibles. Peu importe qu'ils soient gigantesques ou minuscules.

Une remarque méthodologique s'impose. Dans la narration de l'histoire du philosophe P. du chapitre 3, le paramètre g est tout à fait indéterminé . On ne nous donne aucune information sur les détails du cerveau dont le médecin a jugé opportun de tenir compte pour la reconstitution, à part une brève allusion au réseau neuronal et à sa topologie. Par contre le paramètre d est implicitement évalué à la taille du cerveau, puisque le médecin se contente de dupliquer celui-ci. On pourrait donc croire que j'utilise un computationnalisme plus fort (vis-à-vis du niveau) dans certaines expérience par la pensée, ce qui nuirait à la généralité de ma démonstration. A ceci j'ai deux réponses :

- Il est possible de modifier les expériences par la pensée de telle façon qu'elles ne dépendent d'aucune valeur particulière, pourvu qu'elles soient finies, de g et de d. Forcément les expériences par la pensée deviennent plus lourdes, beaucoup plus longues, et plus complexes. C'est dans un souci de ne pas compliquer les

arguments de façon prématurée que je me permettrai régulièrement de borner les paramètres associés au niveau n.

- Il n'a pas été *nécessaire* de modifier les expériences par la pensée. En effet, la démonstration de la proposition principale dépend exclusivement du déployeur universel (DU) et du graphe filmé (GF). On aurait pu introduire explicitement des hypothèses particulières limitant g et d pour les expériences de duplication ou de téléportation, et ensuite on aurait pu les éliminer explicitement avec le DU ou le GF. Mais cela aurait inutilement compliqué la présentation. Cela est dû au fait que, quelle que soit la valeur du niveau n (vu comme un code des paramètres g, d, ...), du fait que n est fini (avec le computationnalisme), le DU aboutira inéluctablement aux reconstitutions pertinentes. Et cela est encore plus clair avec la présentation du GF, puisqu'il rend la réalisation *concrète* du DU non pertinente pour la démonstration.

Les niveaux ne sont pas nécessairement totalement ou linéairement ordonnés. Il n'y a pas a priori de plus haut ou de plus bas niveau de substitution. Il existe par contre a priori des niveaux incomparables. Cela provient du fait qu'un niveau est déterminé par un ensemble de paramètres indépendants. On ne peut pas, par exemple comparer un niveau déterminé par un petit coefficient de granularité g_1 et un petit "diamètre" d_1 avec un niveau déterminé par un grand coefficient de granularité g_2 et un grand "diamètre" d_2. Il est donc opportun de garder à l'esprit une "échelle" qui a la forme d'un treillis ou d'un préordre.

Bibliography

Abramsky, S. (1987). *Domain theory and the logic of observable properties*. PhD thesis, University of London.

Agnes, C. and Rasetti, M. (1988). Chaos and undecidability: A group theoretical view. In Solomon, A. I., editor, *Advances in Statistical Mechanics*, pages 57–73. World Scientific Publ., Singapore.

Albert, D. Z. (1986). How to take a Photograph of another Everett World. In *New Techniques and Ideas in Quantum Measurement Theory*, volume 480, pages 498–502. Annals of the New York Academy of Sciences, New York.

Albert, D. Z. (1990). The quantum mechanics of self-measurement. In Zurek, W. H., editor, *Complexity, Entropy, and the Physics of Information*, volume VIII of *SFI Studies in the Sciences of Complexity*, pages 471–476. Addison Wesley.

Albert, D. Z. (1992). *Quantum Mechanics and Experience*. Harvard University Press, Cambridge, UK.

Albert, D. Z. and Loewer, B. (1989). Two no-collapse interpretations of quantum mechanics. *Noûs*, 12:121–138.

Alechina, N. (1994). Logic with Probabilistic Operators. Preprint.

Artemov, S. (1990). Kolmogorov's logic of problems and a provability interpretation of intuitionistic logic. In Parikh, R., editor, *Proceedings of the Third Conference on Theoretical Aspect of Reasoning about Knowledge (TARK 90)*. Morgan Kaufmann Publishers.

Aspect, A. (1976). Proposed experiment to test the non-separability of quantum mechanics. *Physical review*, D(14):1944–1951.

Aspect, A., Dalibard, J., and Roger, G. (1982). Experimental tests of Bell's inequalities using time-varying analysers. *Physical Review Letters*, 49:1804–1807.

Baertschi, B. (1992). *Les rapports de l'âme et du corps, Descartes Diderot et Maine de Biran*. Vrin, J., Paris.

Barnes, E. (1991). The causal history of computational activity: Maudlin and Olympia. *The Journal of Philosophy*, pages 304–316.

Bell, J. L. (1986). A new approach to quantum logic. *Brit. J. Phil. Sci.*, 37:83–99.

Bell, J. S. (1964). On the Einstein-Podolsky-Rosen paradox. *Physics*, 1:195–200.

Bell, J. S. (1987). *Speakable and unspeakable in quantum mechanics*. Cambridge University Press, Cambridge.

Benacerraf, P. (1967). God, the Devil, and Gödel. *The monist*, 51:9–32.

Bennett, C. H. (1988). Logical Depth and Physical Complexity. In Herken, R., editor, *The Universal Turing Machine A Half-Century Survey*, pages 227–258. Oxford University Press.

Birkhoff, G. and von Neumann, J. (1936). The Logic of Quantum Mechanics. *Annals of Mathematics*, 37(4):823–843.

Bitbol, M. and Ruhnau, E., editors (1994). *Now, Time and Quantum Mechanics*. Editions Frontières, Paris.

Blum, L. and Blum, M. (1975). Toward a mathematical theory of inductive inference. *Information and Control*, 28:125–155.

Bohm, D. (1951). *Quantum Theory*, pages 614–619. Prentice-Hall, New York.

Bohm, D. and Hiley, B. J. (1993). *The undivided universe*. Routledge, London and New York.

Boolos, G. (1979). *The unprovability of consistency*. Cambridge University Press, London.

Boolos, G. (1980a). On Systems of Modal Logic with Provability Interpretations. *Theoria*, 46(1):7–18.

Boolos, G. (1980b). Provability in Arithmetic and a Schema of Grzegorczyk. *Fundamenta Mathematicae*, 96:41–45.

Boolos, G. (1980c). Provability, Truth, and Modal Logic. *Journal of Philosophical Logic*, 9:1–7.

Boolos, G. (1993). *The Logic of Provability*. Cambridge University Press, Cambridge.

Brouwer, L. E. J. (1905). *Leven, Kunst en Mystiek*. Waltman, Delft, Holland. La Vie, l'Art et le Mysticisme.

Brouwer, L. E. J. (1983). Consciousness, philosophy and mathematics. In Benacerraf, P. and Putnam H., editors, *Philosophy of Mathematics*, pages 90–96. Cambridge University Press, Cambridge, second edition. première édition chez Prentice-Hall 1964.

Brown, J. R. (1991). *The laboratory of the mind*. Routledge, London.

Bub, J. (1997). *Interpreting the quantum world*. Cambridge University Press, Cambridge.

Burnyeat, M. (1991). Socrate et le jury: de quelques aspects paradoxaux de la distinction platonicienne entre connaissance et opinion vraie. In Canto-Sperber, M., editor, *Les paradoxes de la connaissance, essais sur le Ménon de Platon*, pages 237–251. Editions Odile Jacob, Paris.

Caillois, R. (1956). *L'incertitude qui vient des rêves*. Gallimard, France.

Carnap, R. (1966). *Philosophical Foundations of Physics*. Basic Books, Inc., New-York. Les fondements philosophiques de la physique, Armand Colin, Paris 1973.

Cartwright, H. M. (1993). On two arguments for the indeterminacy of personal identity. *Synthese*, 95:241–273.

Chellas, B. F. (1980). *Modal Logic, an introduction*. Cambridge University Press, Cambridge.

Chihara, C. S. (1972). On Alleged Refutations of Mechanism Using Gödel's Incompleteness Results. *The Journal of Philosophy*, "LXIX"(17):507–526.

Church, A. (1936). An Unsolvable Problem of Elementary Number Theory. *The American Journal of Mathematics*, 58. Aussi dans Davis 1965.

Church, A. and Kleene, S. C. (1937). Formal definitions in the theory of ordinal numbers. *Fundamenta Mathematicae*, 28:11–21.

Clauser, J. F., Horn, M. A., Shimony, A., and Holt, R. A. (1969). Proposed experiment to test hidden variable theories. *Physical Review Letters*, 23:880–883.

Dalla Chiara, M. L. (1976). A General Approach to non Distributive Logics. *Studia Logica*, XXXV:139–162.

Dalla Chiara, M. L. (1977a). Logical Self-Reference, Set Theoretical Paradoxes and the Measurement Problem in Quantum Mechanics. *Journal of Philosophical Logic*, 6:331–347.

Dalla Chiara, M. L. (1977b). Quantum Logic and Physical Modalities. *Journal of Philosophical Logic*, 6:391–404.

Dalla Chiara, M. L. (1985). Some Foundational Problems in Mathematics Suggested by Physics. *Synthese*, 62:303–315.

Dalla Chiara, M. L. (1986). Quantum logic. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic, Vol. III*, pages 427–469. D. Reidel Publishing Company, Dordrecht.

Davis, M., editor (1965). *The Undecidable*. Raven Press, Hewlett, New York.

Davis, M. (1982). Why Gödel Didn't Have Church's Thesis. *Information and Control*, 54:3–24.

Day, R. A. (1986). *How to write and publish a Scientific Paper*. Cambridge University Press, Cambridge, fourth edition.

de Broglie, L. (1957). *La théorie de la mesure en mécanique ondulatoire*. Gauthier-Villar, Paris.

Delahaye, J. P. (1991a). Le réalisme en mathématiques et en physique. *Pour la science*, 159.

Delahaye, J. P. (1991b). Thermodynamique et informatique théorique. *Pour la science*, 162:17–20.

Delahaye, J. P. (1994). *Informatique, complexité et hasard*. Hermès, Paris.

Dement, W. (1972). *Some must Watch While Some Must Sleep*. Seuil, traduction française 1981, Dormir, rêver, Paris.

Dement, W. and Kleitman, N. (1957). The relation of eye movements during sleep to dream activity: An objective method for the study of dreaming. *Journal of Experimental Psychology*, 53:339–346.

Dennett, D. C. (1978). *Brainstorms*. Harvester Press, Hassocks, Sussex.

Dennett, D. C. and Hofstadter, D., editors (1981). *Mind's I*. Basic Books, Inc., Publishers, New-York. Paru en français sous le titre "Vue de l'esprit", interEditions, Paris, 1987.

Descartes, R. (1953). *Oeuvres et Lettres*. Bibliothèque de la Pléiade. Gallimard, Paris.

d'Espagnat, B. (1971). *Conceptual Foundations of Quantum Mechanics*. Addison-Wesley, Reading, Mass. 2ème éd: 1976.

Deutsch, D. (1985). Quantum theory, the Church-Turing principle and the universal quantum computer. *Proc. R. Soc. Ac.*, 400:97–117.

Deutsch, D. (1986a). On Wheeler's Notion of "law without law" in Physics. *Foundations of Physics*, 16(6):565–572.

Deutsch, D. (1986b). Three connections between Everett's interpretation and experiment. In Penrose, R. and Isham, C., editors, *Quantum Concepts in Space and Time*, pages 215–225. Clarendon Press, Oxford.

Dewdney, A. K. (1988). *The Armchair Universe, An exploration of computer worlds*, pages 135–148. W. H. Freeman and Company, New-York.

Dewitt, B. S. (1970). Quantum mechanics and reality. *Physics Today*, 23(9). Aussi dans DeWitt and Graham, 1973.

DeWitt, B. S. and Graham, N., editors (1973). *The Many-Worlds Interpretation of Quantum Mechanics*. Princeton University Press, Princeton, New-Jersey.

d'Hervey de Saint Denys, J. M. L. (1867). *Les rêves et les moyens de les diriger*. Amyot, Paris. Réédition intégrale aux Editions Oniros 1995.

Dogen (1980). *Shôbôgenzô*. Argenteuil. 1232-1253.

Einstein, A. (1989). *Oeuvres choisies, 1, Quanta*. Editions du Seuil/Editions du CNRS, Paris.

Einstein, A., Podolski, B., and Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47:777–780.

Everett III, H. (1957). "Relative state" formulation of quantum mechanics. *Review of Modern Physics*, 9(3):454–462. Aussi dans DeWitt et Graham 1973.

Everett III, H. (1973). The theory of the universal wave functions. In DeWitt, B. and Graham, N., editors, *The Many-Worlds Interpretation of Quantum Mechanics*, pages 3–140. Princeton University Press, Princeton, New Jersey.

Fattorosi-Barnaba, M. and Amati, G. (1987). Modal Operators with Probabilistic Interpretations i. *Studia Logica*, XLVI(4):383–393.

Feferman, S. (1960). Arithmetisation of Metamathematics in a general Setting. *Fundamenta Mathematicae*, XLIX:35–92.

Ferret, S. (1993). *Le philosophe et son scalpel*. Editions de Minuit, Paris.

Feys, R. (1937a). Les logiques nouvelles des modalités. *Revue néoscolastique de philosophie*, 40:517–553.

Feys, R. (1937b). Les logiques nouvelles des modalités. *Revue néoscolastique de philosophie*, 41:217–252.

Flagg, R. C. (1985). Church's Thesis is consistent with epistemic arithmetic. In Shapiro, S., editor, *Intensional Mathematics*, Studies in Logic and the Foundations of Mathematics 113, pages 121–172. NorthHolland.

Ford, N. M. (1988). *When did I begin?* Cambridge University Press, Cambridge.

Gackenbach, J. and Laberge, S., editors (1988). *Conscious Mind, Sleeping Brain, Perspective on Lucid Dreams*. Plenum Press, New York London (USA).

Galouye, D. (1964). *Simulacron 3*. Bantam Books, Inc., New York. Editions française J'ai lu n 778.

Gandy, R. (1980). Church's Thesis and Principles for Mechanisms. In Barwise, J., Keisler, H. J., and Kunen, K., editors, *The Kleene Symposium*, pages 123–148. North Holland.

Gardner, M. (1980). *Le nombre aléatoire omega semble bien recéler les mystères de l'univers*. Bibliothèque pour la science, Pour la science S.A.R.L., Belin, Paris. Traduit de l'américain dans "Le monde Mathématique de Martin Gardner".

Gardner, M. (1996). *The night is large*. Penguin Books, London.

Gell-mann, M. and Hartle, J. H. (1990). Quantum Mechanics in the light of Quantum Cosmology. In Zurek, W. H., editor, *Complexity, Entropy, and the Physics of Information*, volume VIII of *SFI Studies in the Sciences of Complexity*, pages 425–458. Addison Wesley.

Gödel, K. (1931). Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatsh., Math. Phys.*, 38:173–198. Traduction américaine dans Davis 1965, page 5+.

Gödel, K. (1933). Eine interpretation des intuitionistischen aussagenkalküls. *Ergebnisse eines Mathematischen Kolloquiums*, 4:39–40.

Gödel, K. (1946). Remarks before the princeton bicentennial conference on problems in mathematics. In (Davis, 1965), pages 84–88.

Gödel, K. (1972). Communication orale à R. Rucker. Dans Rucker 1982, page 164.

Gold, E. M. (1965). Limiting recursion. *Journal of Symbolic Logic*, 30(1):27–48.

Goldblatt, R. I. (1974). Semantic Analysis of Orthologic. *Journal of Philosophical Logic*, 3:19–35. Aussi dans Goldblatt 1993, page 81–97.

Goldblatt, R. I. (1978). Arithmetical Necessity, Provability and Intuitionistic Logic. *Theoria*, 44:38–46. Aussi dans Goldblatt 1993, page 105–112.

Goldblatt, R. I. (1993). *Mathematics of Modality*. CSLI Lectures Notes, Stanford California.

Goodstein, R. L. (1963). The significance of incompleteness theorems. *British Journal for the Philosophy of Science*, 14:208–220.

Graham, N. (1973). The measurement of relative frequency. In DeWitt, B. and Graham, N., editors, *The Many-Worlds Interpretation of Quantum Mechanics*, pages 229–252. Princeton University Press, Princeton, New Jersey.

Gregory, R. L., editor (1987). *The Oxford Companion to The Mind*, pages 200–201. Oxford University Press, Oxford, UK.

Grim, P. (1991). *The Incomplete Universe*. The MIT Press, Cambridge, USA.

Griswold, C. L. (1986). *Self-Knowledge in Plato 's Phaedrus*. Yale University Press, New Haven and London.

Grzegorczyk, A. (1964). A philosophicaly plausible formal interpretation of intuitionistic logic. *Indagationes Math.*, 26:596–601.

Grzegorczyk, A. (1967). Some relational systems and the associated topological spaces. *Fundamenta Mathematicae*, LX:223–231.

Guenancia, P. (1986). *Descartes*. Bordas, Paris.

Hacking, I. (1975). *The Emergence of Probability*. Cambridge University Press, Cambridge.

Hardegree, G. M. (1976). The Conditional in Quantum Logic. In Suppes, P., editor, *Logic and Probability in Quantum Mechanics*, volume 78 of *Synthese Library*, pages 55–72. D. Reidel Publishing Company, Dordrecht-Holland.

Hartle, J. B. (1968). Quantum mechanics of individual systems. *American Journal of Physics*, 36(704):27–48.

Hartmanis, J. (1972). Relations between diagonalisation, proof systems, and complexity gaps. Computer Science Department, Cornell University, Ithaca, New York 14853, USA.

Heyting, A. (1980). *Intuitionism, an Introduction*. North-Holland, Amsterdam, third edition. Edition originale: 1956.

Hobson, J. A. and McCarley, R. (1977). The brain as a dream-state generator: An activation-synthesis of the dream process. *American Journal of Psychiatry*, 134:1335–1348.

Hughes, R. I. G. (1989). *The Structure and Interpretation of Quantum Mechanics*. Harvard University Press, Cambridge.

Huisman, D. and Vergez, A. (1966). *Métaphysique*. Fernand Nathan, Nancy, France.

Humes, D. (1987). *A Treatise of Human Nature*. Fontana/Collins, Glasgow. Edition originale: Londres 1739.

Isham, C. J. (1994). Quantum logic and the histories approach to quantum theory. *J. Math. Phys.*, 35(5):2157–2185.

Jacob, F. and Monod, J. (1961). Genetics regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3:318–356.

Jammer, M. (1974). *The Philosophy of Quantum Mechanics*. John Wiley & Sons, New York.

Jouvet, M. (1992). *Le sommeil et le rêve*. Editions Odile Jacob, Paris.

Kafatos, M., editor (1989). *Bell's Theorem, Quantum Theory and Conceptions of the Universe*. Kluwer Academic Publishers, Dordrecht.

Kaplan, D. and Montague, R. (1961). A paradox regained. *Journal of Formal Logic*, 1:79–90.

Kelly, K. (1993). Learning theory and descriptive set theory. *Journal of Logic and Computation*, 3(1):27–45.

Kent, A. (1990). Against many world interpretations. *International Journal of Modern Physics A*, 5(9):1745–1762.

Kleene, S. C. (1952). *Introduction to Metamathematics*. North-Holland, Amsterdam.

Kleene, S. C. (1965). General recursive functions of natural numbers. In (Davis, 1965), pages 237–253. Parution originale: 1936.

Kolmogorov, A. (1932). Zur deutung der intuitionistischen logik. *Math. Zeitschr.*, 35:58–65.

Kripke, S. A. (1963a). Semantical analysis of modal logic i: Normal propositional calculi. *Zeit. Math. Logik. Grund.*, 9:67–96.

Kripke, S. A. (1963b). Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16:83–94.

Kurtz, S. A. (1983). On the random oracle hypothesis. *Information and Control*, 57:40–47.

Kuznetsov, A. V. and Muravitsky, A. Y. (1977). Magari algebras. *Fourteenth All-Union Algebra Conf., Abstract part 2: Rings, Algebraic Structures*, pages 105–106. En Russe.

Laberge, S. (1991). *Le rêve lucide*. collection Rêvéveil. Oniros, Paris. Traduction française.

Ladrière, J. (1957). *Les limitations internes des formalismes*. Nauwelaerts/Gauthier-Villars, Paris/Louvain.

Largeault, J., editor (1992). *Intuitionisme et théorie de la démonstration*. Vrin, J., Paris.

Letovsky, S. (1987). Ecclesiastes: A report from the battlefields of the mind-body problem. *AI Magazine*, 8(3):63–69.

Lewis, D. (1986). *On the Plurality of Worlds*. Basil Blackwell, Oxford.

Löb, M. H. (1955). Solution of a problem of Leon Henkin. *Journal of Symbolic Logic*, 20:115–118.

London, F. and Bauer, E. (1939). *La théorie de l'observation en mécanique quantique*. Hermann et Cie, Paris.

Lucas, J. R. (1961). Minds, Machines and Gödel. *Philosophy*, 36:112–127.

Lucas, J. R. (1968). Satan Stultified: A Rejoinder to Paul Benacerraf. *The Monist*, 52:145–158.

Lucas, J. R. (1970). This gödel is killing me. *Philosophia*, 6:145–148.

Lucas, J. R. (1971). Metamathematics and the philosophy of mind, a rejoinder. *Philosophy of Sciences*, XXXVIII(273):310–313.

Mackay, D. M. (1960). On the logical indeterminacy of a free choice. *Mind*, 69(273):31–40.

Magari, R. (1975). Representation and duality theory for diagonalizable algebras. *Studia Logica*, XXXIV(4):305–313.

Malcolm, N. (1959). *Dreaming*. Routledge Kegan Paul ltd., London.

Malcolm, N. (1968). The conceivability of mechanism. *Phil. Review*, 77:45–72.

Marchal, B. (1978). Introductions au théorème de Bell. Technical report, Laboratoire de Cosmologie, ULB, Plaine des Manoeuvres, Bâtiment NO.

Marchal, B. (1983). L'elaboratore e'un grafo. *L'insegnamento della matematica e delle scienze integrate*, 6(2):43–58.

Marchal, B. (1988). Informatique théorique et philosophie de l'esprit. In *Actes du 3ème colloque international de l'ARC*, pages 193–227, Toulouse.

Marchal, B. (1990). Des fondements théoriques pour l'intelligence artificielle et la philosophie de l'esprit. *Revue Internationale de Philosophie*, 1(172):104–117.

Marchal, B. (1991). Mechanism and personal identity. In De Glas, M. and Gabbay, D., editors, *Proceedings of WOCFAI 91*, pages 335–345, Paris. Angkor.

Marchal, B. (1992a). Amoeba, planaria, and dreaming machines. In Bourgine, P. and Varela, F. J., editors, *Artificial Life, towards a practice of autonomous systems, ECAL 91*, pages 429–440. MIT Press.

Marchal, B. (1992b). Sciences et science-fiction: Simulacron 3. In *La littérature fantastique pour les jeunes de 8 18 ans*, pages 61–68. Inspection de l'éducation permanente et de la culture, Ville de Bruxelles. Transcription des communications des 28, 29, et 30 avril 1992.

Marchal, B. (1995). Conscience et Mécanisme. Technical Report TR/IRIDIA/95, Brussels University.

Maudlin, T. (1989). Computation and Consciousness. *The Journal of Philosophy*, pages 407–432.

Maudlin, T. (1994). *Quantum Non-Locality and Relativity*, volume 13 of *Aristotelian Society Series*. Blackwell, Oxford.

McKinsey, J. C. and Tarski, A. (1948). Some theorems about the sentential calculi of Lewis and Heyting. *Journal of Symbolic Logic*, 13:1–15.

Mittelstaedt, P. (1978). *Quantum Logic*. D. Reidel Publishing Company, Dordrecht, Holland.

Moore, C. M. (1990). Unpredictability and undecidability in dynamical systems. *Physical Review Letters*, 64(20):2354–2357.

Moravec, H. (1988). *Mind Children*. Harvard University Press, Cambridge.

Myhill, J. (1952). Some philosophical implications of mathematical logic. *The review of Metaphysics*, VI(2).

Myhill, J. (1960). Some remarks on the notion of proof. *Journal of Philosophy*, 57:461–471.

Myhill, J. (1964). Abstract theory of self-reproduction. In Mesarovic, M. D., editor, *Views on general systems theory*, pages 106–118. Wiley, New York.

Nagel, T. (1987). *What does it all mean?* Oxford University Press, New York.

Nagel, T. (1994). Consciousness and objective reality. In Warner, R. and Szubka, T., editors, *The Mind-Body Problem*, pages 63–68. Basil Blackwell, Cambridge, USA.

Nozick, R. (1981). *Philosophical Explanations*. Clarendon Press, Oxford.

Ouzoulias, A. (1989). *La conscience*. Editions Quintette, Paris.

Pacherie, E. (1993). *Naturaliser l'intentionnalité*. PUF, Paris.

Parfit, D. (1984). *Reasons and Persons*. Clarendom Press, Oxford.

Pauling, L. (1966). *Chimie générale*. Dunod, Paris, seconde edition. Traduction française de R. Pâris.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford University Press, Oxford.

Penrose, R. (1990). Précis of The Emperor's New Mind : concerning computers, minds and the laws of physics. *Behavioral and Brain Sciences*, 13(4):643–705.

Penrose, R. (1994). *Shadows of the Mind*. Oxford University Press, Oxford.

Penrose, R. and Isham, C. J., editors (1986). *Quantum Concepts in Space and Time*. Clarendon Press, Oxford.

Platon (1950). *Théétète ou de la science*, pages 83–192. Oeuvre de la pléiade. Editions Gallimard, Paris.

Posner, C. H. (1980). A survey of non-r.e. degrees $\leq O'$. In Drake, F. R. and Wainer, S. S., editors, *Recursion Theory: its generalisation and applications*, pages 52–109. Cambridge University Press.

Post, E. (1922). Absolutely unsolvable problems and relatively undecidable propositions : Account of an anticipation. In (Davis, 1965), pages 338–433.

Reinhardt, W. N. (1985). Absolute version of incompleteness theorems. *Noûs*, 19:317–346.

Reinhardt, W. N. (1986). Epistemic theories and the interpretation of Gödel's incompleteness theorems. *Journal of Philosophical Logics*, 15:427–474.

Rucker, R. (1982). *Infinity and the Mind*. The Harverster Press, Brighton, Sussex, Great Britain.

Russell, B. and Whitehead, A. N. (1970). *Principia Mathematica*. Cambridge University Press, Cambridge. Paperback Edition to *56, (Edition originale: 1910).

Schrödinger, E. (1992). *Physique quantique et représentation du monde*. Seuil, Paris. Science et Humanisme 1950, La situation actuelle en mécanique quantique 1935.

Shapiro, S. (1985). Epistemic and intuitionistic arithmetic. In Shapiro, S., editor, *Intensional Mathematics*, Studies in Logic and the Foundations of Mathematics 113, pages 11–46. NorthHolland, Amsterdam.

Shimony, A. (1963). Role of the Observer in Quantum Theory. *Am. J. of Physics*, 31(6):755–773.

Slezak, P. (1982). Gödel's theorem and the mind. *Brit. J. Phil. Sci.*, 33:41–52.

Slezak, P. (1983). Descartes 's diagonal deduction. *Brit. J. Phil. Sci.*, 34:13–36.

Smets, P. (1991). Probability of Provability and Belief Functions. *Logique Analyse*, 133:177–195.

Smiley, T. J. (1963). The logical basis of ethics. *Acta Philosophica Fennica*, 16:237–246.

Smith, C. H. (1980). Applications of classical recursion theory to computer science. In Drake, F. R. and Wainer, S. S., editors, *Recursion Theory: its generalisation and applications*, pages 236–247. Cambridge University Press.

Smoryński, P. (1981). Fifty Years of Self-Reference in Arithmetic. *Notre Dame Journal of Formal Logic*, 22(4):357–374.

Smoryński, P. (1985). *Self-Reference and Modal Logic*. Springer Verlag, New York.

Smullyan, R. (1987). *Forever Undecided*. Knopf, New York.

Solovay, R. M. (1976). Provability Interpretation of Modal Logic. *Journal of Mathematics*, 25:287–304.

Squires, E. (1990). *Conscious Mind in the Physical World*. Adam Hilger, Bristol.

Stalnaker, R. C. (1984). *Inquiry*. The MIT Press, Cambridge, Massachusetts.

Stapp, H. P. (1993). *Mind, Matter and Quantum Mechanics*. Springer-Verlag, Berlin.

Thomas, R. (1978). Logical analysis of systems comprising feedback loops. Université libre de Bruxelles, Preprint.

Thomas, R. and Van Ham, P. (1974). Analyse formelle de circuits de régulation génétique, le contrôle de l'immunité chez les bactériophages lambdoïdes. *Biochimie*, 56(11–12):1529+.

Thomason, R. H. (1980). A note on syntactical treatment of modality. *Synthese*, 44:391–395.

Turing, A. (1936). On computable numbers with an application to the entscheidungsproblem. *Proc. London Math. Soc.*, 42:230–265. Aussi dans Davis 1965, page 116–154.

Turing, A. (1939). Systems of logic based on ordinals. *Proc. London Math. Soc.*, 45:161–228. Aussi dans Davis 1965, page 155–222.

Tye, M. (1995). *Ten problems of consciousness*. The MIT Press, Cambridge, Massachusetts.

van Fraassen, B. C. (1974). The labyrinth of quantum logic. In Cohen, R. and Wartosky, M., editors, *Boston Studies of Philosophy of Sciences*, volume 13, pages 224–254. Reidel, Dordrecht.

van Stigt, W. P. (1990). *Brouwer's Intuitionism*, volume 2 of *Studies in the history and philosophy of Mathematics*. North Holland, Amsterdam.

Vickers, S. J. (1989). *Topology via Logic*. Cambridge Tracts in Theoretical Computer Science 5. Cambridge University Press, Cambridge.

Visser, A. (1985). *Aspects of Diagonalization and Provability*. PhD thesis, University of Utrecht, Department of Philosophy, The Nederland.

Wainwright, R. (1974). Life is universal. In *Proceedings of the Winter Simulation Conference, ACM*, Washington DC.

Wang, H. (1957). A variant to Turing's theory of computing machines. *Journal of the ACM*, 4(1).

Wang, H. (1974). *From Mathematics to Philosophy*. Routledge Kegan Paul, London.

Warner, R. and Szubka, T., editors (1994). *The Mind-Body Problem*. Basil Blackwell, Cambridge, USA.

Watson, J. D. (1968). *Biologie moléculaire du gène*. Ediscience, Paris. Edition française dirigée par François Gros.

Webb, J. C. (1980). *Mechanism, Mentalism and Metamathematics: An essay on Finitism*. D. Reidel Publishing Company, Dordrecht, Holland.

Webb, J. C. (1983). Gödel's theorems and church's thesis: a prologue to mechanism. In Cohen, R. S. and Wartofsky, M. W., editors, *Language, logic, and method*, pages 309–353. D.Reidel Publishing Company, Dordrecht, Holland.

Wheeler, J. A. (1994). *At Home in the Universe*. The Aip Press, New York.

Wigner, E. (1967a). *Symmetries and Reflections*. Indiana University Press, Bloomington.

Wigner, E. (1967b). *The Unreasonable Effectiviness of Mathematics in the Natural Sciences*, pages 222–237. In (Wigner, 1967a).

Wittgenstein, L. (1965). *De la certitude*. Gallimard, Paris. Edition originale: 1951.

Yam, P. (1993). A bus for Scotty. *Scientific American*. Juin.

Zeh, H. D. (1990). Quantum Measurements and Entropy. In Zurek, W. H., editor, *Complexity, Entropy, and the Physics of Information*, volume VIII of *SFI Studies in the Sciences of Complexity*, pages 405–422. Addison Wesley.