ROGER

PENROSE

LES OMBRES DE L'ESPRIT

A la recherche d'une science de la conscience

LES OMBRES DE L'ESPRIT

ROGER PENROSE

LES OMBRES DE L'ESPRIT

A la recherche d'une science de la conscience

Traduit de l'anglais par Christian Jeanmougin

Ouvrage traduit avec le concours du Centre National du Livre

INTEREDITIONS

L'édition originale de cet ouvrage a été publiée en anglais par Oxford University Press, sous le titre *Shadows of the Mind. A Search for the Missing Science of Consciouness.* Traduction publiée en accord avec Oxford University Press.

This translation of *Shadows of the Mind* originally published in English in 1994 is published by arrangement with Oxford University Press.

ISBN 2 7296 0558 4

Table des matières

Préface

Ce livre peut, en un sens, être considéré comme une suite à *l'Esprit, l'ordinateur et les lois de la physique* (qui sera noté en abrégé EOLP). En fait, il développe le thème initié dans EOLP, mais ce que je dirai ici peut être lu tout à fait indépendamment de ce premier livre. À l'origine, j'ai voulu revenir sur ce sujet en partie pour répondre en détail à un certain nombre de questions et de critiques concernant EOLP. Toutefois, les arguments que je développe ici forment un tout autonome et explorent des idées nouvelles allant bien au-delà de celles contenues dans EOLP. L'une des thèses centrales de ce livre était que la conscience nous permet d'accomplir des actions irréductibles à toute forme d'activité de calcul. Cette idée était cependant plus ou moins présentée comme une simple hypothèse, et je suis resté assez vague sur les processus pouvant se ranger sous la rubrique « activité de calcul ». Ce nouveau livre offre ce que je crois être une argumentation bien plus rigoureuse et féconde, qui s'applique à tous les processus de calcul. En outre, je suggère ici, avec plus de crédibilité que dans EOLP, l'existence, dans le cerveau, d'un mécanisme dont l'action physique non réductible au calcul sous-tendrait notre comportement conscient.

Mon argumentation comprend deux aspects distincts. Le premier est essentiellement négatif dans la mesure où il s'oppose fermement au point de vue général selon lequel nos états mentaux conscients — dans leurs diverses manifestations — admettraient en principe une interprétation en termes de modèles numériques. Le second aspect, plus constructif, correspond à une authentique recherche des moyens par lesquels — compte tenu des contraintes imposées par les faits purement scientifiques — le cerveau pourrait utiliser des principes physiques subtils et encore largement inconnus pour accomplir ces actions non réductibles au calcul.

Conformément à cette dichotomie, ce livre se présente en deux parties. La première est une analyse complète et détaillée, étayant fortement ma thèse selon laquelle la conscience, telle qu'elle se manifeste dans cet attribut humain

qu'est la « compréhension », réalise des choses qu'un simple calcul ne peut accomplir. Je précise tout de suite que le mot « calcul » inclut à la fois les systèmes « descendants » opérant selon des procédures algorithmiques spécifiques et parfaitement maîtrisées, et les systèmes « ascendants » programmés avec plus de souplesse pour permettre une amélioration des performances avec l'expérience. La première partie s'articule essentiellement autour du célèbre théorème de Gödel et livre une analyse détaillée des conséquences fondamentales de ce théorème pour le problème de la conscience. Cette analyse élargit considérablement la portée des arguments avancés par Gödel lui-même, par Nagel et Newman, et par Lucas ; en outre, elle réfute en détail toutes les objections dont j'ai eu connaissance et expose minutieusement quelques arguments démontrant que les systèmes ascendants (pas plus que les systèmes descendants) ne pourront jamais acquérir une véritable intelligence. Cette analyse permet de conclure que la pensée consciente met nécessairement en jeu des éléments qu'aucune procédure de calcul ne peut même *simuler*, et qu'un calcul, par lui-même, peut encore moins susciter des intentions ou des sentiments conscients. Ainsi, l'esprit est nécessairement indescriptible en termes de calcul.

Dans la deuxième partie, mes arguments font intervenir la physique et la biologie. Bien qu'ils soient par endroits indéniablement plus spéculatifs que la discussion rigoureuse de la première partie, ils tentent véritablement de comprendre comment de telles actions indépendantes de tout calcul pourraient survenir dans le cadre de lois physiques scientifiquement compréhensibles. J'expose de nouveau les principes fondamentaux de la mécanique quantique — il n'est pas nécessaire que le lecteur possède une connaissance préalable de cette théorie. Les énigmes, les paradoxes et les mystères de cette discipline sont analysés en profondeur à l'aide d'un certain nombre d'exemples nouveaux qui illustrent de manière spectaculaire le rôle important joué par la non-localité et la multiplicité des états observables, ainsi que par l'influence de l'appareillage de mesure sur les phénomènes observés. J'affirmerai avec force la nécessité d'introduire, à un niveau bien précis, un changement radical de notre vision quantique du monde. (Ces idées sont intimement liées à des travaux récents réalisés entre autres par Ghirardi et Diósi.) Les idées que j'exposerai ici diffèrent beaucoup de celles présentées dans EOLP.

Je vais en effet affirmer l'existence au niveau quantique d'une non-calculabilité qui explique celle de nos actes conscients, et que le niveau où cette non-calculabilité physique se manifeste avec le plus d'ampleur joue un rôle déterminant pour l'activité cérébrale. C'est ici que mes propositions diffèrent le plus radicalement de celles avancées dans EOLP. J'affirme que si les signaux neuronaux peuvent effectivement avoir un comportement déterministe classique, le contrôle des liaisons synaptiques entre neurones s'effectue à un niveau plus profond qui semble correspondre à une importante activité physique située à la frontière séparant le monde classique du monde quantique. L'interprétation que je propose met en jeu l'existence, dans les microtubules des cytosquelettes neuronaux, d'un comportement quantique cohérent à grande échelle (en accord avec des suggestions avancées par Fröhlich). L'idée est que

cette activité quantique serait liée, d'une façon non réductible au calcul, à une action assimilable à un calcul qui, selon Hameroff, se déroule dans ces microtubules.

Si, en plusieurs endroits, je montrerai que les théories actuelles sont impuissantes à fournir une interprétation scientifique des états mentaux de l'être humain, je ne crois pas cependant que le phénomène de la conscience restera définitivement inaccessible à la science. J'affirmerai au contraire, comme je l'ai fait dans EOLP, qu'il existe une voie menant à une compréhension scientifique des phénomènes mentaux, et que nous ne pourrons trouver cette voie sans procéder au préalable à un examen approfondi de la nature de la réalité physique. Il importe selon moi que le lecteur consciencieux, qui désire savoir comment on peut appréhender d'un point de vue matérialiste un phénomène aussi mystérieux que l'esprit, comprenne bien l'étrangeté des règles qui probablement gouvernent *réellement* cette « matière » présente dans notre monde physique.

Comprendre est en définitive l'objectif de la science — et la science est loin de se réduire à une simple exécution de calculs mécaniques.

R.P.
Oxford
Avril 1994

Note au lecteur

Les diverses parties de ce livre ne présentent pas toutes le même niveau de difficulté. Les plus techniques d'entre elles sont les appendices A et C, mais la plupart des lecteurs ne perdront pas le fil conducteur de mon exposé en les ignorant purement et simplement. On pourrait dire la même chose des parties les plus techniques du chapitre 2 et, certainement, du chapitre 3. Ces passages sont essentiellement destinés aux lecteurs qui veulent se persuader de la solidité des arguments que j'oppose à toute modélisation de la compréhension humaine fondée uniquement sur le calcul. Au lecteur plus confiant (ou plus pressé) qui se satisferait des points essentiels de mon argumentation, je conseillerais de lire le dialogue imaginaire de la section 3.23, précédé de préférence du chapitre 1, des sections 2.1 à 2.5 et de la section 3.1.

Les mathématiques les plus délicates sont celles qui interviennent lors de la discussion de la mécanique quantique. Elles concernent notamment la description de l'espace de Hilbert donnée aux sections 5.12 à 5.18, et plus particulièrement les discussions figurant aux sections 6.4 à 6.6 et centrées sur la matrice densité, un concept important pour comprendre pourquoi nous aurons un jour besoin d'une version *améliorée* de la théorie quantique ! D'une manière générale, je conseillerais aux lecteurs non mathématiciens (et même, d'ailleurs, aux lecteurs mathématiciens) de sauter toute expression mathématique qui leur apparaît rebutante — une fois acquise la conviction qu'un examen plus approfondi de cette expression ne leur apporterait aucun éclaircissement supplémentaire. S'il est vrai qu'on ne peut pleinement apprécier les subtilités de la mécanique quantique sans posséder une connaissance préalable de ses élégants mais mystérieux fondements mathématiques, on peut cependant parvenir à une vue d'ensemble correcte de cette théorie en ignorant totalement ces aspects.

Je dois en outre m'excuser auprès de mes lecteurs à propos d'un point tout à fait différent. On le verra, il me sera parfois nécessaire de me référer à une personne « abstraite », par exemple un « observateur » ou un « physicien », et

je conçois volontiers que l'on puisse s'offusquer d'un emploi aussi systématique du genre masculin. Il est clair que cela n'implique aucunement de ma part un présupposé sur le sexe de cette personne — loin de moi une telle pensée ! Je ne fais que me conformer ici à un usage contre lequel on peut difficilement aller. Un emploi alterné de termes comme « observateur », « observatrice » ou « physicien », « physicienne » n'introduit selon moi aucun avantage sur le plan humaniste, et toute précaution oratoire systématique telle que « il ou elle » ou « elle ou il » s'avérerait rapidement ennuyeuse.

J'ai donc décidé, lorsque je me référerai à cette personne abstraite, d'utiliser les pronoms masculins « il », « lui », ce qui ne signifie pas que cette personne ne peut selon moi qu'être de sexe masculin. Ainsi, j'espère par exemple que mes lectrices ne s'offenseront pas en constatant que l'être (abstrait) vivant sur α du Centaure est masculin, de même que les individus totalement impersonnels apparaissant aux sections 1.15, 4.4, 6.5, 6.6 et 7.10.

Prologue

Jessica se sentait toujours un peu nerveuse lorsqu'elle pénétrait dans cette partie de la grotte. « Papa, et si ce gros rocher, là, coincé entre les autres, tombait par terre ? Est-ce qu'il ne risquerait pas de bloquer la sortie ? Nous ne pourrions plus jamais rentrer à la maison. »

« Il pourrait, mais il ne le fera pas », répondit son père d'un ton légèrement irrité et inutilement brusque, manifestement plus intéressé par l'acclimatation de ses divers échantillons végétaux à l'humidité et à l'obscurité régnant dans cet endroit reculé de la grotte.

« Mais comment tu sais qu'il ne va pas tomber ? » insista Jessica.

« Ce rocher est probablement là depuis des milliers et des milliers d'années. Il ne va pas tomber juste au moment où nous sommes là. »

Jessica n'était guère rassurée par cette réponse. « Mais s'il doit tomber un jour, plus il est resté longtemps là-haut, plus il a de chances de tomber maintenant, non ? »

Son père s'arrêta de titiller les plantes pour la regarder avec un léger sourire. « Non, ça ne se passe pas du tout comme ça. » Son sourire s'accentua mais se fit plus intérieur. « En fait, on pourrait même dire que plus il est resté longtemps là-haut, *moins* il a de chances de tomber quand nous sommes là. » Jugeant l'explication suffisante, il se repencha sur ses plantes.

Jessica haïssait son père quand il était comme ça — non, bien sûr ; elle l'aimait tout le temps, plus que toute autre chose ou que toute autre personne, mais elle aurait préféré qu'il ne fût jamais de cette humeur-là. Elle savait que ça avait quelque chose à voir avec son métier de scientifique ; pourtant, elle ne comprenait pas. Elle espérait même elle aussi devenir un jour scientifique, mais si elle le devenait, elle veillerait à n'être jamais d'une humeur pareille.

Du moins cessa-t-elle de craindre que le rocher ne tombât en bloquant l'entrée de la grotte. Elle voyait que son père n'avait pas peur, et la confiance qu'elle avait en lui la rendait elle-même confiante. Elle ne comprenait pas l'explication de son père, mais elle savait qu'il avait toujours raison pour ce

genre de choses — du moins presque toujours. Il y avait eu cette discussion sur les horloges en Nouvelle-Zélande, lorsque Maman avait dit une chose mais que Papa soutenait que c'était l'inverse qui était vrai. Puis, trois heures plus tard, Papa était sorti de son bureau en disant qu'il était désolé, qu'il avait tort, que c'était elle qui avait raison depuis le début ! Ça c'était drôle ! « Je suis sûre que Maman elle aussi aurait pu devenir scientifique si elle l'avait voulu » pensa Jessica, « et elle, elle n'aurait jamais été de cette humeur étrange. »

Jessica prit soin de poser sa question suivante à un moment où son père venait de finir ce qu'il était en train de faire et n'avait pas encore commencé ce qu'il allait faire ensuite. « Papa ? Je sais que le rocher ne va pas tomber. Mais imagine qu'il tombe et que nous soyons enfermés ici le reste de notre vie. Toute la grotte deviendrait très sombre, non ? Est-ce que nous pourrions respirer ? »

« Quelle idée saugrenue ! » répondit son père. Il regarda attentivement la forme et la taille du rocher, puis l'ouverture de la grotte. « Mmmm » fit-il, « oui, je pense que ce rocher boucherait presque parfaitement l'entrée. Mais il laisserait certainement un petit espace où l'air pourrait circuler, de sorte que nous ne suffoquerions pas. Quant à la lumière, eh bien je pense que cette petite fissure arrondie en haut de l'entrée la laisserait pénétrer, mais la grotte serait très sombre — bien plus sombre qu'elle ne l'est maintenant. Mais je suis presque certain que nous pourrions voir parfaitement, une fois que nous nous serions habitués. Ce ne serait pas très agréable, bien sûr ! Mais je vais te dire une chose : si je devais vivre ici le reste de ma vie avec quelqu'un, je préférerais que ce soit avec ma merveilleuse Jessica plutôt qu'avec toute autre personne au monde — et avec Maman, bien sûr. »

Jessica se rappela pourquoi elle aimait tant son père ! « Je veux que Maman soit là aussi, dans ma prochaine question, parce que je vais supposer que le rocher soit tombé avant que je sois née et que toi et Maman m'avez eue dans la grotte, et que j'ai grandi avec vous ici... et que nous sommes restés vivants en mangeant toutes tes drôles de plantes. »

Son père la regarda un peu curieusement, mais ne dit rien.

« Alors je n'aurais jamais connu d'autre vie que celle ici, dans cette grotte Comment aurais-je pu savoir à quoi ressemble le monde extérieur ? Aurais-je pu savoir qu'il y a des arbres, et des oiseaux, et des lapins, et d'autres choses ? Bien sûr, tu pourrais me *raconter* ces choses, parce que tu les aurais connues toi-même avant d'avoir été enfermé, mais comment *moi* je les aurais connues — je veux dire, comment je les aurais réellement connues *moi-même*, plutôt qu'en croyant simplement ce que tu me dirais ? »

Son père s'arrêta et réfléchit quelques minutes. Puis il dit : « Eh bien, je suppose que de temps en temps, un jour où il ferait beau, un oiseau passerait exactement dans la ligne entre la fissure et le Soleil, et que nous pourrions alors voir son ombre se projeter sur la paroi arrière de la grotte. Bien sûr, cette ombre serait légèrement déformée parce que la paroi est assez irrégulière, mais nous pourrions apprendre à corriger cela. Si la fissure était assez ronde et assez petite, l'oiseau projetterait une ombre bien nette ; sinon, il nous faudrait faire d'autres corrections. Alors, si ce même oiseau passait de nombreuses fois, nous

pourrions commencer à nous faire une image très précise de ce à quoi il ressemble, comment il vole, et ainsi de suite, simplement à partir de son ombre. Puis, lorsque le Soleil serait bas dans le ciel, il pourrait y avoir un arbre, convenablement placé entre le Soleil et notre fissure, avec ses feuilles balancées par le vent, de sorte que l'on pourrait également obtenir une image de cet arbre, elle aussi à partir de son ombre. Et peut-être que de temps en temps un lapin sauterait jusqu'au niveau de la fissure, de sorte que lui aussi nous pourrions commencer à nous le représenter à partir de son ombre. »

« C'est intéressant » dit Jessica. Elle réfléchit quelques instants puis dit : « Tu crois que nous pourrions faire une vraie découverte scientifique, même si nous restions coincés ici dans cette grotte ? Suppose que nous ayons fait une grande découverte sur le monde extérieur et qu'on réunisse ici plein de savants, comme dans tes conférences, pour les convaincre que nous avons raison — bien sûr, tous les gens de la conférence (et toi aussi) auraient grandi dans cette grotte, autrement, c'est tricher. Mais il n'y aurait pas eu de problème. Tu as tellement de drôles de plantes qu'on aurait pu aussi *tous* les nourrir ! »

Cette fois, le père ne put s'empêcher de sourciller, mais ne dit toujours rien. Il resta pensif plusieurs minutes puis répondit : « Oui, je pense que ce serait possible. Mais vois-tu, le plus dur serait de les persuader que ce monde extérieur existe. Tout ce qu'ils connaîtraient du monde, ce seraient ces ombres, comment elles se déplacent et changent de temps en temps. Pour eux, ces ombres et ces choses qui s'agitent de manière compliquée sur la paroi de la grotte seraient tout ce qui formerait le monde. Alors, une partie de notre travail consisterait à les convaincre que le monde extérieur dont parle notre théorie existe réellement. En fait, ces deux choses iraient ensemble. Avoir une bonne théorie sur le monde extérieur nous permettrait bien plus facilement de les persuader que ce monde extérieur existe réellement ! »

« D'accord Papa, mais quelle serait notre théorie ? »

« Doucement ... laisse-moi une minute ... ah voilà : la Terre tourne autour du Soleil ! »

« Pas très nouveau comme théorie. »

« Non. En fait, elle est à peu près vieille de vingt-trois siècle — presque aussi vieille que le temps que ce rocher a passé coincé près de l'entrée ! Mais comme nous imaginons que nous avons passé toute notre vie dans cette grotte, les gens n'en auraient jamais entendu parler avant. Nous devrions alors les convaincre qu'il y a réellement une chose qui s'appelle le Soleil — et même une Terre, on en a besoin. L'idée est qu'en expliquant avec élégance toutes sortes de petits détails concernant les mouvements de la lumière et des ombres, nous arriverions à persuader la plupart des gens que non seulement il y a un point très brillant dehors — que nous appelons le "Soleil" — mais aussi que la Terre tourne continuellement autour de lui, tout en tournant autour de son axe. »

« Mais ça serait très dur de les persuader ? »

« Oh, certainement ! En fait, il nous faudrait faire deux choses tout à fait distinctes. Premièrement, nous devrions montrer comment notre simple théorie explique de manière très précise une quantité incroyable de détails liés

au déplacement de ce point brillant et de ses ombres sur la paroi de la grotte. Cela convaincrait peut-être certaines personnes, mais il y en aurait d'autres qui affirmeraient qu'il existe une théorie bien plus proche du bon sens, selon laquelle c'est le Soleil qui tourne autour de la Terre. Dans le détail, cette théorie serait cependant plus compliquée que la nôtre. Mais ces gens préféreraient se cramponner à leur théorie compliquée — et ils n'auraient pas tort — simplement parce qu'ils ne pourraient accepter l'idée que leur grotte se déplace dans l'espace à environ cent mille kilomètres à l'heure, comme le prétendrait notre théorie. »

« Quoi ? si vite que ça ? »

« À peu près, oui. Ensuite, pour la seconde partie de notre raisonnement, nous devrions changer complètement de tactique et faire des choses que les gens trouveraient totalement absurdes. Nous ferions rouler des boules le long de plans inclinés, osciller des pendules, et d'autres choses comme ça — juste pour montrer que les lois physiques qui gouvernent le mouvement des objets dans la grotte ne seraient pas affectées si la grotte entière se déplaçait dans n'importe quelle direction et à n'importe quelle vitesse. Cela leur montrerait qu'ils ne sentiraient absolument rien si la grotte se déplaçait à une vitesse énorme. C'est l'une des choses importantes que Galilée a été obligé de démontrer — tu te rappelles qui c'est ? Je t'ai donné un livre sur lui. »

« Oui, bien sûr ! Mais qu'est-ce que c'est compliqué ! Je parie que beaucoup de gens s'endormiraient, comme je les ai vu faire à des vraies conférences quand tu fais un exposé. »

Le père de Jessica rougit légèrement. « Je crois que tu as raison. Oui, mais j'ai bien peur que la science soit souvent comme ça : des tonnes et des tonnes de détails qui, pour la plupart, peuvent paraître très ennuyeux et parfois presque totalement sans aucun rapport avec l'idée que tu essaies de faire passer, même si cette idée peut finalement s'avérer extrêmement simple, comme le fait que la Terre tourne sur elle-même tout en tournant autour d'une chose appelée Soleil. Il y a certaines personnes qui trouvent qu'elles n'ont pas à se soucier de tous ces détails ennuyeux parce que l'idée leur paraît suffisamment plausible. Mais il y a aussi les vrais sceptiques, ceux qui veulent absolument tout vérifier pour s'assurer qu'il n'y a aucun point faible. »

« Merci Papa ! J'aime quand tu me parles de choses comme ça, quand tu deviens parfois tout rouge et tout excité. On peut rentrer maintenant ? Il commence à faire sombre, je suis fatiguée et j'ai faim — et un peu froid. »

« Bien, on y va » répondit le père en lui posant sa veste sur les épaules. Il rassembla ses objets puis enlaça sa fille pour la guider vers la sortie maintenant assombrie de la grotte. Alors qu'ils sortaient, Jessica regarda une nouvelle fois le rocher.

« Tu sais, je crois que je suis d'accord avec toi, Papa. Ce rocher va rester là pendant au moins vingt-trois autres siècles ! »

Première partie

Pourquoi il faut une nouvelle physique pour comprendre l'esprit

La non-calculabilité de la pensée consciente

1
Conscience et calcul

1.1 L'esprit et la science

Quel est le véritable champ d'action de la science ? Doit-elle se borner à comprendre les attributs *matériels* de notre Univers qui sont réductibles à ses méthodes et exclure à jamais de sa démarche notre existence *mentale* ? Ou parviendrons-nous un jour à une authentique compréhension scientifique du ténébreux mystère de l'esprit ? Le phénomène de la conscience humaine échappe-t-il à l'investigation scientifique, ou la puissance de la méthode scientifique résoudra-t-elle un jour le problème de l'existence même de notre moi conscient ?

Certains pensent que nous sommes aujourd'hui sur le point d'aboutir à une compréhension scientifique de la conscience, que ce phénomène ne recèle *aucune* zone d'ombre, voire que nous disposons déjà de tous les éléments pour l'interpréter. Ils affirment que la limitation actuelle de notre compréhension de l'esprit humain résulte simplement de la complexité et de la sophistication organisationnelle extrêmes de notre cerveau — complexité et sophistication qu'il ne faut certes pas sous-estimer, mais qui n'entraînent aucune objection de principe nous obligeant à sortir du cadre scientifique actuel. À l'autre extrême se trouvent ceux qui soutiennent que le problème de la pensée et de l'esprit — et le mystère même de la conscience humaine — ne pourra jamais être abordé de manière convenable à l'aide des froides procédures de calcul d'une science insensible.

Je tenterai dans ce livre d'aborder le problème de la conscience d'un point de vue scientifique. J'affirmerai cependant haut et fort — en *recourant* à un argument scientifique — qu'un élément essentiel manque à notre arsenal scientifique actuel. Cet élément est justement celui qu'il nous faudrait pour

traiter les problèmes cruciaux de l'esprit humain au sein d'une vision du monde scientifiquement cohérente. J'affirmerai que cet élément *ne se situe pas* au-delà de la science — bien qu'il ne fasse aucun doute que nous avons besoin, pour le formuler, d'une représentation scientifique du monde convenablement élargie. Dans la deuxième partie de ce livre, je tenterai de guider le lecteur dans une direction très particulière conduisant à une telle extension de notre représentation actuelle de l'univers physique. Cette direction implique un changement radical de nos lois physiques les plus fondamentales, changement que je décrirai de manière relativement précise et dont j'examinerai les éventuelles conséquences pour la biologie du cerveau humain. Bien que nous ne comprenions encore que partiellement la nature de cet élément manquant, nous pouvons d'ores et déjà entrevoir en quel endroit il devrait poser sa marque et comment il devrait fournir une contribution essentielle à l'étude de tout ce qui sous-tend nos actes et sentiments conscients.

Inévitablement, mes arguments seront parfois relativement complexes. Je me suis néanmoins efforcé de les exposer aussi clairement que possible en recourant au maximum à des concepts élémentaires. J'ai introduit par endroits quelques points techniques mathématiques, mais uniquement lorsqu'ils étaient nécessaires ou qu'ils amélioraient la clarté de l'exposé. Si, par expérience, je ne m'attends pas à une approbation unanime de mes idées, je crois cependant qu'elles méritent d'être considérées avec soin et sans passion : elles expriment en effet une thèse que l'on ne peut ignorer.

Une vision scientifique de l'Univers qui n'intègre pas le problème de l'esprit conscient ne peut sérieusement prétendre être une vision complète. La conscience faisant partie de notre Univers, toute théorie physique qui ne lui accorde pas une place convenable ne peut qu'échouer à donner une véritable description de cet Univers. J'affirme qu'il n'existe encore aucune théorie physique, biologique ou informatique qui puisse expliquer la conscience et l'intelligence qui en découle ; cela ne doit toutefois pas nous décourager d'en chercher une. Les arguments exposés dans ce livre obéissent à une telle aspiration. Peut-être disposerons-nous un jour de toutes les idées nécessaires pour atteindre un tel objectif. Notre horizon philosophique en sera alors profondément modifié. Pourtant, toute connaissance scientifique est une épée à double tranchant. Ce que nous *ferons* de cette connaissance est un autre problème. Essayons maintenant de voir où peut nous conduire notre vision de la science et de l'esprit.

1.2 Les robots sauveront-ils le monde ?

Chaque fois que nous ouvrons un journal ou allumons la télévision, nous sommes assaillis par les conséquences de la stupidité humaine. Des pays, des régions, s'opposent dans des confrontations qui, de temps en temps, dégénèrent en guerres atroces. La ferveur religieuse excessive, le nationalisme, les

intérêts ethniques divergents, de simples différences culturelles ou linguistiques, les ambitions égoïstes de quelques démagogues engendrent une agitation et une violence continuelles, débouchant parfois sur des atrocités indescriptibles. Des régimes autoritaires et oppressifs soumettent encore des peuples, les maintenant sous leur coupe en recourant à la torture ou à des escadrons de la mort. De leur côté les opprimés, qui sembleraient pourtant avoir un objectif commun, s'opposent souvent entre eux et, lorsqu'ils acquièrent la liberté qui leur fut longtemps refusée, choisissent parfois de l'utiliser à des fins terriblement autodestructrices. Même dans les pays privilégiés où règnent paix, prospérité et libertés démocratiques, les ressources et les forces de travail sont dilapidées de manière apparemment insensée. N'est-ce pas là un signe manifeste de la stupidité humaine ? Bien que nous nous considérions comme le summum de l'intelligence au sein du règne animal, cette intelligence semble tristement incapable de traiter nombre des problèmes auxquels notre société continue de nous confronter.

Pourtant, on ne peut nier les succès de notre intelligence. Voyez par exemple la science et la technologie. De fait, même si nous devons reconnaître que certains des fruits de cette technologie correspondent à des valeurs clairement contestables à long terme (ou à court terme) — ainsi qu'en témoignent de nombreux problèmes environnementaux et l'inquiétude véritable que suscite l'idée d'une catastrophe planétaire d'origine technologique —, c'est cette même technologie qui a produit notre société moderne et son confort, qui nous a libérés de la peur, de la maladie et du besoin, qui a favorisé l'énorme développement de nos capacités intellectuelles et de notre sens esthétique et qui nous permet de communiquer avec le monde entier pour le plus grand enrichissement de notre esprit. Si cette technologie nous a ouvert tant d'horizons et, en un sens, a étendu le champ et le pouvoir de notre moi physique individuel, ne pouvons-nous pas imaginer qu'elle fera bien davantage dans l'avenir ?

Notre technologie, tant ancienne que moderne, a immensément développé nos capacités sensorielles. Notre vue est décuplée grâce aux lunettes, miroirs, télescopes et microscopes de toutes sortes, grâce aux caméras vidéo, à la télévision, etc. Notre ouïe, qui ne s'aidait autrefois que de simples cornets acoustiques, bénéficie aujourd'hui de minuscules dispositifs électroniques et se trouve relayée par le téléphone, les communications radio et satellitaires. Nous avons des bicyclettes, des trains, des automobiles, des bateaux, des avions pour accroître et surpasser nos moyens naturels de locomotion. Nos souvenirs ne s'estompent plus grâce aux livres, aux films — et aux énormes capacités de stockage des ordinateurs. Nos calculs, qu'ils soient simples et routiniers ou gigantesques et sophistiqués, sont également amplement facilités par la puissance des ordinateurs modernes. Ainsi, outre qu'elle accroît considérablement le champ de notre moi *physique*, notre technologie élargit nos capacités *mentales* en améliorant grandement nos aptitudes à effectuer de nombreuses tâches routinières. Qu'en est-il des tâches mentales qui ne sont pas routinières — de celles qui exigent une véritable *intelligence* ? Seront-elles, elles aussi, facilitées par la technologie informatique ?

Il est selon moi évident que notre société technologique (souvent informatisée) contient au moins un domaine pouvant énormement renforcer notre intelligence. Je pense ici à nos méthodes éducatives. Celles-ci pourraient tirer un grand parti de différents aspects de la technologie — à condition que cette dernière soit utilisée avec discernement — grâce à une production judicieuse de livres, de films, d'émissions télévisées, et au recours à divers types de systèmes interactifs gérés par ordinateur. Ces moyens — et bien d'autres — représentent autant d'occasions d'éveiller notre esprit — ou de l'endormir. L'esprit humain peut accomplir bien plus qu'il ne lui est souvent donné de réaliser. Malheureusement, ces occasions sont trop fréquemment perdues, et notre esprit, que nous soyons jeune ou vieux, ne se voit pas offrir les ouvertures qu'indéniablement il mérite.

Nombre de lecteurs ne manqueront pas de faire remarquer qu'il existe aussi une autre piste, à savoir cette étrange « intelligence » électronique qui ne fait encore qu'émerger des extraordinaires progrès effectués par la technologie informatique. C'est vrai. Nous avons d'ailleurs déjà recours à l'assistance des ordinateurs pour accomplir certaines tâches intellectuelles. Par exemple, placée devant un choix, l'intelligence humaine échoue souvent à en évaluer les diverses issues possibles, leur détermination pouvant dépasser considérablement sa puissance de calcul. Il est donc à prévoir que dans les situations où la prise de décision dépend de la valeur brute d'un résultat numérique, les ordinateurs de l'avenir accroîtront considérablement les capacités de l'intelligence humaine.

Se pourrait-il toutefois que les ordinateurs accomplissent un jour bien plus ? Nombre d'experts soutiennent qu'en effet ces machines recèlent, du moins en principe, la possibilité de créer une intelligence *artificielle* qui finira par être supérieure à notre propre intelligence[1]. Lorsque les robots commandés par ordinateur auront atteint le niveau de l'« équivalent humain », nous n'attendrons pas longtemps — affirment ces spécialistes — avant qu'ils ne dépassent largement ce misérable niveau. *Alors* seulement, prétendent-ils, nous disposerons d'une entité dotée d'une intelligence, d'une sagesse et d'une compréhension suffisantes pour résoudre les problèmes engendrés dans ce monde par l'humanité.

Combien de temps s'écoulera avant que ne survienne cette situation bénie ? Les experts ne sont pas véritablement unanimes sur ce point. Certains parlent de plusieurs siècles, d'autres prétendent que cet équivalent humain est une affaire de décennies[2]. Ces derniers appuient leurs arguments sur la croissance « exponentielle » qu'a connue la puissance des ordinateurs et sur des comparaisons entre la vitesse et la précision des transistors d'une part, et la relative nonchalance de l'activité neuronale de l'autre. De fait, certains circuits électroniques sont déjà plus d'un million de fois plus rapides que les neurones du cerveau (ces vitesses étant de quelque 10^9 signaux/s pour les transistors et seulement 10^3/s pour les neurones*) et possèdent, tant au niveau de la coordina-

---

* La puce Pentium d'Intel contient, logés dans une « tranche de silicium » qui tiendrait sur l'ongle du pouce, plus de trois millions de transistors, chacun d'eux étant capable d'exécuter 113 millions d'instructions par seconde.

tion des tâches qu'à celui de leur exécution, une haute précision totalement absente de nos neurones. En outre, le « câblage » du cerveau est passablement aléatoire et pourrait apparemment être largement dépassé par une organisation précise et judicieuse de circuits imprimés.

Si la structure neuronale du cerveau fournit cependant encore, dans certains domaines, un avantage numérique sur les ordinateurs actuels, cela ne durera probablement pas. On considère qu'avec ses quelques centaines de milliers de millions de neurones, le cerveau humain est actuellement supérieur aux ordinateurs, du point de vue du nombre de transistors qu'ils renferment. D'autre part, il y a en moyenne bien plus de *connexions* entre les différents neurones qu'il n'en existe entre les transistors d'un ordinateur. Dans le cervelet en particulier, les cellules de Purkinje peuvent avoir jusqu'à 8 000 terminaisons synaptiques (les jonctions entre neurones), tandis que dans un ordinateur, le nombre correspondant s'élève au plus à trois ou quatre. (Je ferai plus loin quelques commentaires sur le cervelet ; *cf.* §1.14, §8.6.) De surcroît, la plupart des transistors des ordinateurs actuels ont simplement pour fonction de mémoriser, sans intervenir directement dans les calculs, tandis que le calcul serait une activité plus répandue dans le cerveau.

Cette supériorité actuelle du cerveau pourrait être remise en question dans l'avenir, notamment lorsqu'on aura davantage développé les systèmes informatiques massivement « parallèles ». L'un des atouts de l'ordinateur est que l'on peut associer des unités différentes pour former des entités de plus en plus grosses permettant, en principe, d'accroître pratiquement sans limite le nombre des transistors. De plus, certaines révolutions technologiques pointent le bout de leur nez — telles le remplacement des câbles et des transistors de nos ordinateurs actuels par des dispositifs optiques (des lasers) —, promettant d'énormes progrès en termes de vitesse, de puissance et de miniaturisation. À un niveau plus fondamental, nous subissons de nombreuses contraintes, telles celle de devoir croître à partir d'une seule cellule ou d'avoir un cerveau apparemment figé sur ses caractéristiques actuelles. Les ordinateurs, en revanche, peuvent être délibérément construits de manière à satisfaire tout besoin éventuel. Bien que ces considérations ne prennent pas en compte certains facteurs importants que j'indiquerai plus loin (tout particulièrement l'existence d'une activité significative sous-jacente à celle des neurones), il n'en reste pas moins que sur le seul problème de la puissance de calcul, nombre d'arguments démontrent de manière convaincante que si les ordinateurs ne sont pas déjà supérieurs au cerveau, ils le *seront* certainement avant longtemps.

Ainsi, à en croire les adeptes les plus déterminés de l'intelligence artificielle, les capacités des ordinateurs et des robots informatisés dépasseront un jour — voire bientôt — celles de l'être humain, et ces ordinateurs pourront faire immensément plus que simplement assister *notre* intelligence. Ils disposeront réellement d'une intelligence propre, et celle-ci sera incommensurable. Nous pourrons *alors* nous tourner vers ces intelligences supérieures pour leur demander conseil ou invoquer leur autorité à propos de tout — et les problèmes introduits par l'humanité sur cette planète pourront enfin être résolus !

Ces développements potentiels semblent toutefois entraîner une autre conséquence logique, conséquence qui pourrait se révéler inquiétante. Ces ordinateurs ne finiront-ils pas un jour par rendre les êtres humains eux-mêmes superflus ? Si les robots informatisés nous surpassent à tous égards, n'estimeront-ils pas qu'ils peuvent mieux faire marcher le monde en se dispensant complètement de nous ? L'humanité elle-même deviendra alors obsolète. Peut-être, si nous avons de la chance, nous garderont-ils comme animaux de compagnie, comme l'a suggéré Edward Fredkin ; ou peut-être, si nous sommes suffisamment intelligents, pourrons-nous transformer en robots les « schémas d'information » *que nous sommes*, ainsi que le croit Hans Moravec (1988) ; ou peut-être ne serons-nous pas aussi chanceux, ou simplement pas assez intelligents...

1.3 L'$\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ du calcul et de la pensée consciente

Mais les véritables problèmes sont-ils simplement ceux de la puissance de calcul, de la vitesse, de la précision, de la mémoire, voire du détail du « câblage » entre les divers éléments ? Se peut-il que notre cerveau ait en fait des activités indescriptibles en termes de calcul ? Comment la connaissance immédiate consciente — nos sentiments de bonheur, de douleur, d'amour, de sensibilité esthétique, de volonté, de compréhension, etc. — s'intègre-t-elle dans une telle représentation numérisée ? Les ordinateurs de l'avenir auront-ils réellement un *esprit* ? La présence d'un esprit conscient influe-t-elle réellement en quoi que ce soit sur le comportement ? Est-il sensé de parler de ces choses en termes scientifiques, ou la science est-elle totalement incompétente pour aborder les problèmes liés à la conscience humaine ?

Il me semble que l'on peut raisonnablement avancer au moins quatre points de vue différents[3] — quatre extrêmes — sur ce sujet :

$\mathcal{A}$. Toute pensée se réduit à un calcul ; en particulier, le sentiment de connaissance immédiate consciente naît simplement de l'exécution de calculs appropriés.

$\mathcal{B}$. La connaissance immédiate est un produit de l'activité physique du cerveau ; mais bien que toute action physique puisse être simulée par un calcul, une telle simulation ne peut par elle-même susciter la connaissance immédiate.

$\mathcal{C}$. La connaissance immédiate est suscitée par une action physique du cerveau, mais aucun calcul ne peut simuler, même à la perfection, cette action physique.

$\mathcal{D}$. On ne peut expliquer la connaissance immédiate à l'aide du langage de la physique, de l'informatique, ni de quelque autre discipline scientifique que ce soit.

Le point de vue $\mathscr{D}$, qui nie résolument la position physicaliste et considère l'esprit comme une entité totalement inexplicable en termes scientifiques, est celui du mystique ; l'acceptation d'une doctrine religieuse semble d'ailleurs mettre en jeu au moins quelques éléments de $\mathscr{D}$. Bien qu'il occupe une situation très inconfortable dans la connaissance scientifique actuelle, le problème de l'esprit ne doit pas, selon moi, être considéré comme définitivement hors de portée de la science. Si celle-ci est encore incapable de dire grand-chose de sensé sur ce problème, elle finira probablement par élargir son champ d'action de manière à le prendre en compte — au besoin en modifiant ses propres procédures. Si je rejette le mysticisme pour sa négation des critères scientifiques qui permettent le progrès de la connaissance, je suis persuadé qu'en élargissant son champ, la science éclaircira quantité de mystères, dont celui de l'esprit. Je reviendrai plus loin sur ces idées ; il me suffit pour l'instant de dire que je rejette $\mathscr{D}$ et de tenter de progresser en suivant la voie que la science a tracée pour nous. Si vous êtes fermement convaincu de la vérité de $\mathscr{D}$, sous une forme ou une autre, je vous demande un peu de patience. Voyez jusqu'où on peut aller en suivant la voie scientifique — et essayez de percevoir à quel endroit, selon moi, cette voie devrait finalement nous mener.

Considérons maintenant ce qui me semble être l'autre extrême, à savoir le point de vue $\mathscr{A}$. Ceux qui adoptent la thèse souvent qualifiée d'*IA forte* (intelligence artificielle forte) ou parfois d'*IA dure*, ou encore de *fonctionnalisme*[4], se rangent sous cette bannière — bien que le mot « fonctionnalisme » prenne parfois un sens englobant certaines versions de $\mathscr{C}$. Le point de vue $\mathscr{A}$ est considéré par certains comme le seul digne d'une véritable attitude scientifique. D'autres le tiennent pour une absurdité méritant à peine qu'on s'y arrête. $\mathscr{A}$ admet de nombreuses formulations (on en trouvera une longue liste dans Sloman 1992). Certaines diffèrent par leur définition des mots « calcul » ou « exécution » d'un calcul. Il existe d'ailleurs des partisans de $\mathscr{A}$ qui nient catégoriquement être des « partisans de l'IA forte » dans la mesure où ils ont du mot « calcul » une interprétation différente de celle admise par l'IA conventionnelle (*cf.* Edelman 1992). J'examinerai plus en détail ces problèmes à la section 1.4. Il nous suffira pour l'instant de désigner par « calcul » toute opération susceptible d'être effectuée par un ordinateur standard. D'autres partisans de $\mathscr{A}$ s'opposent sur le sens des termes « connaissance immédiate » ou « conscience ». Certains nient même l'*existence* d'un phénomène tel que la connaissance immédiate consciente, tandis que d'autres l'acceptent en la considérant toutefois comme une simple « propriété émergente » (*cf.* aussi §4.3 et §4.4) se manifestant dès lors qu'un calcul met en jeu un niveau de complexité (ou de sophistication, ou d'auto-référence, etc.) suffisant. Je donnerai ma propre interprétation des termes « conscience » et « connaissance immédiate » à la section 1.12. D'ici là, ces différences d'interprétation ne seront pas essentielles pour les considérations que je vais développer.

Le point de vue de l'IA forte — *i.e.* le point de vue $\mathscr{A}$ — représente ce contre quoi portaient plus particulièrement les arguments que j'ai exposés dans *l'Esprit, l'ordinateur et les lois de la physique* (EOLP). La seule épaisseur de ce livre montre que, bien que je ne croie pas personnellement à la justesse de

$\mathcal{A}$, je le tiens *réellement* pour une possibilité sérieuse méritant une attention considérable. $\mathcal{A}$ résulte d'une attitude scientifique hautement opérationaliste qui va jusqu'à considérer que le monde physique fonctionne entièrement comme un ordinateur. Selon une version extrême de ce point de vue, l'Univers lui-même serait en fait un gigantesque ordinateur[5], et les sous-calculs effectués par cet ordinateur susciteraient les sentiments de « connaissance immédiate » qui constituent notre esprit conscient.

J'imagine que ce point de vue — qui assimile les systèmes physiques à de simples entités informatiques — a son origine en partie dans le rôle essentiel que jouent les simulations numériques dans la science du XX^{e} siècle, en partie aussi dans la conviction que les objets physiques eux-mêmes sont, en un sens, de simples « schémas d'information » soumis à des lois mathématiques. La plupart des matériaux constituant notre corps et notre cerveau étant en permanence renouvelés, c'est uniquement leur *schéma d'information* qui reste inchangé. En outre, la matière elle-même peut se transformer et semble n'avoir qu'une existence éphémère. Même la *masse* d'un corps matériel, qui donne une mesure physique précise de la quantité de matière contenue dans ce corps, peut, selon les circonstances, se convertir entièrement en énergie (selon la fameuse équation d'Einstein $E = mc^2$) — et donc devenir une pure abstraction mathématique. De surcroît, la théorie quantique semble nous dire que les particules matérielles sont de simples « ondes » d'information. (Nous examinerons plus en profondeur ces problèmes dans la deuxième partie.) Ainsi, si la matière elle-même a un caractère insaisissable et éphémère, il n'est absolument pas déraisonnable de penser que la persistance du « moi » puisse être davantage liée à la préservation de *schémas* qu'à la conservation de particules matérielles concrètes.

Même si l'on ne considère pas l'Univers comme un simple ordinateur, on peut néanmoins, par inclination opérationaliste, être séduit par le point de vue $\mathcal{A}$. Supposons que nous ayons un robot informatisé répondant aux questions exactement comme le ferait un être humain. Quand nous lui demandons comment il se sent, nous constatons qu'il répond exactement comme s'il éprouvait effectivement des sentiments. Il nous dit qu'il est conscient, qu'il est gai ou triste, qu'il perçoit la couleur rouge, et que des questions comme l'« esprit » et le « moi » le préoccupent. Il dit même douter de l'existence, chez d'*autres* êtres (en particulier chez les humains), d'une conscience analogue à celle qu'il affirme lui-même posséder. Pourquoi ne pas le croire lorsqu'*il* prétend avoir l'expérience de la connaissance immédiate, de l'émerveillement, de la joie, de la douleur, alors que parfois nous-mêmes avons si peu en commun avec des êtres humains auxquels cependant nous *reconnaissons* une conscience ? L'argument opérationaliste possède, me semble-t-il, une force considérable, même s'il n'est pas totalement convaincant. Si un système entièrement géré par ordinateur peut de fait parfaitement imiter toutes les manifestations *externes* d'un cerveau conscient — notamment les réponses à un interrogatoire continu —, pourquoi alors ne pas accepter que ses manifestations *internes* — la conscience elle-même — soient également présentes dans une telle simulation ?

L'acceptation de ce type d'argument, qui est fondamentalement ce que l'on appelle un *test de Turing*[6], est en substance ce qui distingue $\mathcal{A}$ de $\mathcal{B}$. Selon $\mathcal{A}$, tout robot informatisé qui, lors d'un interrogatoire soutenu, se comporte de manière convaincante *comme s'il* possédait une conscience, doit être considéré comme *réellement* conscient — tandis que selon $\mathcal{B}$, un robot pourrait fort bien se comporter exactement comme une personne douée de conscience sans en réalité posséder lui-même la moindre parcelle de cet attribut mental. Si $\mathcal{A}$ comme $\mathcal{B}$ reconnaissent qu'un robot informatisé peut *se comporter* de manière convaincante comme une personne consciente, le point de vue $\mathcal{C}$ en revanche n'admet même pas qu'il puisse parvenir ne serait-ce qu'à une parfaite simulation de ce comportement. Ainsi, selon $\mathcal{C}$, un interrogatoire suffisamment long permettrait de révéler l'absence de conscience chez ce robot. En fait, $\mathcal{C}$ correspond davantage que $\mathcal{B}$ à une conception *opérationaliste* — et est, à cet égard, plus proche de $\mathcal{A}$ que de $\mathcal{B}$.

Qu'en est-il donc de $\mathcal{B}$? Je pense que pour la majorité des gens, ce point de vue est celui du « bon sens scientifique ». On le désigne parfois sous le nom d'*IA faible* (ou *douce*). Comme $\mathcal{A}$, il affirme que tous les objets physiques de ce monde se comportent selon une science qui, en principe, permet leur simulation numérique. Il nie fermement, en revanche, le point de vue opérationaliste, à savoir qu'une chose affichant le comportement d'un être conscient est elle-même nécessairement consciente. Comme l'a souligné le philosophe John Searle[7], la simulation numérique d'un processus physique est très différente du processus lui-même — la simulation d'un ouragan n'est pas un ouragan ! D'après $\mathcal{B}$, la présence ou l'absence de conscience dépend fortement de l'objet physique « produisant la pensée » et des actions physiques particulières que cet objet effectue. La considération des calculs mis en jeu dans ces actions est secondaire. Ainsi, si l'activité d'un cerveau biologique peut susciter la conscience, une simulation électronique précise de cette activité ne le peut pas. Selon $\mathcal{B}$, cette impossibilité ne se fonde pas nécessairement sur la distinction entre biologie et physique. Ce qui importe avant tout, c'est la constitution *matérielle* réelle de l'objet en question (disons, un cerveau), et non les opérations qu'il effectue.

$\mathcal{C}$ m'apparaît plus proche de la vérité. Il est plus opérationaliste que $\mathcal{B}$ dans la mesure où il affirme l'existence de manifestations externes d'objets conscients (par exemple du cerveau) qui diffèrent des manifestations externes d'un ordinateur : il dit qu'un ordinateur ne peut simuler correctement les manifestations externes de la conscience. Je donnerai en temps utile les raisons qui me poussent à préférer ce point de vue. Reconnaissant, comme $\mathcal{B}$, la position physicaliste selon laquelle l'esprit est un produit du comportement de certains objets physiques (le cerveau — mais pas nécessairement seulement le cerveau), $\mathcal{C}$ affirme qu'il existe des actions physiques qu'un ordinateur ne peut correctement simuler.

La physique actuelle autorise-t-elle l'existence d'une action impossible à simuler sur ordinateur ? D'un point de vue rigoureusement mathématique, la réponse à cette question ne m'apparaît pas complètement claire. On ne dispose pas encore sur ce point de théorèmes mathématiques précis[8]. Je suis

toutefois fermement convaincu qu'une telle action non numérique est à rechercher dans un domaine *qui échappe* aux lois physiques actuellement connues. J'y reviendrai ; de solides arguments, issus de la physique elle-même, donnent à penser qu'il existe un domaine physique encore inexploré situé entre le niveau « à petite échelle » — régi par les lois quantiques — et le niveau « ordinaire » de la physique classique. Toutefois, l'existence de ce domaine, et donc la nécessité d'une nouvelle théorie physique qui en rendrait compte, sont loin d'être universellement acceptées par les physiciens d'aujourd'hui.

Ainsi, $\mathcal{C}$ recouvre deux points de vue très différents. Certains partisans de $\mathcal{C}$ soutiennent que le cadre physique actuel est parfaitement adéquat et suffit pour rechercher des comportements subtils échappant totalement à toute simulation numérique (*e.g.*, ainsi que nous l'examinerons plus loin : le comportement chaotique (§1.7), les subtilités de l'opposition entre processus continus/processus discrets (§1.8) et le hasard quantique). D'autres en revanche affirment que la physique d'aujourd'hui ne nous offre aucune perspective raisonnable de découvrir la non-calculabilité recherchée. Je donnerai plus loin ce que je pense être d'excellentes raisons pour adopter cette version forte de $\mathcal{C}$, à savoir celle qui exige l'introduction d'une physique fondamentalement nouvelle.

Certaines personnes pensent que puisque j'affirme que si jamais nous trouvons une explication au phénomène de la conscience, elle se situera hors du domaine connu de la science, je me place en fait dans le camp $\mathcal{D}$. Il existe cependant une différence essentielle entre la version forte de $\mathcal{C}$ et le point de vue $\mathcal{D}$ — en particulier au regard de la *méthodologie*. Selon $\mathcal{C}$, le problème de la connaissance consciente est un problème scientifique, même si nous ne disposons pas encore du cadre théorique approprié. Je partage résolument ce point de vue. Je suis convaincu que ce sont les méthodes de la science — convenablement élargies en un sens que nous ne pouvons qu'à peine entrevoir aujourd'hui — qui nous permettront de résoudre le problème de la conscience. C'est là que réside, par-delà les similitudes de jugement que ces deux points de vue peuvent avoir sur ce dont est capable la science *actuelle*, la différence capitale entre $\mathcal{C}$ et $\mathcal{D}$.

Je l'ai dit, tels qu'ils sont formulés plus haut, les énoncés $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ et $\mathcal{D}$ représentent les formes extrêmes — les polarités — des diverses positions que l'on peut choisir d'adopter. J'admets cependant que certains puissent considérer qu'aucune de ces quatre positions ne traduit fidèlement leur point de vue, et que celui-ci se situe quelque part entre elles. Il existe sûrement de nombreuses gradations entre $\mathcal{A}$ et $\mathcal{B}$ (voir Sloman 1992). Il existe même un point de vue, qui est loin d'être exceptionnel, correspondant à une combinaison de $\mathcal{A}$ et $\mathcal{D}$ (ou peut-être $\mathcal{B}$ et $\mathcal{D}$) — il jouera d'ailleurs un rôle marquant dans nos réflexions ultérieures. Selon ce point de vue, l'activité du cerveau est identique à celle d'un ordinateur, mais d'un ordinateur d'une complexité si extraordinaire que son imitation surpasse l'entendement de l'homme et de la science, d'un ordinateur, en somme, qui est nécessairement une création de Dieu — le « meilleur programmeur de la profession »[9] !

1.4 Physicalisme et mentalisme

Je dois faire ici quelques brèves remarques sur l'emploi des mots « physicaliste » et « mentaliste » auxquels on recourt souvent pour décrire des divergences entre les thèses résumées par $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ et $\mathcal{D}$. $\mathcal{D}$ représentant une négation totale du physicalisme, il ne fait aucun doute que l'on doive ranger ses partisans parmi les mentalistes. Il en va autrement avec les trois autres points de vue, pour lesquels la frontière entre physicalisme et mentalisme ne m'apparaît pas nettement définie. Je pense que les partisans de $\mathcal{A}$ doivent être normalement considérés comme des physicalistes, et je suis sûr que la grande majorité d'entre eux est d'accord avec moi. Il y a là cependant quelque chose qui tient du paradoxe. Selon $\mathcal{A}$, la structure *matérielle* d'un dispositif pensant n'influe en rien sur ses attributs mentaux. Ceux-ci sont uniquement et totalement déterminés par les calculs effectués. Ces calculs sont eux-mêmes un jeu mathématique abstrait qui n'est en rien lié à des corps matériels particuliers. Ainsi, les attributs mentaux n'ayant, selon $\mathcal{A}$, aucun lien avec des objets physiques, le qualificatif de « physicaliste » peut paraître légèrement inadéquat. Les points de vue $\mathcal{B}$ et $\mathcal{C}$ en revanche affirment que la composition physique d'un objet joue un rôle essentiel dans la détermination de la présence ou de l'absence d'une conscience véritable dans cet objet. Il s'ensuit que l'on peut considérer que $\mathcal{B}$ et $\mathcal{C}$ représentent, plus que $\mathcal{A}$, les positions physicalistes possibles. Toutefois, il semble qu'une telle terminologie diffère d'un certain usage commun, le mot « mentaliste » étant souvent jugé plus approprié pour $\mathcal{B}$ et $\mathcal{C}$, puisqu'ici, les qualités mentales sont perçues comme des « choses réelles » et non comme de simples « épiphénomènes » survenant incidemment lors de l'exécution de calculs (d'un certain type). Face à de telles confusions, je m'efforcerai autant que possible de ne pas recourir aux mots « physicaliste » et « mentaliste » dans la discussion qui va suivre et me référerai explicitement aux points de vue $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ et $\mathcal{D}$ tels qu'ils sont définis plus haut.

1.5 Le calcul : procédure descendante et procédure ascendante

Je n'ai jusqu'ici rien dit de ce que j'entends par « calcul » dans les définitions de $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ et $\mathcal{D}$ à la section 1.3. Qu'est-ce qu'un calcul ? D'une manière générale, on peut désigner par ce terme l'activité d'un ordinateur standard. Mais pour être plus précis, nous allons lui donner un sens convenablement idéalisé : un *calcul* est l'action d'une *machine de Turing*.

Qu'est-ce qu'une machine de Turing ? C'est en vérité un ordinateur mathématiquement idéalisé (le précurseur théorique de l'ordinateur moderne) — idéalisé en ce sens qu'il ne commet jamais d'erreur, peut fonctionner aussi

longtemps que nécessaire et possède une capacité de stockage illimitée. J'aborderai un peu plus précisément la définition d'une machine de Turing à la section 2.1 et dans l'appendice A. (Pour une introduction plus détaillée, le lecteur intéressé pourra par exemple consulter le chapitre 2 de EOLP, ou Kleene 1952, ou encore Davis 1978.)

L'action d'une machine de Turing est souvent appelée « algorithme ». Ce terme est pour moi totalement synonyme de « calcul ». Cela demande un petit éclaircissement dans la mesure où certaines personnes adoptent un point de vue plus restrictif et appellent « algorithme » ce que je désigne sous l'expression « algorithme descendant ». Voyons donc ce que recouvrent le qualificatif « descendant » et son antithèse « ascendant » dans le contexte d'un calcul.

Un algorithme est dit *descendant* s'il a été élaboré selon une procédure de calcul fixe, bien définie et parfaitement maîtrisée (pouvant inclure un ensemble de connaissances pré-installé), et si cette procédure est expressément destinée à fournir une solution précise à un problème donné. (L'algorithme d'Euclide décrit p. 34 dans EOLP et permettant de trouver le plus grand commun diviseur de deux entiers naturels est un exemple simple d'algorithme descendant.) Cette organisation descendante s'oppose à l'organisation dite *ascendante*, qui ne spécifie à l'avance aucune règle clairement définie ni ne met en jeu aucune connaissance préalable, mais est conçue de telle sorte que le système « apprend » et améliore ses performances en fonction de son « expérience ». Ainsi, dans un algorithme ascendant, les règles opératoires sont sujettes à de constantes modifications. On exécute de nombreuses fois l'algorithme en le faisant à chaque fois opérer sur d'autres données de départ. Chaque exécution donne lieu à une évaluation — éventuellement effectuée par l'algorithme lui-même — puis, en fonction de cette évaluation, l'algorithme modifie ses opérations afin d'améliorer la qualité du résultat. Par exemple, les données de départ du système peuvent être un certain nombre de photographies de visages humains convenablement numérisées, et le système a pour tâche de trouver quelles photographies représentent le même individu. À chaque exécution de l'algorithme, on compare la performance du système aux réponses correctes, puis on modifie les règles opératoires en vue d'une amélioration des résultats lors de l'exécution suivante.

Les détails de la procédure d'amélioration ne nous intéressent pas ici. Il y a de nombreuses possibilités. Parmi les plus connus des systèmes ascendants figurent ce que l'on appelle les *réseaux de neurones formels* (parfois simplement appelés, de manière légèrement trompeuse, des « réseaux neuronaux »), des programmes informatiques capables d'apprendre — ou encore des dispositifs électroniques spécifiquement construits — reposant sur ce que l'on sait de la façon dont un système de connexions de neurones du cerveau améliore son organisation à mesure qu'il acquiert de l'expérience. (La façon dont un système de connexions de neurones du cerveau se modifie *réellement* va revêtir pour nous une certaine importance ; *cf.* §7.4 et §7.7.) Bien entendu, un algorithme peut également associer des éléments descendants et ascendants.

L'important pour nous ici est que ces procédures, descendantes comme ascendantes, peuvent être chargées sur un ordinateur standard et doivent donc

être rangées à la rubrique que je qualifierai d'*algorithmique*. Ainsi, la *façon* dont les systèmes ascendants (voire mixtes) modifient leurs procédures est elle-même fournie par une procédure prédéfinie entièrement *basée sur le calcul*. Cela explique justement que ces systèmes puissent être chargés sur un ordinateur standard. La différence essentielle entre un système ascendant (ou mixte) et un système descendant réside dans le fait que la procédure de calcul des systèmes ascendants garde le « souvenir » de ses résultats antérieurs — elle acquiert de l'« expérience » —, de sorte que ce souvenir peut être intégré dans les opérations de calcul ultérieures. Nous reviendrons plus en détail sur ce point au paragraphe 3.11.

L'intelligence artificielle s'efforce d'imiter un comportement intelligent, quel qu'en soit le niveau, en recourant à des procédures de calcul. Les organisations descendante et ascendante ont souvent été utilisées. Si, au début, les systèmes descendants parurent particulièrement prometteurs[10], les systèmes ascendants de type réseau de neurones formels sont ensuite devenus très populaires. Il semble que les systèmes d'IA les plus efficaces soient ceux qui *associent* les procédures descendante et ascendante. Chacune de ces procédures a ses avantages. Les succès de la procédure descendante tendent à se restreindre aux domaines où données et règles opératoires sont clairement définies et possèdent une formulation bien précise : problèmes mathématiques, jeu d'échecs ou diagnostics médicaux dans lesquels on donne un ensemble de règles permettant d'identifier différentes maladies, règles qui reposent sur des procédures médicales reconnues. L'organisation ascendante, quant à elle, est plutôt utile lorsque les critères de décision sont peu précis ou insuffisamment compris, par exemple dans la reconnaissance de visages ou de sons, voire dans la prospection minière, où ce que l'on recherche est l'amélioration des performances en fonction de l'expérience. Dans de nombreux cas interviennent en fait des éléments appartenant aux *deux* types de procédure, descendante et ascendante (par exemple lorsqu'un ordinateur jouant aux échecs se perfectionne en se fondant sur son expérience, ou lorsqu'on injecte des résultats de géologie théorique dans un dispositif de calcul servant à la prospection minière).

Pour être honnête, il faut dire que c'est uniquement dans certains cas de procédures descendantes (ou essentiellement descendantes) que les ordinateurs ont démontré une supériorité significative sur les êtres humains. L'exemple le plus évident est le calcul numérique pur, où ils gagnent haut la main — et aussi les jeux « mathématiques », tels les échecs ou les dames, où il semble que très peu de joueurs humains parviennent à battre les meilleures machines (je reviendrai sur ce point aux sections 1.15 et 8.2). Avec la procédure ascendante (le réseau de neurones formels), l'ordinateur atteint seulement, et dans un nombre limité de cas, un niveau équivalent à celui d'êtres humains ordinaires bien entraînés.

Une autre distinction entre les différents types de systèmes informatiques est celle qui oppose architecture *séquentielle* et architecture *parallèle*. Un ordinateur séquentiel est une machine qui effectue ses calculs l'un après l'autre, pas à pas, tandis qu'un ordinateur parallèle accomplit simultanément de

nombreux calculs indépendants, dont les résultats sont ensuite regroupés une fois qu'un nombre suffisant d'entre eux a été obtenu. Ce sont aussi les théories sur le fonctionnement du cerveau qui ont contribué au développement de certains systèmes parallèles. Il faut cependant souligner qu'il n'y a réellement aucune distinction *de principe* entre machine séquentielle et machine parallèle. Il est toujours possible de simuler séquentiellement des opérations parallèles, même s'il existe quelques types de problèmes (mais pas tous, loin de là) pour lesquels la résolution en parallèle s'avère plus efficace — en termes de temps de calcul, etc. — qu'une résolution séquentielle. Comme je vais surtout me concentrer sur des questions de principe, la distinction entre calculs séquentiels et calculs parallèles ne sera pas essentielle.

1.6 Le point de vue $\mathcal{C}$ viole-t-il la thèse de Church-Turing ?

Rappelons que selon le point de vue $\mathcal{C}$, l'activité du cerveau conscient échappe à toute simulation algorithmique, qu'elle soit descendante, ascendante, ou mixte. Certains critiques de $\mathcal{C}$ affirment notamment que ce point de vue contredit un argument (généralement admis) appelé *thèse de Church* (ou thèse de Church-Turing). Que dit la thèse de Church ? Telle qu'elle fut initialement formulée en 1936 par le logicien américain Alonzo Church, elle affirme que tout processus mathématique que l'on peut raisonnablement qualifier de « purement mécanique » — *i.e.* tout ce qui est *algorithmique* — peut s'obtenir à l'aide d'une procédure particulière découverte par Church lui-même et appelée *lambda-calcul* (λ-calcul)[11] (une procédure d'une élégance et d'une économie conceptuelle extrêmes ; voir EOLP, p. 72-78, pour une brève introduction à ce sujet). Peu après, en 1936-1937, le mathématicien anglais Alan Turing découvrit une description bien plus convaincante des processus algorithmiques, reposant sur le fonctionnement de « machines à calculer » théoriques que nous connaissons aujourd'hui sous le nom de *machines de Turing*. Le logicien américain d'origine polonaise Emil Post conçut également (1936) une procédure quelque peu similaire à celle de Turing. Church et Turing démontrèrent rapidement et indépendamment l'équivalence des procédures de Church et de Turing (et donc de Post). Ce furent les conceptions mêmes de Turing qui permirent, dans une très large mesure, l'apparition des ordinateurs. Nous l'avons dit plus haut, l'action d'une machine de Turing est en fait totalement équivalente à celle d'un ordinateur — à condition d'idéaliser ce dernier en lui accordant une capacité de stockage illimitée. Ainsi, la thèse de Church est vue aujourd'hui comme la simple assertion que les algorithmes mathématiques sont précisément ce que peut effectuer un ordinateur idéal — ce qui, compte tenu de la *définition* que l'on donne aujourd'hui du mot « algorithme », devient une pure tautologie. L'acceptation de cette

formulation de la thèse de Church ne renferme certainement aucune contradiction avec $\mathcal{C}$*.

Il est toutefois probable que Turing lui-même ait eu quelque chose de plus à l'esprit, à savoir que les moyens de calcul de tout dispositif *physique* sont (dans l'idéal) nécessairement équivalents au fonctionnement d'une machine de Turing. Une telle affirmation va bien au-delà de ce que Church semble avoir initialement imaginé. Les motivations qui ont poussé Turing à développer le concept de « machine de Turing » se fondaient sur ses idées concernant ce qu'un calculateur humain peut en principe accomplir (voir Hodges 1983). Il semble probable qu'il ait considéré que l'action physique en général — y compris celle du cerveau humain — était toujours réductible, d'une façon ou d'une autre, à l'action d'une machine de Turing. Peut-être devrait-on appeler « thèse de Turing » cette affirmation (physique), afin de la distinguer de l'affirmation (purement mathématique) qu'est la « thèse de Church », thèse qui n'est aucunement contredite par $\mathcal{C}$. C'est d'ailleurs la terminologie que j'adopterai dans ce livre. Ainsi, c'est la *thèse de Turing*, et non la thèse de Church, qui serait contredite par le point de vue $\mathcal{C}$.

1.7 Le chaos

Ces dernières années ont vu se développer, corrélativement à la découverte de phénomènes physiques semblant se comporter de manière extravagante et imprévisible (Fig. 1.1), un grand intérêt pour une discipline mathématique baptisée « chaos ». Le chaos représente-t-il la base physique non algorithmique qu'exige $\mathcal{C}$?

Les *systèmes chaotiques* sont soit des systèmes physiques concrets, soit des simulations mathématiques de ces systèmes, soit encore de simples modèles mathématiques étudiés pour eux-mêmes, dont le comportement dépend de manière extrêmement cruciale de leur état initial. Si les systèmes chaotiques ordinaires sont totalement déterministes et algorithmiques, ils peuvent, *dans la pratique*, se comporter comme s'ils ne l'étaient pas. La raison en est que la précision avec laquelle on doit connaître leur état initial pour pouvoir établir une prédiction déterministe de leur comportement excède, et de loin, toute mesure concevable.

---

* On rencontre parfois, dans certaines discussions mathématiques, une procédure qui est « manifestement » de nature algorithmique bien que l'on ne voie pas forcément immédiatement comment la formuler en termes de machine de Turing ou de lambda-calcul. On peut alors affirmer qu'une telle formulation existe nécessairement « en vertu de la thèse de Church ». Voir par exemple Cutland (1980). Ce raisonnement n'a rien d'erroné et ne contient assurément aucune contradiction avec $\mathcal{C}$. En fait, ce type d'utilisation de la thèse de Church est omniprésent dans la majeure partie de la discussion du chapitre 3.

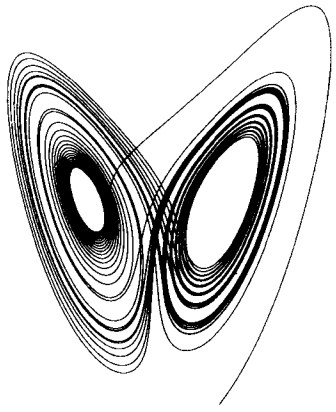


Figure 1.1. L'attracteur de Lorentz — l'un des premiers exemples de systèmes chaotiques. En suivant les courbes, on effectue des allers et retours entre le lobe de gauche et le lobe de droite, de façon apparemment aléatoire, et le lobe dans lequel on se trouve à un instant donné dépend de manière critique du point de départ. Pourtant, la courbe est définie par une simple équation mathématique (différentielle).

Un exemple célèbre de cette situation est la prévision météorologique détaillée à long terme. On connaît parfaitement les lois gouvernant le mouvement des molécules d'air et les autres grandeurs physiques nécessaires à la détermination du temps qu'il fera. Toutefois, les situations météo qui peuvent effectivement survenir en seulement quelques jours dépendent de manière si subtile des conditions initiales qu'il n'existe aucune possibilité de mesurer ces conditions avec une précision suffisante pour établir des prédictions fiables. On pourrait penser que l'énorme quantité de paramètres intervenant dans l'évolution d'une situation météo justifie à elle seule la quasi-impossibilité de telles prédictions, mais le problème est en fait plus profond.

Ce comportement chaotique peut en effet également se manifester pour des systèmes extrêmement simples, consistant seulement en un petit nombre de particules. Imaginons par exemple qu'au billard américain, on vous demande de mettre dans un trou la cinquième boule E d'un ensemble de cinq boules A, B, C, D, E disposées en zigzag* et très espacées les unes des autres, cela en frappant la boule A de sorte qu'elle heurte B qui heurte C qui heurte D qui enfin heurte E qui tombe alors dans le trou. La précision nécessaire pour réussir ce coup dépasse en général de loin les capacités d'un virtuose du billard. En présence de vingt boules, même parfaitement sphériques et élastiques, l'envoi de la dernière boule dans le trou est irréalisable par le plus précis des dispositifs que permettrait de concevoir la technologie moderne. En effet, le comportement des dernières boules est totalement aléatoire, en dépit du fait que les lois

* Lors du premier jet de ce livre, je n'avais pas précisé « en zigzag ». Si les boules sont parfaitement alignées, le problème s'avère tout à fait facile, ainsi que je l'ai vérifié à ma grand surprise directement sur le tapis. L'alignement des boules fait apparaître une stabilité fortuite qui n'existe pas dans le cas général.

newtoniennes gouvernant le mouvement des boules sont mathématiquement déterministes et ne posent en principe aucun problème de calcul. Aucun algorithme ne peut prédire le comportement *réel* des dernières boules, simplement parce qu'on n'a aucun moyen de déterminer avec une précision suffisante la position et la vitesse initiales vraies de la queue, ou les positions des premières boules. En outre, d'infimes effets externes, tels la respiration de quelqu'un dans le village voisin, influent suffisamment sur cette précision pour rendre totalement vain l'accomplissement d'un tel calcul.

Qu'il soit bien clair cependant qu'en dépit d'obstacles aussi incontournables, tous les systèmes ordinaires que l'on qualifie de « chaotiques » *font partie* de ce que j'appelle les systèmes algorithmiques. Pourquoi ? Comme dans d'autres situations que nous rencontrerons plus tard, il suffit pour décider de la calculabilité d'une procédure de se demander : peut-elle être exécutée par un ordinateur standard ? La réponse est dans ce cas manifestement « oui », pour la simple raison que les systèmes chaotiques mathématiquement descriptibles sont de fait habituellement étudiés à l'aide d'un ordinateur !

Bien sûr, si nous tentons de prédire numériquement les conditions météo qui régneront sur l'Europe durant une semaine, ou les collisions successives de vingt boules de billard très espacées et disposées en zigzag, notre simulation risque de ne ressembler en rien à ce qui se passera *en réalité*. C'est là la caractéristique des systèmes chaotiques. Il est en pratique impossible de prédire sur ordinateur le comportement *réel* du système. Cependant, on peut parfaitement parvenir à simuler un comportement *typique* de ce système. Le temps prédit n'est peut-être pas le temps qu'il fera effectivement, mais il est parfaitement *plausible* ! De même, le comportement prédit pour les boules de billard est tout à fait acceptable en tant que comportement possible, même si en fait les boules peuvent subir des collisions très différentes de celles qui sont calculées. Un autre point qui met en relief la nature parfaitement algorithmique de ces opérations est que si l'on exécute plusieurs fois la simulation numérique en utilisant les mêmes données de départ, on obtient chaque fois un résultat *exactement* identique ! (Cela suppose que l'ordinateur ne commette aucune erreur ; mais les erreurs de calcul sont exceptionnelles chez les ordinateurs modernes.)

L'intelligence artificielle ne cherche pas à simuler le comportement de l'individu X ; elle serait amplement satisfaite si elle parvenait à simuler celui d'*un* individu ! Il n'est donc pas du tout déraisonnable d'adopter le point de vue que je défends, à savoir que les systèmes chaotiques se classent indéniablement dans la catégorie des systèmes algorithimiques. La simulation numérique d'un système chaotique correspond en fait à un « cas typique » parfaitement raisonnable, même si elle peut différer du « cas réel ». Si les manifestations extérieures de l'intelligence humaine étaient le résultat d'une évolution dynamique chaotique — évolution algorithmique dans le sens que je viens de décrire —, cela serait conforme aux points de vue $\mathcal{A}$ et $\mathcal{B}$, mais *pas* à $\mathcal{C}$.

On a parfois suggéré que la présence du chaos dans l'activité cérébrale expliquerait pourquoi notre cerveau a un comportement *apparemment* différent de l'activité algorithmiquement déterministe d'une machine de Turing, même si, ainsi que nous l'avons souligné, ce comportement est techniquement algorith-

mique. J'aurai à revenir sur ce sujet (*cf.* §3.22). Pour le moment, il doit être clair que les systèmes chaotiques *appartiennent* à la catégorie de systèmes que j'appelle « algorithmiques », car pouvant en théorie être simulés par le *calcul*. Le fait que, en pratique, quelque chose puisse ou non être simulé est un autre problème, distinct des problèmes *de principe* examinés ici.

1.8 Le calcul analogique

J'ai jusqu'ici uniquement considéré le mot « calcul » dans le sens où on l'entend pour les ordinateurs numériques modernes ou, plus précisément, pour leurs précurseurs théoriques : les machines de Turing. Il existe d'autres types de machines à calculer, utilisés surtout dans les décennies précédentes, dans lesquels les opérations sont représentées non pas par les classiques états discrets « ouvert/fermé » des calculs numériques, mais par des paramètres physiques continus. La plus familière de ces machines est la règle à calcul dont le paramètre de base est la distance linéaire (le long de la règle) correspondant au logarithme des nombres à multiplier ou à diviser. Il existe de nombreux types de calculateurs analogiques, associés à d'autres paramètres physiques tels que le temps, la masse ou le potentiel électrique.

Les systèmes analogiques nous confrontent au problème technique suivant : les notions standard de calcul et de calculabilité ne s'appliquent, strictement parlant, qu'aux systèmes *discrets* (donc concernant des opérations « numériques ») et non aux systèmes *continus*, comme par exemple les distances ou les potentiels électriques intervenant en physique classique. L'application des notions ordinaires de calcul à un système dont la description exige des paramètres continus et non discrets (« numériques ») nécessite donc des *approximations*. De fait, lors des simulations de systèmes physiques sur ordinateur, la procédure normale consiste à faire une approximation en discrétisant tous les paramètres continus pertinents. Outre que cette procédure entraîne bien sûr une certaine erreur, l'étude de certains systèmes physiques ne peut se satisfaire du degré de précision dont dispose l'ordinateur. Ainsi, la simulation d'un système physique continu sur calculateur numérique peut conduire à des conclusions erronées sur le comportement du système.

En principe, on peut toujours accroître cette précision jusqu'à un niveau adéquat pour simuler correctement le système. Toutefois, et particulièrement dans le cas des systèmes chaotiques, le temps de calcul et la capacité de mémoire alors exigés peuvent se révéler en pratique totalement prohibitifs. De surcroît, on n'est jamais certain que le degré de précision choisi *est* suffisant. Il faut introduire dans la simulation des tests signalant que la précision *est* suffisante et donne un comportement qualitatif réaliste. Cette situation soulève un certain nombre de problèmes mathématiques passablement délicats sur lesquels je ne m'attarderai pas ici.

L'étude des systèmes continus admet cependant d'autres approches. Celles-ci traitent le système comme une structure mathématique à part entière, avec sa *propre* notion de « calculabilité » — notion qui généralise, du cas discret au cas continu, celle de calculabilité au sens de Turing[12]. Grâce à cette notion, l'application de la calculabilité au sens de Turing n'exige plus le recours à l'approximation d'un système continu par des paramètres discrets. Pour intéressantes qu'elles soient au niveau mathématique, ces idées ne semblent malheureusement pas pour l'instant avoir atteint l'unicité et le caractère naturel propres à la notion standard de calculabilité au sens de Turing pour les systèmes discrets. En outre, certaines anomalies provoquent l'apparition d'une « non-calculabilité » technique chez des systèmes simples pour lesquels il n'est pas évident qu'une telle terminologie soit réellement appropriée (*e.g.*, même pour la simple « équation d'onde » de la physique ; *cf.* Pour-El et Richards 1981, EOLP, p. 202-203). Indiquons cependant qu'un travail relativement récent (Rubel 1989) a montré que les calculateurs analogiques théoriques, appartenant à une classe relativement vaste, ne peuvent dépasser la calculabilité au sens de Turing. Ce sont là, à mon sens, des problèmes importants et intéressants que des recherches plus approfondies permettront d'éclaircir. Je n'ai toutefois pas le sentiment que ces travaux soient, dans l'ensemble, suffisamment avancés pour qu'on puisse les appliquer de manière définitive aux problèmes examinés ici.

Dans ce livre, je m'intéresserai particulièrement au problème de la nature calculable de l'activité mentale, « calculable » signifiant *calculable au sens de Turing*. Les ordinateurs actuels sont de fait des machines numériques, et c'est cela qui importe aujourd'hui pour la recherche en IA. Rien n'interdit d'imaginer qu'existeront dans l'avenir des « ordinateurs » d'un type différent, qui feront un usage *crucial* des paramètres physiques continus — tout en demeurant dans le cadre théorique standard de la physique actuelle —, ce qui leur permettra de se comporter de façon fondamentalement différente d'un ordinateur numérique.

Ces problèmes concernent toutefois principalement la distinction entre les versions « forte » et « faible » du point de vue $\mathcal{C}$. Selon la version *faible* de $\mathcal{C}$, il y aurait des actions physiques sous-tendant le comportement du cerveau humain conscient, non calculables au sens de Turing, mais que l'on pourrait totalement interpréter dans le cadre des théories physiques actuelles. Ces actions dépendraient de paramètres physiques continus, d'une façon qui les rendrait impropres à une simulation par les procédures numériques standard. Selon la version *forte* de $\mathcal{C}$ en revanche, cette non-calculabilité proviendrait d'une théorie physique non calculable — encore à découvrir — dont les conséquences seraient des éléments essentiels de l'activité cérébrale consciente. Si cette seconde possibilité leur semble tirée par les cheveux, les partisans de $\mathcal{C}$ n'ont donc d'autre alternative que de trouver une action continue issue des lois physiques connues et n'admettant aucune simulation numérique, même approchée. Toutefois, presque tout le monde s'accorde actuellement à penser que tous les types de systèmes analogiques fiables qui ont été à ce jour sérieusement envisagés *devraient* admettre — du moins en principe — une simulation numérique correcte.

Aujourd'hui, même en négligeant ces problèmes théoriques d'ordre général, les avantages sont plus du côté des ordinateurs numériques que des ordinateurs analogiques. L'action numérique est bien plus précise, essentiellement parce que la précision s'accroît en augmentant simplement la longueur des nombres, ce que l'on obtient facilement en recourant à un accroissement modeste (logarithmique) de la capacité de l'ordinateur, tandis que la précision des machines analogiques (du moins de celles qui sont *entièrement* analogiques et ne contiennent aucun concept numérique) ne peut s'accroître que par une augmentation relativement énorme (linéaire) de leur capacité. Il se peut que des idées nouvelles fassent un jour basculer l'avantage en faveur des machines analogiques, mais avec la technologie actuelle, la plupart des avantages pratiques significatifs semblent se situer fortement du côté du calcul numérique.

1.9 Quel type d'action pourrait-il échapper au calcul ?

La plupart des types d'action bien définis venant spontanément à l'esprit seraient donc à inclure sous la rubrique que j'ai baptisée « calculable » (dans le sens de « numériquement calculable »). Le lecteur commence peut-être à se demander si $\mathcal{C}$ a encore quelque chose à se mettre sous la dent. Je n'ai toutefois rien dit des actions strictement *aléatoires* pouvant résulter par exemple d'une donnée issue d'un système quantique. (La mécanique quantique sera examinée en détail dans la deuxième partie, aux chapitres 5 et 6.) Il est cependant difficile de voir quel serait l'avantage d'un système possédant des données de départ *véritablement* aléatoires plutôt que simplement *pseudo-aléatoires*, *pouvant être* entièrement générées par le calcul (*cf.* §3.11). En fait, bien qu'il y ait, strictement parlant, quelques différences techniques entre « aléatoire » et « pseudo-aléatoire », elles semblent avoir peu d'importance pour les problèmes de l'IA. Plus loin (§.3.11, 3.18 *et sq.*), je donnerai de solides arguments montrant qu'une donnée « purement aléatoire » ne présente pour nous aucune utilité ; il convient plutôt de s'en tenir au caractère pseudo-aléatoire du comportement chaotique — et, comme il a été souligné plus haut, tous les types ordinaires de comportement chaotique sont modélisables par le calcul.

Qu'en est-il du rôle de l'environnement ? Lors de son développement, chaque être humain se trouve doté d'un environnement unique, qu'il ne partage avec aucun autre être humain. Se peut-il que cet environnement personnel et original fournisse à chacun d'entre nous des données de départ qui échappent au calcul ? J'avoue que je vois difficilement l'avantage que pourrait offrir dans ce contexte l'« originalité » de notre environnement. Cette discussion est semblable à celle que nous avons eue sur le chaos (*cf.* §1.7). La simulation d'un environnement (chaotique) *plausible* étant disponible, elle suffit pour « éduquer » un robot géré par ordinateur. Ce robot n'a pas besoin d'acquérir

ses compétences au contact d'un environnement réel ; un environnement *typique* (plutôt que réel) simulé sur ordinateur lui suffit certainement.

N'y aurait-il pas toutefois une impossibilité intrinsèque qui empêcherait toute simulation numérique d'un environnement, même seulement plausible ? Peut-être *y a-t-il* dans le monde physique extérieur un élément qui se dérobe effectivement à toute simulation numérique. Quelques partisans de $\mathcal{A}$ ou de $\mathcal{B}$ attribuent parfois l'impossibilité apparente de modéliser certaines actions humaines à une non-calculabilité affectant cet environnement extérieur. Il serait cependant imprudent, pour les partisans de $\mathcal{A}$ ou de $\mathcal{B}$, de s'appuyer sur cet argument, car accepter l'existence, *quelque part* dans le monde physique, d'un élément dont le comportement ne puisse être simulé numériquement remettrait en question ce qui est probablement leur principale raison de douter de $\mathcal{C}$. S'il existe, dans l'environnement extérieur, des actions qui échappent à toute simulation numérique, pourquoi n'existerait-il pas dans le cerveau des actions *internes* elles aussi numériquement insaisissables ? L'organisation physique interne du cerveau humain est somme toute bien plus sophistiquée que la majeure partie (au moins) de son environnement — sauf peut-être lorsque cet environnement est lui-même fortement influencé par les actions d'autres cerveaux humains. L'acceptation d'une action physique *externe* non calculable invalide le principal argument opposé à $\mathcal{C}$. (Voir aussi les sections 3.9 et 3.10.)

L'idée suggérée par le point de vue $\mathcal{C}$, selon laquelle il peut exister quelque chose d'impossible à simuler par le calcul, mérite un autre commentaire. Elle n*e signifie pas* simplement qu'il existe quelque chose qui échappe à tout calcul *concret*. On pourrait en effet affirmer que la simulation précise de tout environnement plausible — ou de tous les processus physiques et chimiques intervenant dans le cerveau —, bien qu'en principe possible, exigerait un temps de calcul ou un espace mémoire si considérable qu'il n'y aurait aucun espoir de la réaliser sur un ordinateur, actuel ou à venir. Peut-être la simple rédaction d'un programme informatique approprié est-elle absolument inenvisageable à cause du grand nombre de facteurs dont il faut tenir compte. Pour pertinentes que soient ces considérations (elles seront examinées à la section 2.6, en **Q8** et à la section 3.5), elles *ne correspondent pas* à ce que j'entends par la « non-calculabilité » affirmée par $\mathcal{C}$. Par « non calculable », je veux dire « échappant en principe à tout calcul » — je vais préciser ce point dans un instant. Les calculs que l'on ne peut exécuter à l'aide des méthodes ou des machines actuelles ou à venir sont encore des « calculs » au sens technique du terme.

Le lecteur peut bien sûr se demander : s'il n'y a rien de « non calculable » dans l'aléatoire ni dans les influences extérieures, voire dans la complexité pure et indomptable, que puis-je donc avoir à l'esprit lorsque j'utilise cet adjectif ? Ce que j'ai à l'esprit repose sur certains types de problèmes mathématiques bien définis dont on peut *démontrer* qu'ils échappent à tout calcul. En l'état actuel des connaissances, aucune activité mathématique de ce type n'est requise pour décrire le monde physique. Cependant, c'est une possibilité logique ; c'est même *plus* qu'une possibilité logique. J'affirme ici qu'il existe quelque chose de totalement non calculable qui *est* inhérent aux lois

physiques, même s'il n'a pas encore été rencontré dans la physique que l'on connaît. Certaines formes remarquablement simples de ce type de problèmes vont me permettre d'illustrer ce que j'ai en tête.

Je commencerai par décrire quelques exemples de problèmes mathématiques bien définis qui — en un sens que je vais expliquer dans un moment — n'admettent aucune solution par le calcul. A partir de n'importe lequel de ces problèmes, on peut construire de toutes pièces un « modèle » d'univers physique dont le comportement, bien qu'entièrement déterministe, se dérobe à toute simulation numérique.

Le premier exemple d'un tel problème est le plus célèbre de tous. Il s'agit du « dixième problème de Hilbert ». Formulé en 1900 par le grand mathématicien David Hilbert, il fait partie d'une liste de problèmes mathématiques qui ont énormément contribué au développement des mathématiques du début du XX^{e} siècle (et même d'aujourd'hui). Le dixième problème de Hilbert consistait à chercher une procédure de calcul permettant de décider si un système d'équations *diophantiennes* admet une solution.

Qu'est-ce qu'une équation diophantienne ? C'est une équation polynomiale à un nombre quelconque de variables dont tous les coefficients et toutes les solutions doivent être des entiers relatifs. (Un entier relatif est simplement un nombre de la suite ..., -3, -2, -1, 0, 1, 2, 3, 4, ... Les équations diophantiennes furent pour la première fois systématiquement étudiées au III^{e} siècle avant notre ère par le mathématicien grec Diophante.) Un système d'équations diophantiennes est par exemple

$$6w + 2x^2 - y^3 = 0\,, \quad 5xy - z^2 + 6 = 0\,, \quad w^2 - w + 2x - y + z - 4 = 0$$

ou encore

$$6w + 2x^2 - y^3 = 0\,, \quad 5xy - z^2 + 6 = 0\,, \quad w^2 - w + 2x - y + z - 3 = 0$$

Le premier système admet notamment comme solution

$$w = 1\,,\ x = 1\,,\ y = 2\,,\ z = 4$$

tandis que le second n'en admet aucune (parce que, par sa première équation, y doit être pair, par sa seconde équation, z doit être également pair, conditions qui contredisent la troisième équation quel que soit w, parce que $w^2 - w$ est toujours pair et que 3 est impair). Le problème posé par Hilbert consistait donc à trouver une procédure mathématique — un *algorithme* —permettant de décider quels sont les systèmes d'équations diophantiennes qui, comme notre premier exemple, possèdent des solutions, et quels sont ceux qui n'en possèdent pas. Rappelons (*cf.* §1.5) qu'un algorithme est simplement une procédure de calcul — l'action d'une machine de Turing. Ainsi, le dixième problème de Hilbert demande une procédure de calcul permettant de déterminer à quelles conditions un système d'équations diophantiennes est soluble.

Ce problème fut historiquement très important dans la mesure où il soulevait une question inédite. Que *signifie* réellement, en termes mathématiques précis, qu'une classe de problèmes possède une solution algorithmique ?

Qu'est-ce précisément qu'un algorithme ? Ce fut cette question même qui, en 1936, amena Alan Turing à proposer sa propre définition d'un algorithme — en termes de machine de Turing. D'autres mathématiciens (notamment Church, Kleene, Gödel, Post ; *cf.* Gandy 1988) proposèrent vers la même époque des procédures légèrement différentes. Church et Turing montrèrent rapidement qu'elles étaient toutes équivalentes, mais l'approche particulière de Turing se révéla plus féconde. (Il fut le seul à introduire le concept de machine algorithmique généraliste — appelée machine de Turing *universelle* — pouvant exécuter *n'importe quelle* opération algorithmique, concept qui devait donner naissance à celui d'ordinateur aujourd'hui si familier.) Turing put montrer qu'il existe des classes de problèmes n'admettant *aucune* solution algorithmique (en particulier, le « problème de l'arrêt » que je décrirai bientôt). Le dixième problème de Hilbert dut toutefois attendre 1970 pour que le mathématicien russe Yuri Matiyasevich — établissant un théorème complétant certains arguments avancés antérieurement par les Américains Julia Robinson, Martin Davis et Hilary Putnam — montre qu'aucun programme informatique (aucun algorithme) ne peut décider systématiquement si un système d'équations diophantiennes possède ou non une solution. (Pour un récit accessible de cette histoire, voir Davis 1978 et Devlin 1988, chapitre 6.) Il faut remarquer que si un système admet des solutions, cela peut en principe être contrôlé par un programme informatique qui essaie alors servilement tous les entiers relatifs les uns après les autres. En revanche, une réponse négative interdit tout traitement systématique. S'il existe diverses règles permettant d'aboutir à une réponse négative — comme l'argument sur la parité des nombres pour le second système proposé plus haut —, le théorème de Matiyasevich montre que ces ensembles de règles ne peuvent *jamais* être exhaustifs.

Un autre exemple de problème mathématique bien défini ne possédant pas de solution algorithmique est constitué par le *pavage du plan.* Il se formule ainsi : étant donné un ensemble de formes polygonales, quelles sont celles qui pavent le plan ? Autrement dit, est-il possible, sans laisser de vide ni provoquer de chevauchements, de recouvrir entièrement le plan euclidien en utilisant uniquement cet ensemble de formes ? En 1966, se fondant sur une extension de travaux publiés en 1961 par le mathématicien sino-américain Hao Wang (voir Grünbaum et Shephard 1987), le mathématicien américain Robert Berger montra que ce problème est insoluble algorithmiquement. En fait, tel que je l'ai énoncé, il présente un point délicat dans la mesure où les « pavés » polygonaux se définissent à l'aide de nombres réels (nombres présentant une partie décimale infinie), tandis que les algorithmes ordinaires opèrent sur des nombres entiers. On peut contourner cette difficulté en se restreignant à des pavés faits d'une juxtaposition de carrés. Ces pavés sont appelés des *polyminos* (voir Golomb 1965 ; Gardner 1965, chapitre 13 ; Klarner 1981). La figure 1.2 montre des exemples de polyminos. (Pour d'autres exemples d'ensembles de pavés, voir EOLP, p. 143-147, Fig. 4.6-4.12.) Curieusement, l'insolubilité algorithmique du problème du pavage dépend de l'existence de certains ensembles de polyminos appelés ensembles *apériodiques* — qui pavent le plan de manière *uniquement non périodique* (*i.e.* de sorte que le motif obtenu avec

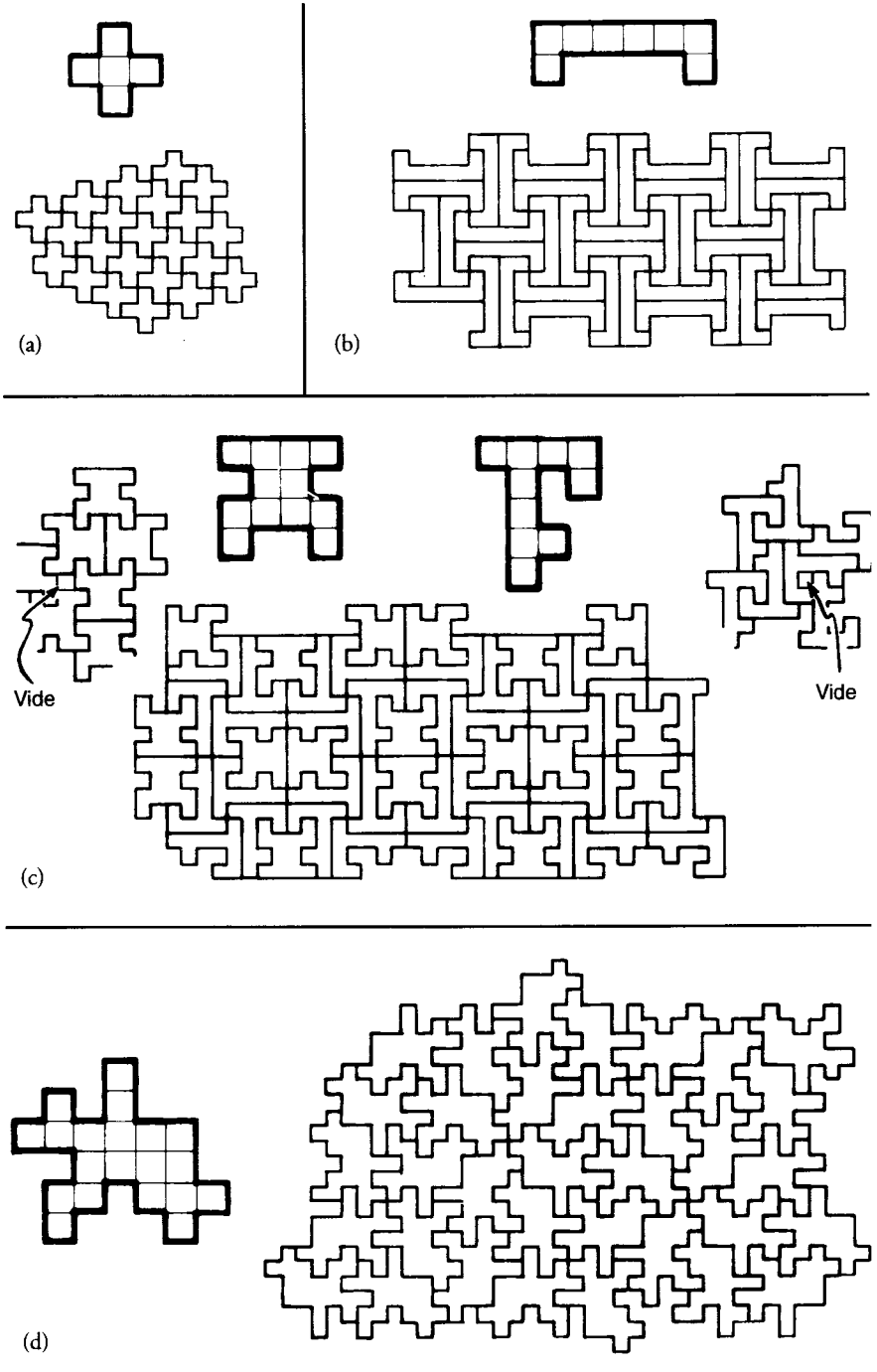


Figure 1.2. Ensembles de polyminos pavant le plan euclidien infini (la rotation des éléments du pavage est autorisée). Remarquez qu'en (c), aucun des polyminos de base ne pave à lui seul le plan entier.

ces polyminos ne se répète jamais, quelle que soit son étendue). La figure 1.3 montre un ensemble apériodique de trois polyminos (constitué à partir d'un ensemble de pavés découvert par Robert Ammann en 1977, *cf.* Grünbaum et Shephard 1987, Fig. 10.4.11-10.4.13, p. 555-556).

Les démonstrations de l'insolubilité algorithmique du dixième problème de Hilbert et du problème du pavage sont loin d'être évidentes, et je ne tenterai certainement pas ici d'en donner les arguments[13]. Le point central de chacun de ces raisonnements consiste en fait à montrer comment on peut encoder une machine de Turing pour qu'elle exécute un pavage du plan ou résolve un système d'équations diophantiennes. Le problème se ramène alors à celui que Turing avait traité dans sa discussion originale : l'insolubilité algorithmique du *problème de l'arrêt* — qui consiste à savoir dans quels cas l'action d'une machine de Turing ne connaît jamais de fin. Je donnerai à la section 2.3 divers calculs explicites qui ne s'arrêtent *jamais*, et à la section 2.5 une argumentation relativement simple — s'inspirant essentiellement du raisonnement original de Turing — qui montre, entre autres choses, que le problème de l'arrêt est effectivement insoluble par le calcul. (Les conséquences des « autres choses » démontrées par cette argumentation joueront un rôle central dans toute la discussion de la deuxième partie !)

Voyons maintenant comment utiliser une classe de problèmes, tels les équations diophantiennes ou le pavage du plan, pour construire un modèle d'univers déterministe mais non calculable. Supposons que ce modèle soit doté d'un *temps discret* paramétré par les entiers naturels (les entiers non négatifs) 0, 1, 2, 3, 4, ... Au temps n, on spécifie l'état de l'univers par un problème de la classe considérée — disons par un ensemble de polyminos. Deux règles définissent l'état de l'univers au temps $n + 1$ en fonction de l'ensemble de polyminos représentant l'univers au temps n. On adopte la première règle si l'ensemble de polyminos *pave* le plan, la seconde lorsque l'ensemble de polyminos *ne pave pas* le plan. Ces deux règles — dont le détail nous importe peu ici — permettent par exemple de former une liste S_0, S_1, S_2, S_3, S_4, S_5, ... de tous les ensembles possibles de polyminos telle que ceux utilisant un nombre total *pair* de carrés possèdent des indices pairs — S_0, S_2, S_4, S_6, ... — et ceux utilisant un nombre total *impair* de carrés possèdent des indices impairs — S_1, S_3, S_5, S_7, ... (Cet arrangement s'obtient relativement facilement à l'aide d'une procédure de calcul.) L'« évolution dynamique » de notre modèle d'univers est alors donnée par :

> L'état d'univers S_n au temps n devient S_{n+1} au temps $n + 1$ si l'ensemble de polyminos S_n *pave* le plan, et S_{n+2} si l'ensemble S_n *ne pave pas* le plan.

Un tel univers se comporte de manière totalement déterministe. Mais comme il n'existe pas de procédure de calcul générale pour s'assurer qu'un ensemble de polyminos S_n pave le plan (que le nombre total de carrés soit pair ou impair), il n'existe aucune simulation numérique permettant de suivre l'évolution réelle de cet univers. (Fig. 1.4.)

Bien sûr, cet « univers » n'est pas un modèle du véritable Univers dans lequel nous vivons. Il est simplement présenté ici (et dans EOLP, p. 182) pour

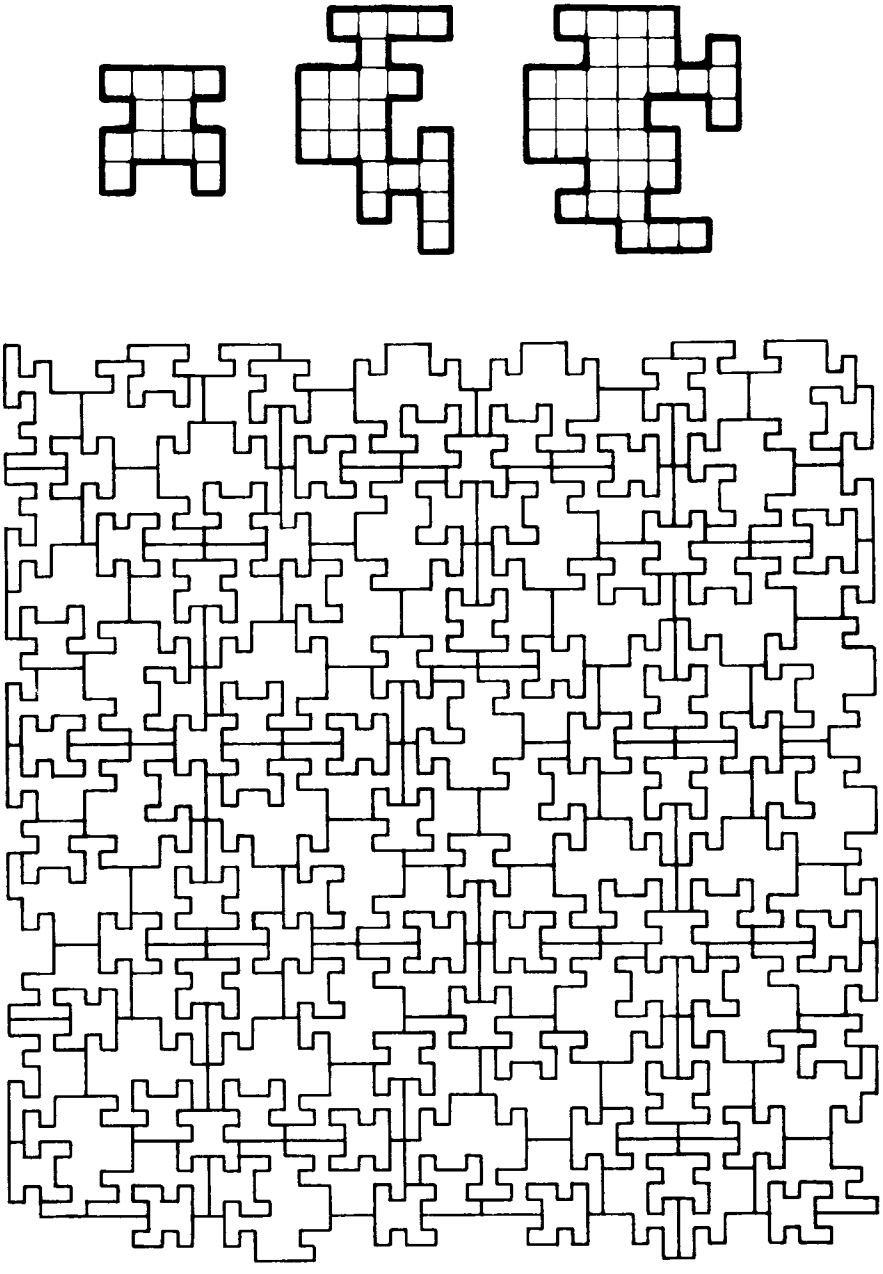


Figure 1.3. Les trois polyminos ci-dessus pavent le plan entier, mais de manière non périodique (éléments de pavages dus à Robert Ammann).

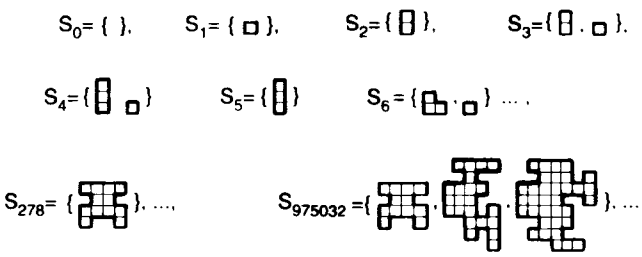


Figure 1.4. Modèle d'univers non calculable. Les différents états de ce modèle d'univers déterministe mais non calculable s'expriment à l'aide des ensembles finis possibles de polyminos numérotés par S_n de sorte que les indices n pairs correspondent à un nombre total pair de carrés et les indices impairs à un nombre total impair de carrés. L'évolution temporelle suit l'ordre numérique (S_0, S_2, S_3, S_4, ..., S_{278}, S_{280}, ...), sauf que l'on saute un nombre chaque fois que l'ensemble précédent ne pave pas le plan.

illustrer la différence nette — mais souvent mal perçue — entre déterminisme et calculabilité. *Il existe des modèles d'univers complètement déterministes, avec des règles d'évolution bien définies, qui sont impossibles à simuler numériquement.* Nous le verrons à la section 7.9, les types très particuliers de modèles que je viens de considérer ne suffisent pas à fournir ce que requiert le point de vue $\mathcal{C}$. Mais nous verrons aussi à la section 7.10 que ce qui est requis, curieusement, pourrait bien être apporté par la physique !

1.10 Qu'en est-il de l'avenir ?

Que nous disent les points de vue $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ et $\mathcal{D}$ sur l'avenir de notre planète ? Selon $\mathcal{A}$, un jour viendra où des superordinateurs convenablement programmés atteindront — puis dépasseront — le niveau des capacités mentales de l'être humain. Bien sûr, les partisans de $\mathcal{A}$ ne datent pas unanimement ce jour. Certains, constatant le peu de connaissances actuelles sur les calculs que le cerveau doit (selon eux) doit effectuer pour parvenir à notre indéniable subtilité d'action — subtilité nécessaire avant que puisse émerger une « conscience » digne de ce nom —, estiment raisonnable d'envisager que de nombreux siècles s'écouleront avant que les ordinateurs atteignent notre niveau. D'autres penchent pour une échéance plus courte. En particulier, Hans Moravec, dans son livre *Mind Children* (1988), présente un argument — fondé sur l'accélération du développement de la technologie informatique au cours des cinquante dernières années et sur la part de l'activité cérébrale qui est, selon lui, déjà correctement simulée — visant à montrer que l'« équivalent humain » sera déjà dépassé vers l'an 2030. (D'autres encore avancent une échéance bien plus courte[14] — affirmant parfois que l'équivalent humain a

déjà été dépassé !) Que le lecteur consterné à l'idée d'être supplanté par l'ordinateur dans moins de (disons) quarante ans se rassure. On lui accorde l'espoir — en fait la promesse — que nous pourrons à notre guise transférer nos « programmes mentaux » dans les corps métalliques (ou plastiques) de rutilants robots et nous assurer ainsi une forme d'immortalité (Moravec 1988, 1994).

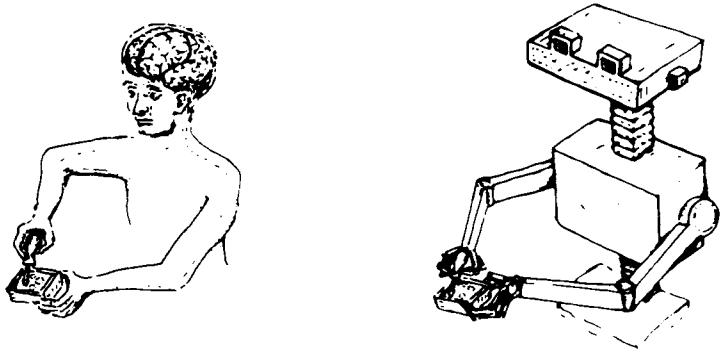


Figure 1.5. Le point de vue $\mathcal{B}$ affirme la possibilité de principe d'une simulation sur ordinateur de l'activité d'un cerveau humain conscient. Des robots contrôlés numériquement pourraient alors atteindre et surpasser de loin toutes les capacités de l'homme.

Les défenseurs du point de vue $\mathcal{B}$ n'affichent toutefois pas un tel optimisme. Certes ils s'accordent avec les partisans de $\mathcal{A}$ sur ce que les ordinateurs finiront par accomplir extérieurement. Mais s'ils reconnaissent qu'une *simulation* adéquate de l'activité cérébrale humaine suffirait en elle-même pour commander un robot (Fig. 1.5), ils nient la présence d'une connaissance consciente associée à cette simulation. Qu'elle mette des siècles ou moins de quarante ans à se réaliser, une telle simulation, selon $\mathcal{B}$, deviendra un jour une possibilité technique. Les ordinateurs auront alors atteint le niveau de l'« équivalent humain » et dépasseront par la suite ce niveau d'aptitude. Il ne nous sera pas possible, cependant, de fusionner avec ces robots et il semble que nous devions nous résigner à la perspective d'une planète finalement gouvernée par des machines insensibles ! Des quatre points de vue $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ et $\mathcal{D}$, $\mathcal{B}$ me semble offrir la vision la plus pessimiste pour l'avenir de notre planète — bien qu'elle soit apparemment inspirée par le « sens commun » !

Selon $\mathcal{C}$ ou $\mathcal{D}$ en revanche, les ordinateurs resteront — ou devraient rester — éternellement à notre service, quels que soient les progrès qu'ils connaîtront en termes de vitesse, de capacité et de structure logique. Le point de vue $\mathcal{C}$ accepte cependant la perspective de développements scientifiques futurs pouvant déboucher sur la construction de dispositifs — basés *non pas* sur les ordinateurs tels que nous les concevons aujourd'hui, mais sur précisément l'action physique non calculable dont $\mathcal{C}$ affirme qu'elle sous-tend nos processus de pensée consciente — doués d'une intelligence et d'une conscience

réelles. Peut-être seront-ce *ces* dispositifs, plutôt que les « ordinateurs » — dans leur conception actuelle — qui surpasseront toutes les capacités humaines. Toutefois, de telles spéculations me semblent actuellement extrêmement prématurées. Notre manque de connaissances scientifiques est sur ce point quasi total, pour ne rien dire de notre savoir-faire technologique. Je reviendrai sur cette question dans la deuxième partie (*cf.* §8.1).

1.11 Les ordinateurs ont-ils des droits et des responsabilités ?

Un problème connexe — qui pourrait avoir une importance pratique un peu plus immédiate — a commencé d'attirer l'attention des législateurs[15]. Ce problème est de savoir si, dans un futur pas trop lointain, on ne devra pas envisager que les ordinateurs aient, sur un plan légal, des responsabilités ou des droits. Il est certain que si, à la longue, ils doivent approcher, voire dépasser, les niveaux de compétence humaine dans de nombreuses situations sociales, les questions de ce genre ne pourront manquer de se poser. Si l'on adhère au point de vue $\mathscr{A}$, on se résout clairement à accepter que les ordinateurs (ou les robots informatisés) puissent avoir à la fois des droits et des responsabilités. Selon ce point de vue en effet, il n'y a aucune différence essentielle — hormis celles, fortuites, concernant les matériaux constitutifs — entre nous-mêmes et des robots suffisamment perfectionnés. La situation semble moins claire pour les partisans de $\mathscr{B}$. On peut raisonnablement affirmer que le problème des droits ou des responsabilités repose sur la possession ou la non-possession de certaines qualités mentales authentiques — par exemple la souffrance, la colère, l'esprit de vengeance, la méchanceté, la foi, la confiance, l'intention, la croyance, la compréhension ou la passion. Selon $\mathscr{B}$, un robot informatisé ne possède aucune de ces qualités et n'aurait dès lors, me semble-t-il, ni droits ni responsabilités. Pourtant, toujours selon $\mathscr{B}$, il n'existe aucun moyen sûr de contrôler que ces qualités sont effectivement absentes, de sorte que l'on risque de se trouver confronté à un dilemme dans le cas où les robots parviendraient à imiter suffisamment fidèlement le comportement humain.

Ce dilemme semble éliminé par $\mathscr{C}$ (et probablement aussi par $\mathscr{D}$) parce que, selon ces deux points de vue, les ordinateurs ne peuvent *se comporter* de façon à faire croire à de telles qualités mentales — et ne pourront certainement jamais en posséder. Il s'ensuit que les ordinateurs n'ont *ni* droits *ni* responsabilités. C'est là selon moi un point de vue très raisonnable. Dans ce livre, je m'opposerai fermement à $\mathscr{A}$ et $\mathscr{B}$. L'acceptation des arguments que j'expose devrait simplifier le problème légal : les ordinateurs ou les robots gérés par ordinateur n'ont *jamais* de droits ni de responsabilités. En outre, il n'y a pas à les impliquer lorsque les choses vont mal — c'est ailleurs que l'on doit alors chercher !

Il faut cependant être bien clair : ces arguments ne s'appliquent pas nécessairement aux « dispositifs » hypothétiques qui parviendraient un jour à tirer avantage de la physique non algorithmique. Mais la perspective de tels dispositifs — s'ils sont un jour construits — étant loin de se concrétiser, nous n'aurons pas à nous préoccuper, dans un avenir prévisible, de problèmes légaux à leur sujet.

Cette question de la « responsabilité » soulève de profonds problèmes philosophiques concernant les causes profondes de notre comportement. On peut en effet affirmer que chacun de nos actes est en définitive déterminé par notre patrimoine génétique et notre environnement — voire par les multiples éléments de hasard intervenant continuellement dans le déroulement de notre vie. *Toutes* ces influences n'échappent-elles pas à notre contrôle et donc finalement à notre responsabilité ? La notion de responsabilité est-elle simplement une terminologie commode, ou existe-t-il quelque chose d'autre — un « moi » se dérobant à toutes ces influences — qui exerce un contrôle sur nos actes ? D'un point de vue légal, le problème de la « responsabilité » *semble* impliquer que chacun de nous recèle une sorte de « moi » indépendant doté de ses *propres* responsabilités — et donc de droits propres — dont les actes *ne sont pas* imputables à l'héritage génétique, à l'environnement ou au hasard. Si parler d'un tel « moi » indépendant est plus qu'une simple commodité de langage, il existe alors un élément manquant dans notre compréhension actuelle de la physique. La découverte d'un tel élément modifierait certainement radicalement notre vision de la science.

Ce livre n'apportera pas de réponse à ces problèmes, mais je suis persuadé qu'il peut permettre de les aborder — ne serait-ce que très modestement. Il ne dira pas qu'il y a nécessairement un « moi » dont les actes ne sont imputables à aucune cause extérieure, mais il invitera à élargir notre vision de ce que peut être la nature même d'une « cause ». Une « cause » est peut-être une entité non calculable en pratique ou en principe. Je tenterai de montrer que lorsqu'une « cause » est le produit de nos actes conscients, elle est alors nécessairement quelque chose de très subtil, certainement irréductible non seulement au calcul et au chaos, mais aussi à toute influence purement aléatoire. Et le futur dira si un tel concept de cause peut nous rapprocher d'une compréhension du profond problème (ou de l'« illusion » ?) de notre libre arbitre.

1.12 « Connaissance immédiate », « compréhension », « conscience », « intelligence »

Je n'ai rien dit jusqu'ici sur le sens que je donne aux concepts passablement flous associés au problème de l'« esprit ». Les énoncés $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ et $\mathcal{D}$ de la section 1.3 ont vaguement fait référence à la « connaissance immédiate », sans toutefois mentionner d'autres qualités mentales. Je vais

maintenant tenter de clarifier ma propre terminologie, et en particulier les termes « compréhension », « conscience » et « intelligence », qui auront une certaine importance dans les discussions qui vont suivre.

Je ne crois cependant pas nécessaire de donner des définitions précises et je me limiterai à quelques commentaires. Je suis souvent surpris de constater qu'un emploi de ces mots qui me semble aller de soi est contesté par d'autres personnes. Par exemple, le mot « compréhension » implique indiscutablement selon moi la présence, à un certain degré, d'une *connaissance immédiate.* On ne peut véritablement comprendre un argument si on n'en a pas la moindre connaissance immédiate. Cela me semble une évidence. Pourtant, dans certains contextes, les partisans de l'IA donnent parfois aux mots « compréhension » et « connaissance immédiate » un sens qui nie cette implication. Certains adeptes de l'IA (qu'ils défendent les points de vue $\mathcal{A}$ ou $\mathcal{B}$) affirment qu'un robot commandé par ordinateur « comprend » ses instructions, même si l'on ne peut prétendre qu'il en a une « connaissance immédiate ». C'est là à mon avis un usage abusif du verbe « comprendre », même s'il possède une réelle valeur heuristique pour la description du fonctionnement d'un ordinateur. Lorsque, parlant d'une activité exigeant la présence d'une connaissance immédiate, je voudrai indiquer clairement que je n'emploie pas le mot « compréhension » dans ce sens heuristique, j'emploierai des expressions comme « compréhension véritable » ou « comprend véritablement ».

On peut certes objecter qu'il n'existe pas de différence nette entre ces deux emplois du mot « compréhension », et donc admettre que la connaissance immédiate est elle-même un concept mal défini — ce que je ne nie pas. Il me semble clair cependant que la connaissance immédiate est réellement *quelque chose*, et que ce quelque chose peut être présent ou absent, du moins à un certain degré. Si l'on admet que la connaissance immédiate *est* quelque chose, il semble alors naturel de reconnaître également que ce « quelque chose » est indissociable de toute compréhension véritable. Dès lors, conformément au point de vue $\mathcal{A}$, ce « quelque chose » qu'est la connaissance immédiate pourrait s'inscrire dans le cadre d'une activité de calcul pur et simple.

Il m'apparaît également évident que l'on doit réserver le mot « intelligence » aux seules activités mettant en jeu un certain degré de compréhension. Ici encore, certains partisans de l'IA affirment que leurs robots n'ont pas besoin de « comprendre » pour être « intelligents ». Et, bien que l'expression « intelligence artificielle » signifie la possibilité d'une activité numérique intelligente, certains affirment que l'IA ne cherche pas à créer une compréhension authentique — ni donc une connaissance immédiate. Selon moi, une « intelligence » sans réelle compréhension est une pure fumisterie. Certes on peut parfois, jusqu'à un certain degré, simuler une véritable intelligence sans qu'il y ait de réelle compréhension. (De fait, il n'est pas rare de se laisser abuser par des êtres *humains* qui parviennent à simuler la compréhension d'une situation, alors qu'il apparaît finalement qu'ils ne comprennent rien !) L'existence d'une frontière bien nette entre l'intelligence authentique (ou la compréhension authentique) et une simulation reposant totalement sur le calcul jouera

un rôle important dans les discussions qui vont suivre. Selon ma propre terminologie, la possession d'une intelligence *authentique* exige la présence d'une compréhension authentique. Ainsi, l'utilisation que je fais du terme « intelligence » (en particulier lorsque je la qualifie d'« authentique ») implique la présence d'une réelle connaissance immédiate.

Cela me semble une terminologie naturelle. Mais nombre de partisans de l'IA[16] (en tout cas ceux qui *n'adhèrent pas* au point de vue $\mathcal{A}$) nient fermement qu'ils tentent de créer une « connaissance immédiate » artificielle, même si, comme semble l'indiquer l'expression, ils tentent en fait de construire une « intelligence » artificielle. Peut-être ces gens veulent-ils dire (en accord avec le point de vue $\mathcal{B}$) qu'ils se bornent à *simuler* l'intelligence — ce qui ne met en jeu aucune compréhension ou connaissance immédiate *réelles* — sans chercher à créer ce que j'appelle l'intelligence *authentique*. Peut-être ne font-ils aucune distinction entre intelligence authentique et intelligence simulée — comme le suggère le point de vue $\mathcal{A}$. Les arguments que je vais développer tenteront notamment de montrer qu'il existe en fait une forme de « compréhension authentique » qu'aucun calcul ne peut correctement simuler, et donc qu'il y a effectivement une distinction entre l'intelligence authentique et toute simulation numérique de cette intelligence.

Bien sûr, je n'ai défini *aucun* des termes « intelligence », « compréhension » ou « connaissance immédiate ». Je pense qu'il serait peu avisé d'en donner ici des définitions *précises*. Dans une certaine mesure, il nous suffira de nous appuyer sur le sens intuitif que nous accordons à ces mots. Si la « compréhension » nous apparaît intuitivement nécessaire à l'« intelligence », tout raisonnement établissant que la « compréhension » n'est pas réductible à un calcul établira automatiquement le même résultat pour l'« intelligence ». De plus, si la « connaissance immédiate » est nécessaire à la « compréhension », alors un processus physique non réductible au calcul qui sous-tendrait le phénomène de connaissance immédiate pourrait également rendre compte de la « compréhension ». Ainsi, mon propre usage de ces termes (qui, je l'affirme, est également l'usage commun) se résume en deux énoncés :

(a) l'« intelligence » *exige* la « compréhension »

et

(b) la « compréhension » *exige* la « connaissance immédiate ».

La connaissance immédiate est pour moi un aspect — l'aspect *passif* — de la *conscience*. La conscience a également un aspect *actif*, à savoir le sentiment de *libre arbitre*. Je ne tenterai pas non plus de donner ici une définition précise de la « conscience » (et encore moins du « libre arbitre »), même si mes arguments visent à interpréter le phénomène de la conscience en termes scientifiques mais non sous forme de calcul — c'est le point de vue $\mathcal{C}$. Je ne prétends pas avoir beaucoup progressé sur cette voie, mais j'espère que les arguments exposés dans ce livre (et dans EOLP) constitueront à tout le moins des jalons utiles pour les réflexions à venir. Je considère qu'en définissant trop précisément à ce stade la notion de « conscience », nous risquerions de rendre insai-

sissable le concept même que nous voulons cerner. Ainsi, plutôt que d'en livrer une définition prématurée et inadéquate, je me limiterai à quelques commentaires sur mon propre usage de ce mot. Mais en définitive, nous devrons là aussi nous en remettre à notre appréhension intuitive de ce concept.

Je ne veux pas dire par là que nous « savons intuitivement » ce qu'*est* réellement la « conscience » ; je veux simplement dire que ce concept existe — qu'il est un authentique phénomène scientifiquement descriptible et jouant un double rôle, actif et passif, dans le monde physique. Certains pensent apparemment que ce concept est trop vague pour mériter une étude sérieuse. Pourtant ces mêmes personnes[17], fréquemment, n'hésitent pas à parler du concept d'« esprit » comme s'il était mieux défini. Selon l'acception ordinaire du mot « esprit », il existerait (et en fait existe) une chose que nous qualifions souvent d'« esprit inconscient ». À mon avis, le concept d'esprit inconscient est plus obscur, bien plus obscur même que celui d'esprit conscient. Bien que personnellement je ne me prive pas d'employer le mot « esprit », je ne cherche toutefois pas à en donner une définition précise. Le concept d'« esprit », *sauf* dans la mesure où il est déjà partiellement inclus dans celui de « conscience », ne jouera pas un rôle central dans mon raisonnement.

Qu'est-ce donc que j'entends par conscience ? Je l'ai dit, il existe deux aspects — passif et actif — de la conscience. Toutefois, la distinction entre eux n'est pas toujours évidente. La perception de la couleur rouge, par exemple, requiert certainement une conscience passive, tout comme la sensation de douleur ou le sentiment esthétique à l'égard d'une mélodie. La conscience active intervient dans l'acte volontaire de se lever, ou dans la décision délibérée de cesser une activité physique. L'évocation d'un souvenir lointain met en jeu les deux aspects, actif et passif, de la conscience. La conscience, tant active que passive, est aussi normalement présente dans la formulation de projets d'avenir et il me semble que l'activité mentale que l'on classe ordinairement sous la rubrique « compréhension » exige la présence d'une forme de conscience. En outre, nous pouvons, jusqu'à un certain degré, être (passivement) conscients, même lorsque nous sommes endormis : il nous arrive en effet de rêver (parfois même, l'aspect actif de la conscience intervient dès l'instant du réveil).

Certains contesteront qu'un seul concept de conscience puisse englober toutes ces manifestations. Celles-ci, affirment-ils, font intervenir de nombreux concepts de « conscience », très différents les uns des autres — et non simplement la conscience « active » et la conscience « passive » —, et ces diverses activités mentales correspondent à de nombreux attributs mentaux distincts. Ainsi, invoquer le terme général de « conscience » à propos de toutes ces activités relèverait, au mieux, de la banalité. Selon moi en revanche, il existe un concept unifié de « conscience » au centre de toutes ces formes d'activités mentales. Si je reconnais que l'on peut parfois distinguer les aspects passif et actif de la conscience — l'aspect passif étant lié aux sensations (aux « qualia ») et l'aspect actif à tout ce qui concerne le « libre arbitre » —, je considère qu'ils forment les deux faces d'une même médaille.

Dans la première partie de ce livre, je m'intéresserai particulièrement à ce que permet d'accomplir la qualité mentale que j'appelle « compréhension ». Je

l'ai dit, je ne tenterai pas de définir le sens de ce mot ; j'espère cependant qu'il sera suffisamment clair pour que le lecteur soit persuadé que cette qualité — quelle qu'elle soit — joue un rôle réel dans l'activité mentale nécessaire à l'acceptation des arguments exposés à la section 2.5. Je montrerai que le jugement que l'on peut porter sur *ces* arguments repose nécessairement sur l'existence d'un processus qui ne doit rien au calcul. Si mon raisonnement ne concerne pas aussi directement les concepts d'« intelligence », de « connaissance immédiate », de « conscience » ou d'« esprit », il entretient cependant avec eux un rapport évident dans la mesure où la terminologie « de sens commun » que j'ai évoquée plus haut fait de la connaissance immédiate un élément primordial de la compréhension humaine et que cette compréhension est indissociable de toute intelligence authentique.

1.13 L'argumentation de John Searle

Avant d'exposer mon propre raisonnement, il me faut mentionner brièvement une argumentation, très différente tant par sa nature que par ses motivations, à savoir la célèbre « chambre chinoise » du philosophe John Searle[18]. Elle porte elle aussi sur le problème de la « compréhension » et examine si un ordinateur suffisamment sophistiqué peut manifester une telle qualité mentale. Je me bornerai ici à décrire les grandes lignes du raisonnement de Searle.

Il s'agit d'un ordinateur censé simuler la « compréhension » en répondant à des questions posées sur une histoire qu'on vient de lui raconter. Questions et réponses sont toutes formulées en chinois. Searle envisage alors un sujet humain, ignorant le chinois et manipulant laborieusement des compteurs de manière à reproduire en détail tous les calculs effectués par l'ordinateur. Or si, par ses réponses, l'ordinateur donne l'impression de comprendre les questions posées, l'être humain qui effectue les manipulations ne comprend, lui, absolument rien de l'histoire racontée. Searle affirme donc que la compréhension ne se réduit pas à une procédure de calcul — car le sujet humain (ignorant la langue chinoise) exécute chacune des opérations accomplies par l'ordinateur, mais n'a, quelle que soit l'histoire racontée, aucun sentiment de compréhension. Searle admet que l'on puisse *simuler* les réponses témoignant d'une compréhension — c'est le point de vue $\mathcal{B}$ — et que cette simulation puisse s'obtenir avec un ordinateur reproduisant chacun des processus physiques (quels qu'ils soient) survenant dans le cerveau d'un homme en train de comprendre quelque chose. Mais, souligne-t-il, sa chambre chinoise démontre qu'une *simulation* ne peut, en elle-même, « éprouver » la moindre compréhension. Ainsi, aucune simulation numérique ne peut donner naissance à une *authentique* compréhension.

L'argumentation de Searle s'oppose directement au point de vue $\mathcal{A}$ (selon lequel toute « simulation » de la compréhension équivaut à une compréhen-

sion « réelle ») et vise à conforter le point de vue $\mathscr{B}$ (bien qu'il conforte également $\mathscr{C}$ ou $\mathscr{D}$). S'il concerne les aspects *passif, interne* ou *subjectif* de la compréhension, il ne nie pas cependant la possibilité d'une simulation de ses aspects *actif, externe* ou *objectif*. Searle l'a d'ailleurs dit : « Évidemment, le cerveau est un ordinateur numérique : puisque tout est ordinateur numérique, le cerveau l'est aussi. »[19] Il semble donc qu'il accepte la possibilité d'une simulation parfaite de l'activité d'un cerveau conscient s'efforçant de « comprendre », autrement dit, qu'il admette que les manifestations externes de cette simulation soient identiques à celles d'un être humain conscient — ce qui est le point de vue $\mathscr{B}$. Mes propres arguments, en revanche, sont dirigés contre ces manifestations externes de la « compréhension » : j'affirme qu'aucune simulation numérique ne peut traduire correctement ces manifestations. Je n'examinerai pas en détail l'argumentation de Searle, car elle ne conforte pas directement le point de vue $\mathscr{C}$ (dont je cherche justement à établir la pertinence). Je tiens cependant à préciser que si je ne la considère pas comme totalement concluante, elle me semble constituer une objection sérieuse au point de vue $\mathscr{A}$. Pour plus de détails et divers contre-arguments, voir Searle (1980) et Hofstadter et Dennett (1981) ; voir aussi Dennett (1990) et Searle (1992). Pour connaître mon propre jugement, voir EOLP, p. 18-25.

1.14 Quelques difficultés soulevées par le modèle numérique

Avant d'analyser plus avant ce qui sépare $\mathscr{C}$ de $\mathscr{A}$ et de $\mathscr{B}$, je vais examiner quelques autres difficultés auxquelles se heurte toute interprétation de la conscience inspirée par le point de vue $\mathscr{A}$. Selon $\mathscr{A}$, la simple « exécution » d'algorithmes appropriés suscite la conscience. Mais qu'est-ce que cela signifie exactement ? Le terme « exécution » signifie-t-il que l'on doit manipuler de petits objets matériels conformément aux opérations successives de l'algorithme ? Imaginons que l'on écrive ligne après ligne toutes ces opérations dans un gros livre[20]. L'acte d'écriture ou d'impression de ces lignes constitue-t-il une « exécution » de l'algorithme ? La simple existence statique du livre suffit-elle ? Et que se passe-t-il si l'on fait simplement glisser le doigt le long des lignes, l'une après l'autre ? Cela compte-t-il comme une « exécution » ? Et qu'en est-il du déplacement du doigt sur les symboles, si ceux-ci sont écrits en braille ? Ou de la projection des pages successives du livre sur un écran ? Le simple *exposé* des opérations successives d'un algorithme constitue-t-il une exécution de ce dernier ? Ou faut-il en revanche qu'une personne vérifie l'enchaînement correct de chaque ligne avec la précédente, selon les règles de l'algorithme en question ? J'imagine que *cela* du moins mettrait un terme au problème, car ce processus ne nécessiterait aucune réflexion (consciente) de la

part de cette personne. On le voit, la définition des actions physiques censées exécuter un algorithme est loin d'être évidente. Peut-être ces actions ne sont-elles aucunement nécessaires et le point de vue $\mathcal{A}$ admet-il que la simple existence mathématique platonicienne de l'algorithme (*cf.* §1.17) suffit pour que sa « connaissance immédiate » soit présente.

Quoi qu'il en soit, $\mathcal{A}$ n'estime probablement pas que l'exécution de *n'importe quel* algorithme complexe puisse susciter une connaissance immédiate (appréciable). Pour parvenir à ce stade, il semble que l'algorithme doive posséder quelques caractéristiques particulières telles par exemple qu'un « haut degré d'organisation », l'« universalité », l'« autoréférence », la « simplicité/complexité algorithmique »[21]. En outre, il y a un problème délicat : quelles sont les qualités particulières d'un algorithme qui sont responsables des divers « qualia » constituant la connaissance immédiate ? Par exemple, quel type de calcul provoque une sensation de « rouge » ? Quels calculs déclenchent une sensation de « douleur », de « saveur sucrée », d'« harmonie », d'« âcreté », etc. ? Les tentatives effectuées par les partisans de $\mathcal{A}$ pour répondre à ces questions (*cf.* par exemple Dennett 1991) ne m'apparaissent guère convaincantes.

De plus, tous les algorithmes précis et raisonnablement simples conçus dans ce but (tels ceux que l'on trouve actuellement dans la littérature) présentent l'inconvénient de pouvoir être exécutés sans grande difficulté par un ordinateur actuel. Or l'exécution de ces algorithmes devrait susciter *réellement* le qualium concerné. Aucun partisan du point de vue $\mathcal{A}$, même parmi les plus déterminés, ne peut sérieusement admettre que ces algorithmes — exécutés sur un ordinateur actuel et reposant sur les connaissances actuelles en IA — éprouvent *réellement* la moindre expérience sensorielle. Cela me semble donc obliger les partisans de ces théories à reconnaître que c'est uniquement la *complexité* des calculs mis en jeu dans l'activité de notre cerveau qui nous permet d'avoir des expériences mentales appréciables.

Cela soulève d'autres problèmes que je n'ai pas vu véritablement abordés. Si l'on croit que l'immense complexité du « câblage » constituant le réseau de synapses et de neurones cérébraux est la condition préalable fondamentale à l'existence de notre activité mentale consciente, on doit alors expliquer pourquoi l'activité consciente n'est pas uniformément répartie dans le cerveau. Lorsqu'on utilise le mot « cerveau » sans autre précision, on pense spontanément (c'est du moins le cas des non-spécialistes) aux grandes régions périphériques convolutées composant ce que l'on appelle le *cortex cérébral* — la substance grise qui forme la couche superficielle du *cerveau* proprement dit. Or si le cortex cérébral contient environ cent mille millions (10^{11}) de neurones — ce qui autorise effectivement une énorme complexité —, l'encéphale est loin de se réduire au cortex cérébral. Sa base arrière est constituée d'un autre enchevêtrement de neurones appelé *cervelet* (Fig. 1.6). Le cervelet joue un rôle décisif dans la qualité du contrôle moteur et intervient lorsqu'une aptitude motrice a été maîtrisée — pour devenir une « seconde nature » et cesser d'être une activité à laquelle l'on doit penser consciemment. Au début, l'acquisition d'une nouvelle aptitude met en jeu le contrôle conscient des mouvements,

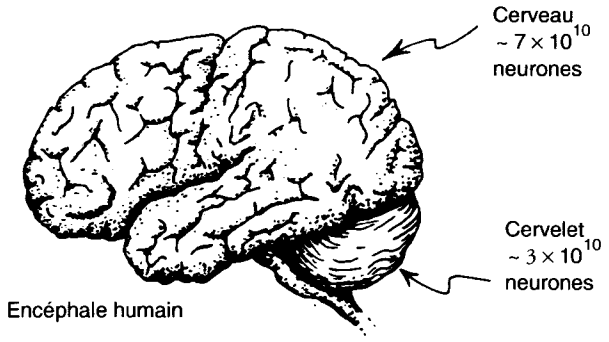


Figure 1.6. Le cervelet et le cerveau contiennent des nombres proches de neurones et de connexions neuronales. Cette simple constatation incite à se demander pourquoi l'activité du cervelet est entièrement inconsciente.

essentiellement par l'intermédiaire du cortex cérébral. Par la suite pourtant, lorsque ces mouvements sont devenus « automatiques », l'activité inconsciente du cervelet est largement prépondérante. Il est cependant remarquable, si l'on songe que l'activité du cervelet semble entièrement inconsciente, que cet organe contienne un nombre de neurones pouvant atteindre la moitié du nombre de neurones présents dans le cerveau. En outre, le cervelet comprend les cellules de Purkinje — dont nous avons parlé à la section 1.2 ; celles-ci sont des neurones possédant jusqu'à 80 000 connexions synaptiques, de sorte que le nombre total de connexions neuronales est du même ordre de grandeur dans le cervelet et dans le cerveau. Si donc la complexité du réseau neuronal est la condition préalable à la conscience, on doit se demander pourquoi cette dernière semble être entièrement absente de l'activité du cervelet. (Je ferai quelques commentaires sur ce point à la section 8.6.)

Bien entendu, les points de vue $\mathcal{B}$ et $\mathcal{C}$ rencontrent eux aussi des difficultés analogues à celles que je viens d'évoquer pour $\mathcal{A}$. Tout point de vue scientifique, quel qu'il soit, doit de toute façon expliquer ce qui sous-tend la conscience et comment surviennent les qualia. Dans les dernières sections de la deuxième partie, je tenterai d'esquisser une interprétation de la conscience du point de vue de $\mathcal{C}$.

1.15 Les insuffisances actuelles de l'IA favorisent-elles le point de vue $\mathcal{C}$?

Mais pourquoi $\mathcal{C}$? Y a-t-il un témoignage qui conforte directement ce point de vue ? $\mathcal{C}$ est-il une alternative sérieuse à $\mathcal{A}$ ou $\mathcal{B}$, voire à $\mathcal{D}$? Pour répondre à ces questions, je vais examiner ce que nous faisons réellement avec notre cerveau (ou notre esprit) lorsque nous nous adonnons à des réflexions

conscientes — et je tenterai de convaincre le lecteur que (parfois, du moins) ce que nous faisons avec notre pensée consciente diffère grandement de tout ce que l'on peut accomplir au moyen d'un calcul. Rappelons que les partisans de $\mathscr{A}$ — comme ceux de $\mathscr{B}$ si l'on se cantonne aux manifestations extérieures de la conscience — soutiennent que « tout est calcul » (sous une forme ou une autre) et que les partisans de $\mathscr{D}$, bien que reconnaissant avec $\mathscr{C}$ que les actions conscientes échappent à tout calcul, nient cependant la possibilité d'une explication scientifique de la conscience. Ainsi, pour conforter $\mathscr{C}$, il nous faut trouver des exemples d'activité mentale irréductible à toute forme de calcul, puis voir comment une telle activité peut résulter de processus physiques appropriés. C'est à ce premier objectif que va être consacré le reste de la première partie. Je présenterai dans la deuxième partie mes réflexions concernant le second objectif.

Quel type d'activité mentale est donc irréductible au calcul ? Pour tenter de répondre à cette question, on peut examiner la situation actuelle de l'intelligence artificielle et voir de quoi sont capables les systèmes commandés numériquement et de quoi ils sont incapables. Bien sûr, la situation actuelle de l'IA n'est pas forcément une bonne indication de ce qu'il sera finalement possible, en principe, de réaliser. Même dans seulement cinquante ans, les choses risquent d'être fort différentes de ce qu'elles sont aujourd'hui. Durant les seules cinquante *dernières* années, le développement des ordinateurs et de leur champ d'application a été extraordinairement rapide. Tout cela connaîtra certainement d'énormes progrès — peut-être plus rapides encore que nous ne pouvons l'imaginer. Dans ce livre, je m'intéresserai non pas à la vitesse à laquelle ces progrès pourront survenir, mais à certaines limitations *de principe*, fondamentales, auxquelles ils sont soumis. Ces limitations valant quel que soit le nombre de siècles sur lesquels portent nos spéculations, nous devons donc fonder nos arguments sur des principes généraux, sans nous laisser influencer outre mesure par les succès obtenus à ce jour. Toutefois, bien qu'il n'existe guère, aujourd'hui encore, d'intelligence artificielle que l'on puisse considérer comme véritablement convaincante — ce que même les plus résolus des partisans de l'IA admettent d'ailleurs volontiers —, les succès et les échecs de l'intelligence artificielle actuelle fournissent des indices permettant d'orienter nos réflexions.

Curieusement, les principaux échecs de l'intelligence artificielle ne concernent pas tant les domaines où le pouvoir de l'intellect humain apparaît parfois très impressionnant — comme par exemple lorsque certains spécialistes humains nous sidèrent par leur savoir ou leur aptitude à émettre des jugements reposant sur des procédures de calcul extrêmement complexes —, que ceux liés aux activités de « bon sens » auxquelles s'adonnent, durant la majeure partie de leur existence éveillée, les plus humbles d'entre nous. Actuellement, aucun robot commandé numériquement ne rivalise même avec un jeune enfant accomplissant les plus simples des activité quotidiennes, telles la prise de conscience qu'un crayon de couleur se trouvant sur le sol à l'autre bout de la pièce est l'objet qu'il lui faut pour achever un dessin, puis l'utilisation de ce crayon à cette fin. Dans ce domaine, même les aptitudes d'une fourmi accom-

plissant ses tâches quotidiennes surpassent de loin ce que peut réaliser le plus sophistiqué des systèmes de contrôle numérique actuels. En revanche, le développement de puissants ordinateurs jouant aux échecs offre un exemple frappant de situation où ces machines *peuvent* faire preuve d'une terrible efficacité. Le jeu d'échecs est indéniablement une activité où le pouvoir de l'intellect humain est particulièrement manifeste — bien qu'exploité à merveille par quelques-uns seulement d'entre nous —, mais à ce jeu, les ordinateurs actuels battent régulièrement la plupart des joueurs humains. Même les grands maîtres sont mis à rude épreuve et ne conserveront probablement pas longtemps la supériorité qu'ils possèdent encore sur les meilleurs ordinateurs[22]. Il existe également d'autres domaines d'activités intellectuellement exigeantes où les ordinateurs concurrencent — complètement ou partiellement — les spécialistes humains. En outre, dans certains de ces domaines, tels le calcul numérique pur, les aptitudes des ordinateurs surpassent de loin celles des humains.

Dans tous ces contextes cependant, on peut difficilement soutenir que l'ordinateur *comprenne* réellement ce qu'il fait. Dans le cas d'une organisation descendante, la raison de son succès n'est pas qu'il comprend tout, mais que les programmeurs humains ont utilisé leur compréhension (ou celle des spécialistes humains auxquels ils ont fait appel) pour élaborer le programme. Une organisation ascendante, quant à elle, ne semble exiger aucune compréhension particulière, ni de la part du dispositif lui-même, ni de celle des programmeurs. Nous ne parlons pas, bien entendu, de la compréhension humaine bien présente dans la conception d'algorithmes capables d'améliorer les performances du système et dans l'idée même qu'un système peut améliorer ses performances avec l'expérience, dès qu'on y intègre une rétroaction appropriée. Bien sûr, le mot « compréhension » n'admettant pas de définition précise, certains peuvent soutenir que ces ordinateurs possèdent réellement une forme de « compréhension ».

Cette position est-elle justifiée ? Pour illustrer l'absence de toute véritable compréhension chez les ordinateurs actuels, considérons l'exemple (extrait de Jane Seymore et David Norwood 1993) représenté à la figure 1.7. Cette figure (de William Hartston) montre un échiquier sur lequel les noirs, avec deux tours et un fou, sont dans une position très favorable. Toutefois, les blancs peuvent facilement éviter la défaite, simplement en déplaçant sans cesse le roi de leur côté de l'échiquier. La rangée de pions blancs constituant une défense imprenable, le roi blanc est à l'abri des tours et du fou noirs. Cela est tout à fait évident pour un joueur humain relativement familier des échecs. Pourtant, lorsque cette configuration, avec « aux blancs de jouer », fut présentée à *Deep Thought* — qui était à l'époque le plus puissant ordinateur jouant aux échecs et avait à son palmarès un certain nombre de victoires sur des grands maîtres humains —, il prit la tour noire, rompant ainsi sa rangée de pions et se retrouvant dans une situation désespérée !

Comment un joueur d'échecs aussi chevronné a-t-il pu exécuter un coup aussi stupide ? La réponse est qu'en dehors de son immense « savoir livresque », *Deep Thought* a été uniquement programmé pour calculer les suites de coups possibles jusqu'à un nombre extrêmement élevé et tenter à partir

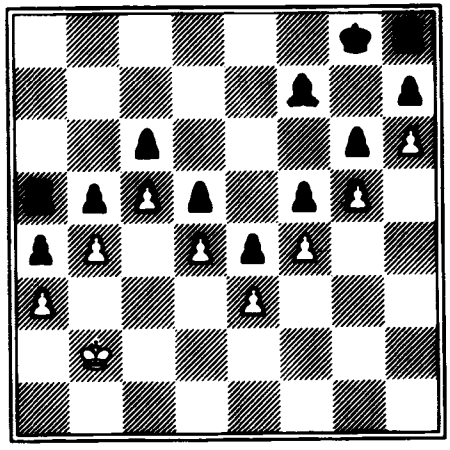


Figure 1.7. Les blancs jouent et obtiennent le nul. Évident pour les humains — mais *Deep Thought* a pris la tour !

de là d'améliorer sa position. À aucun moment il n'a réellement compris l'importance stratégique de sa rangée de pions — ni, de fait, n'a eu la moindre compréhension de son geste, à quelque niveau que ce fût.

Quiconque connaît suffisamment le principe général sur lequel *Deep Thought* et d'autres programmes d'échecs sont construits n'est pas vraiment surpris par leur réaction peu inspirée face à des configurations analogues à celle de la figure 1.7. Non seulement *nous* pouvons comprendre aux échecs des choses qui échappent à *Deep Thought,* mais nous pouvons également comprendre des choses sur les procédures (descendantes) qui ont présidé à la conception de *Deep Thought.* Nous comprenons ainsi pourquoi il ne pouvait que commettre une telle gaffe — tout comme nous comprenons pourquoi il se montrait si efficace dans la plupart des autres situations. On peut toutefois se demander si *Deep Thought* — ou tout autre système d'IA — ne parviendra pas *un jour* à nous égaler sur le plan de la compréhension — que ce soit au jeu d'échecs ou dans un autre domaine. Selon certains partisans de l'IA, un tel système parviendra à une compréhension « réelle » lorsque son programme reposera, à un niveau plus fondamental encore que celui que l'on trouve chez les ordinateurs jouant aux échecs, sur des procédures *ascendantes.* Ainsi la « compréhension », chez un tel système, émergerait progressivement, par accumulation d'« expériences », et non plus par la possession de règles algorithmiques descendantes spécifiques. Les règles descendantes qui sont suffisamment simples pour que nous les comprenions facilement ne peuvent, par elles-mêmes, servir de fondement algorithmique à une compréhension réelle : nous pouvons en effet utiliser notre propre compréhension de ces règles pour prendre conscience de leurs limitations fondamentales.

Ce point sera développé plus amplement à travers les arguments exposés aux chapitres 2 et 3. Mais qu'en est-il des procédures de calcul ascendantes ? Peuvent-elles constituer le fondement de la compréhension ? Au chapitre 3, je

montrerai qu'il n'en est rien. Notons simplement pour l'instant que les ordinateurs à stratégie ascendante actuels ne sont aucunement un substitut à une véritable compréhension humaine, dans toute activité intellectuelle où la compréhension et l'intuition humaines semblent jouer un rôle important. Je suis sûr que cela est largement reconnu de nos jours. La plupart des ambitions initiales[23] formulées par les adeptes de l'intelligence artificielle et les fabricants de systèmes experts ne se sont toujours pas concrétisées.

Mais eu égard à ses objectifs, l'IA en est encore à ses débuts. Ses adeptes (qu'ils adhèrent aux points de vue $\mathscr{A}$ ou $\mathscr{B}$) soutiennent que l'apparition d'une compréhension appréciable dans le comportement de leurs systèmes commandés par ordinateur n'est qu'une question de temps, qu'elle dépend des progrès significatifs que ne manquera pas d'accomplir leur discipline. Je tenterai plus loin de démontrer avec rigueur que ce n'est pas le cas, que tout système purement numérique, qu'il soit descendant ou ascendant, est soumis à des limitations fondamentales. Si un tel système construit avec suffisamment d'intelligence peut, durant un temps considérable, donner (comme *Deep Thought*) l'illusion de comprendre ce qu'il fait, j'affirme que l'absence de compréhension générale chez ce système doit finir — du moins en principe — par se révéler d'elle-même.

Ma démonstration visera à établir que la compréhension des *mathématiques* est irréductible au calcul. Cela surprendra peut-être certains partisans de l'IA qui font remarquer[24] que ce qui est venu en dernier dans l'évolution humaine — comme l'exécution de calculs algébriques ou arithmétiques — est ce que les ordinateurs font le mieux, dépassant déjà très largement les êtres humains ; tandis que la plupart des aptitudes tôt survenues dans l'évolution — comme la marche ou l'interprétation de scènes visuelles complexes — sont des tâches que nous accomplissons sans effort mais pour lesquelles les ordinateurs actuels ont des compétences extraordinairement limitées. Mon point de vue est très différent : *si* elle s'exprime par des règles de calcul bien définies, une activité complexe — qu'il s'agisse de calculs mathématiques, de parties d'échecs ou d'actions ordinaires — est une activité à laquelle excellent les ordinateurs modernes ; mais la compréhension même qui sous-tend ces règles de calcul est quelque chose qui échappe au calcul.

1.16 Une argumentation inspirée par le théorème de Gödel

Comment être sûr qu'une telle compréhension n'est pas réductible à des règles de calcul ? Je donnerai brièvement (aux chapitres 2 et 3) quelques très bonnes raisons de croire qu'aucune procédure de calcul — qu'elle soit ascendante, descendante ou mixte — ne peut simuler correctement les manifestations de la compréhension (ou de certaines formes de compréhension),

autrement dit, que la « compréhension » est une faculté qui s'acquiert par une activité non algorithmique du cerveau ou de l'esprit. Rappelons-le (*cf.* §1.5, §1.9), « non algorithmique » signifie ici que cette activité échappe à toute simulation effectuée à l'aide d'un ordinateur conçu d'après les principes logiques sous-jacents à tous les dispositifs de calcul électroniques et mécaniques actuels. Cette « non-simulabilité » *n'implique pas* toutefois que cette activité se situe en dehors du champ interprétatif de la science et des mathématiques. Ce qu'elle *implique*, c'est que les points de vue $\mathcal{A}$ et $\mathcal{B}$ ne peuvent expliquer comment nous accomplissons les tâches résultant d'une activité mentale consciente.

Au niveau *logique*, rien n'interdit que le cerveau conscient (ou l'esprit conscient) puisse agir selon des lois non algorithmiques (*cf.* §1.9). Mais est-ce bien le cas ? L'argumentation que je présenterai au prochain chapitre (§2.5) démontre effectivement, à mon avis, de manière extrêmement claire la présence d'un élément non réductible au calcul dans notre pensée consciente. Cette argumentation s'appuie sur le célèbre et puissant théorème de logique mathématique que nous devons au grand logicien américain d'origine autrichienne Kurt Gödel. Je n'aurai toutefois besoin que d'une forme extrêmement simplifiée de ce théorème, n'exigeant que très peu de mathématiques (j'utiliserai également une importante idée due à Alan Turing), de sorte que tout lecteur raisonnablement attentif ne devrait pas éprouver de grandes difficultés à le comprendre. Je dois dire cependant que cette utilisation du théorème de Gödel a parfois été vigoureusement contestée[25]. Certains lecteurs ont donc pu avoir l'impression que cette argumentation avait été totalement réfutée. Je tiens à préciser que *ce n'est pas le cas.* Il est vrai que de nombreux contre-arguments ont été avancés au fil des ans. Nombre d'entre eux visaient un raisonnement ancien — privilégiant le mentalisme au détriment du physicalisme — avancé par le philosophe John Lucas (1961). Lucas avait déduit du théorème de Gödel que les facultés mentales étaient nécessairement irréductibles au calcul. (D'autres, tels Nagel et Newman (1958), avaient antérieurement émis une idée analogue.) Ma propre argumentation, bien que suivant une idée similaire, diffère passablement de celle de Lucas — et ne conforte pas nécessairement le mentalisme. Je crois qu'elle est plus apte à résister aux diverses critiques soulevées à l'encontre de celle de Lucas, et à faire ressortir les insuffisances de ces critiques.

Plus loin (aux chapitre 2 et 3), j'analyserai en détail *tous* les contre-arguments dont j'ai eu connaissance. J'espère que cette analyse permettra non seulement de rectifier quelques idées fausses apparemment répandues sur la signification du théorème de Gödel, mais aussi de pallier la brièveté manifestement regrettable de ma discussion dans EOLP. Je démontrerai que bon nombre de ces contre-arguments reposent simplement sur des malentendus ; quant aux autres — qui correspondent à des points de vue fondés méritant d'être considérés en détail —, s'ils permettent peut-être aux points de vue $\mathcal{A}$ et $\mathcal{B}$ d'éluder le raisonnement gödelien, je démontrerai toutefois qu'ils *ne constituent pas* des interprétations plausibles de ce que notre faculté de « compréhension » nous permet de réaliser, et qu'en définitive, de telles échappatoires ne sont

guère utiles à l'IA. Quiconque — qu'il soit partisan de $\mathscr{A}$ ou de $\mathscr{B}$ — maintient que toutes les manifestations externes de la pensée consciente *peuvent* être correctement simulées par voie numérique devra réfuter en détail les arguments que je vais donner.

1.17 Platonisme ou mysticisme ?

Certains avanceront cependant qu'en nous forçant apparemment à adopter les points de vue $\mathscr{C}$ ou $\mathscr{D}$, l'argumentation gödelienne a des connotations « mystiques » qui ne sont assurément pas plus acceptables qu'une quelconque échappatoire. En ce qui concerne $\mathscr{D}$ — rappelons que ce point de vue affirme l'incompétence de la science pour expliquer le problème de l'esprit —, je suis d'accord avec eux. Les raisons qui me poussent à rejeter $\mathscr{D}$ résultent de cette observation que c'est seulement grâce à la science que l'on a pu accomplir des progrès réels dans la compréhension du monde. En outre, les seuls esprits dont nous ayons une connaissance directe sont ceux qui sont intimement associés à des objets physiques particuliers — les *cerveaux* —, et les différences d'état d'esprit semblent clairement découler de différences d'états physiques du cerveau. Les états mentaux constituant l'*esprit conscient* semblent résulter eux-mêmes de certains types spécifiques d'activité cérébrale. N'étaient les aspects troublants de l'esprit conscient liés à la présence de la « connaissance immédiate », voire de notre sentiment de « libre arbitre » — aspects qui à ce jour semblent se dérober à une description physique —, rien ne nous pousserait à rechercher hors du champ standard de la science une explication de l'esprit considéré comme une caractéristique du comportement physique du cerveau.

D'un autre côté, il est clair que la science a révélé un monde plein de mystères. Plus notre compréhension scientifique s'approfondit, plus ces mystères s'épaississent. Il n'est peut-être pas inutile de remarquer que les physiciens, étant ceux qui connaissent le mieux les aspects déconcertants et énigmatiques du comportement *réel* de la matière, tendent plus que les biologistes à adopter une vision du monde qui se démarque de la mécanique classique. Au chapitre 5, je décrirai quelques-uns des aspects les plus mystérieux du comportement quantique, dont certains sont de découverte relativement récente. Nous ne pourrons probablement pas aborder le problème de l'esprit sans élargir préalablement notre concept actuel de « science », mais je ne vois aucune raison d'opérer une rupture nette avec des méthodes qui se sont avérées extraordinairement efficaces. Si, comme je le pense, l'argumentation gödelienne nous oblige à accepter une certaine forme du point de vue $\mathscr{C}$, nous devrons alors également accepter certaines de ses autres conséquences, et notamment envisager une vision *platonicienne* du monde. Selon Platon, les concepts et les vérités mathématiques résident dans un monde réel dépourvu de toute notion de localisation spatio-

temporelle. Le monde de Platon, distinct du monde physique, est un monde idéal de formes parfaites à partir duquel nous devons comprendre ce monde physique. Bien que l'univers platonicien ne se laisse pas réduire à nos constructions mentales imparfaites, notre esprit y a toutefois directement accès, grâce à une « connaissance immédiate » des formes mathématiques et à une capacité de raisonner sur ces formes. Nous verrons que si notre perception platonicienne peut à l'occasion s'aider du calcul, elle n'est pas limitée par ce dernier. C'est ce potentiel de « connaissance immédiate » des concepts mathématiques, cet accès direct au monde platonicien, qui confère à l'esprit un pouvoir supérieur à celui de tout dispositif dont l'action repose uniquement sur le calcul.

1.18 Pourquoi considérer la compréhension mathématique ?

Tous cela est bien joli, diront sans doute certains lecteurs, mais y a-t-il un lien sérieux entre les problèmes subtils des mathématiques et de la philosophie mathématique et, par exemple, la plupart des questions concernant directement l'intelligence artificielle ? De fait, nombre de philosophes et de partisans de l'IA estiment très raisonnablement que si le théorème de Gödel joue indéniablement un rôle important dans son contexte originel — la logique mathématique —, il n'a au mieux que des conséquences limitées pour l'IA ou la philosophie de l'esprit. Et il semble que l'activité mentale de l'homme soit en définitive très peu concernée par le problème qui inspira les travaux de Gödel, à savoir les fondements axiomatiques des mathématiques. À cela je répondrai qu'une grand part de l'activité mentale humaine met en jeu la conscience et la compréhension. L'utilisation que je vais faire du théorème de Gödel visera à montrer que la compréhension humaine ne peut être une activité algorithmique. Il me suffira d'ailleurs de démontrer cela dans *un* contexte particulier. Si en effet on parvient à démontrer que certaines formes de compréhension mathématique échappent à toute description algorithmique, on établira par là même que notre esprit peut accomplir *quelque chose* qui ne relève pas du calcul. Une fois cela accepté, on en conclura de manière naturelle que cette action non algorithmique se retrouve nécessairement dans d'autres aspects de l'activité mentale. La voie sera ouverte !

La formulation du théorème de Gödel qui sera donnée au chapitre 2 peut effectivement sembler n'avoir qu'un rapport très lointain avec la plupart des aspects de la conscience. De fait, la démonstration de la non-calculabilité de certaines formes de compréhension mathématique n'apparaît guère avoir de lien avec les processus intervenant, par exemple, dans la perception de la couleur rouge, tout comme il ne semble pas évident que la plupart des autres aspects de la conscience correspondent à des aspirations mathématiques. Par exemple, même les mathématiciens ne pensent pas habituellement aux mathé-

matiques lorsqu'ils rêvent ! Il semble que les chiens rêvent et soient, dans une certaine mesure, également conscients lorsqu'ils rêvent ; et je suis très tenté de croire qu'ils peuvent être conscients à d'autres moments. Mais ils ne font pas de mathématiques. La méditation mathématique est indéniablement très loin d'être la *seule* activité animale exigeant une conscience ! C'est une activité humaine très particulière et hautement spécialisée. (Certains cyniques vont même jusqu'à dire que c'est une activité propre à certains êtres humains très particuliers.) La conscience, en revanche, est probablement présente dans une bonne part de l'activité mentale, humaine comme non humaine, et certainement chez les humains non mathématiciens, tout comme chez les humains mathématiciens quand ils ne sont pas en train de faire des mathématiques (autrement dit, la plupart du temps). La réflexion mathématique est un aspect extrêmement réduit de l'activité consciente et n'est pratiquée que par une infime minorité d'êtres conscients durant une fraction limitée de leur vie consciente.

Pourquoi alors envisager d'emblée, comme je le fais, le problème de la conscience dans un contexte mathématique ? Parce que je considère que c'est seulement à l'intérieur des mathématiques que l'on peut espérer démontrer avec un minimum de rigueur qu'*une part* au moins de l'activité consciente *est* irréductible au calcul. Par sa nature même, le problème du calcul *est* un problème mathématique. On ne peut escompter apporter une « démonstration », même approchée, de la non-calculabilité d'une activité si l'on ne se tourne pas vers les mathématiques. Je vais tenter de convaincre le lecteur que quoi que nous fassions avec notre cerveau ou notre esprit lorsque nous comprenons les *mathématiques*, cela est en fait différent de tout ce que nous pouvons accomplir à l'aide d'un ordinateur ; le lecteur devrait ensuite être plus à même d'accepter l'importance de l'activité non algorithmique pour la pensée consciente en général.

Nombre d'entre vous feront toutefois remarquer que la simple exécution d'un calcul ne peut *manifestement* pas susciter, par exemple, la sensation de « rouge ». À quoi bon en rechercher une démonstration mathématique s'il est parfaitement évident que les « qualia » — les expériences subjectives — n'ont rien à voir avec le calcul ? On peut répondre à cela que cet argument d'« évidence » (pour lequel j'ai une sympathie considérable) concerne uniquement les aspects *passifs* de la conscience. Comme la chambre chinoise de Searle, il permet de contester le point de vue $\mathcal{A}$, mais ne distingue pas $\mathcal{C}$ de $\mathcal{B}$.

En outre, il me faut attaquer le point de vue fonctionnaliste (le point de vue $\mathcal{A}$), pour ainsi dire, sur son propre terrain ; car les fonctionnalistes soutiennent que tous les qualia sont *nécessairement* suscités par le simple accomplissement de calculs appropriés, quelque improbable qu'une telle conception puisse apparaître au premier abord. En effet, disent-ils, à quoi sert notre cerveau sinon à effectuer un certain type de calcul ? Peut-il être autre chose qu'une sorte de système de contrôle numérique — en l'occurrence hautement sophistiqué ? Quels que soient les « sentiments de connaissance immédiate » suscités par l'activité cérébrale, ils sont, affirment les fonctionnalistes, le

résultat de cette activité opératoire. Ils soutiennent souvent que refuser le modèle algorithmique pour *toute* forme d'activité mentale, y compris la conscience, relève du *mysticisme* (sous-entendu : la seule alternative au point de vue $\mathcal{A}$ est le point de vue $\mathcal{D}$!). Dans la deuxième partie de ce livre, je suggérerai quelques tâches dont devrait être capable un cerveau décrit scientifiquement. Je ne nie pas que certains éléments de la partie « constructive » de mon argumentation restent spéculatifs. Mais je suis persuadé que l'argument en faveur de l'existence d'une *certaine* forme d'activité non algorithmique est extrêmement convaincant. Et c'est pour montrer à quel point il est convaincant que je dois maintenant examiner la réflexion mathématique.

1.19 Quel est le lien entre le théorème de Gödel et le comportement de sens commun ?

Supposons un instant démontré le fait qu'une activité ne relevant pas du calcul sous-tende effectivement nos réflexions et décisions mathématiques conscientes. En quoi cela nous aiderait-il à comprendre les raisons limitant les compétences des robots dans des domaines qui, comme je l'ai mentionné plus haut, sont plus proches des actions élémentaires « de sens commun » que du comportement sophistiqué de spécialistes avertis ? Au premier abord, il semble que mes conclusions seront presque *inverses* de celles qui expliquent les limitations de l'intelligence artificielle — du moins ses limitations actuelles. Car je semble dire que le comportement non algorithmique se situe dans des domaines hautement sophistiqués de la compréhension mathématique et non dans le comportement de sens commun. Mais ce n'est pas ce que j'affirme. J'affirme que la « compréhension » met en jeu le même type de processus non réductible au calcul, qu'elle corresponde à une authentique intuition mathématique — par exemple, à la perception de l'infinitude de l'ensemble des entiers naturels —, à la découverte qu'un objet de forme oblongue peut servir à maintenir une fenêtre ouverte, à l'observation que la manipulation d'une corde selon des mouvements choisis permet d'attacher ou de libérer un animal, à la compréhension du sens des mots « bonheur », « combat » ou « demain », ou à la prise de conscience que lorsque Abraham Lincoln avait son pied gauche à Washington, son pied droit avait de fortes chances d'être également dans cette ville — pour citer un exemple qui s'est avéré étonnamment troublant pour un système d'IA[26] ! Ce processus indépendant de tout calcul est ce qui nous rend spontanément conscients de quelque chose. Cette connaissance immédiate nous permet de visualiser les déplacements géométriques d'un morceau de bois, ou les propriétés topologiques d'une corde, ou l'unité du corps d'Abraham Lincoln. Elle nous permet également d'avoir une forme d'accès direct aux expériences d'une autre personne, de sorte que l'on peut « savoir » ce qu'elle entend par des mots tels que « bonheur », « combat » ou

« demain », même si l'explication de ces mots risque d'être totalement inadéquate. Le « sens » des mots peut effectivement se transmettre d'une personne à une autre, non parce que l'on en donne des explications adéquates, mais parce que l'autre personne a déjà une perception directe — une « connaissance immédiate » — de ce sens, de sorte que des explications très imparfaites suffisent pour que cette personne « saisisse » le sens correct. C'est la possession d'une forme commune de « connaissance immédiate » qui permet la communication entre deux personnes. Et c'est cela qui handicape sérieusement un robot insensible commandé par ordinateur. (En fait, le *sens* même du concept de « sens » d'un mot est une chose dont nous avons spontanément une certaine idée, et on voit difficilement comment on pourrait donner une description adéquate de *ce* concept à notre robot insensible.) Ce sens des mots se transmet uniquement de personne à personne parce que les personnes ont sur les choses des expériences intérieures ou des sentiments similaires. On pourrait imaginer que ces « expériences » constituent simplement une sorte de mémoire enregistrant des événements, mémoire dont notre robot pourrait être facilement équipé. Mais j'affirme qu'il n'en est rien : il est capital que l'« être » dont nous parlons, qu'il soit humain ou robot, ait une *connaissance immédiate* de l'expérience.

Pourquoi cette « connaissance immédiate », quelle que soit sa nature, échappe-t-elle nécessairement, selon moi, à tout calcul, de sorte qu'aucun robot, commandé par un ordinateur lui-même conçu d'après les idées logiques standard d'une machine de Turing (ou d'un équivalent) — qu'elle soit organisée selon une procédure descendante ou ascendante — ne peut la posséder ni même la *simuler* ? C'est ici que l'argumentation gödelienne intervient de manière cruciale. Si nous ne savons pas grand-chose aujourd'hui sur notre « connaissance immédiate » de la couleur rouge (par exemple), *nous pouvons dire* cependant quelque chose de précis sur notre connaissance immédiate de l'infinitude des entiers naturels. C'est la « connaissance immédiate » qui permet à un enfant de « savoir » ce que l'on entend par « zéro », « un », « deux », « trois », « quatre », etc., et ce que signifie l'infinitude de cette suite, alors qu'on ne lui en a fait que des descriptions — ridiculement restreintes et n'ayant apparemment presque aucun rapport avec les entiers naturels — à l'aide de quelques oranges et de quelques bananes. Ces exemples imparfaits suffisent pour que l'enfant se fasse une idée du concept « trois » ; en outre, il peut également comprendre que ce concept appartient à une suite infinie de concepts analogues (« quatre », « cinq », « six », etc.). En un certain sens platonicien, l'enfant « sait » déjà ce que sont les entiers naturels.

Cela paraît un peu mystique, mais ne l'est pas vraiment. Il est essentiel pour les discussions qui vont suivre de distinguer entre cette forme de connaissance platonicienne et le mysticisme. Les concepts que nous « connaissons » en ce sens platonicien sont des choses qui nous sont « évidentes » — qui se réduisent à une appréhension « de sens commun » —, même si nous sommes incapables de les caractériser totalement à l'aide de règles de calcul. Nous le verrons lors des discussions de l'argumentation gödelienne, il n'existe aucun moyen de caractériser totalement les propriétés des entiers naturels à l'aide de telles règles

de calcul. Comment se fait-il néanmoins que la description de nombres en termes de pommes et de bananes permette à un enfant de savoir ce que « trois jours » signifie, et que c'est le même concept « trois » que l'on retrouve dans « trois oranges » ? Bien sûr, cette association n'est pas forcément spontanée et l'enfant peut parfois ne pas la percevoir correctement, mais là n'est pas la question. Le point essentiel est que ce genre d'association est possible. J'affirme que c'est uniquement grâce à sa propre connaissance immédiate que chacun d'entre nous peut comprendre le concept abstrait « trois » et l'appartenance de ce concept à une suite infinie de concepts correspondants — à savoir les entiers naturels.

J'affirmerai pareillement que lorsque nous visualisons les déplacements d'un morceau de bois, ou d'une corde, ou d'Abraham Lincoln, nous n'utilisons pas de règles de calcul. Il existe en fait de très performantes simulations numériques des mouvements de corps rigides tels que le morceau de bois. Ces simulations peuvent être si précises et si fiables qu'elles sont habituellement bien meilleures que la visualisation directe par l'homme. On peut de même simuler sur ordinateur les mouvements d'une corde, même si — cela surprendra peut-être le lecteur — cette simulation s'avère bien plus difficile à réaliser que celle des mouvements d'un corps rigide. (Cela tient en partie au fait que la définition de la position d'une « corde mathématique » met en jeu un nombre infini de paramètres, alors qu'un corps rigide en nécessite seulement six.) Il existe des algorithmes numériques qui déterminent le nombre de nœuds d'une corde, mais ils sont totalement différents de ceux qui décrivent des mouvements rigides (et ne sont guère efficaces). Toute simulation numérique de l'apparence extérieure d'Abraham Lincoln serait certainement encore plus difficile à réaliser. Je ne dis pas que pour ces divers objets, la visualisation humaine est « meilleure » ou « pire » qu'une simulation numérique ; je dis seulement qu'elle est d'une nature totalement *différente.*

L'un des problèmes essentiels est selon moi que la visualisation met en jeu une certaine appréciation de ce qui est visualisé ; je veux dire par là qu'elle met en jeu la *compréhension.* Pour illustrer cela, considérez un fait arithmétique élémentaire, à savoir que deux entiers naturels quelconques a et b (*i.e.* deux nombres quelconques de la suite 0, 1, 2, 3, 4, ...) vérifient la propriété

$$a \times b = b \times a.$$

Cette propriété n'a rien de trivial, car les deux membres de l'égalité ont une signification différente. À gauche, $a \times b$ se réfère à un ensemble de a groupes de b objets, tandis qu'à droite, $b \times a$ se réfère à b groupes de a objets. Dans le cas particulier où $a = 3$ et $b = 5$, $a \times b$ correspond à l'arrangement

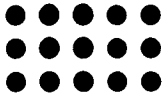
(●●●●●) (●●●●●) (●●●●●)

tandis que pour $b \times a$, on a

(●●●) (●●●) (●●●) (●●●) (●●●).

La présence dans les deux cas du même nombre total de points exprime le fait que $3 \times 5 = 5 \times 3$.

On peut voir que cette relation est vraie en considérant simplement l'arrangement suivant :



Du point de vue des lignes, on constate que l'on a trois lignes contenant chacune cinq points, ce qui exprime la quantité 3×5. En revanche, du point de vue des colonnes, on a cinq colonnes contenant chacune trois points, ce qui exprime la quantité 5×3. L'égalité de ces deux quantités se voit immédiatement par le fait que les deux cas donnent exactement le même arrangement rectangulaire ; seul diffère le point de vue de lecture. (On peut également, si l'on préfère, faire tourner mentalement ce motif de 90 degrés : on constate alors que l'arrangement représentant 5×3 a le même nombre d'éléments que l'arrangement représentant 3×5.)

L'important dans cette visualisation est qu'elle nous donne immédiatement quelque chose de bien plus général que l'égalité particulière $3 \times 5 = 5 \times 3$. Elle ne repose pas en effet sur les valeurs particulières $a = 3$ et $b = 5$. Elle s'applique pareillement si, par exemple, $a = 79\,797\,000\,222$ et $b = 50\,000\,123\,555$, et nous pouvons affirmer en toute confiance que

$$79\,797\,000\,222 \times 50\,000\,123\,555 = 50\,000\,123\,555 \times 79\,797\,000\,222$$

bien que nous n'ayons aucune chance de visualiser avec précision un arrangement rectangulaire aussi grand (aucun ordinateur actuel ne pourrait d'ailleurs en énumérer les éléments). Nous pouvons légitimement conclure que l'égalité ci-dessus est vraie — ou, en fait, que l'égalité générale $a \times b = b \times a$ est vraie* — à partir de la même visualisation que celle utilisée pour le cas particulier $3 \times 5 = 5 \times 3$. Il nous suffit de « voir confusément » dans notre esprit les nombres de lignes et de colonnes mis en jeu pour que l'égalité devienne évidente.

Je ne veux pas suggérer ici que toutes les relations mathématiques peuvent être directement perçues comme « évidentes » pour peu qu'on les visualise correctement — ni même que l'on peut toujours les comprendre en recourant à tout autre moyen que nous dicterait l'intuition. Loin de là. Certaines relations mathématiques exigent de longs raisonnements avant de pouvoir être considérées comme valides. Le but de la démonstration mathématique est justement de fournir des raisonnements dont chaque *étape* peut en fait être considérée comme évidente. Ainsi, la conclusion d'un raisonnement doit être acceptée comme *vraie*, même si, en elle-même, elle peut n'être en rien évidente.

On pourrait imaginer de dresser, une fois pour toutes, la liste de toutes les étapes de raisonnement « évidentes » possibles, de sorte que tout se réduirait

* Indiquons que cette égalité *n'est pas* vérifiée par divers types de « nombres » relativement étranges que l'on rencontre en mathématiques, par exemple les nombres ordinaux mentionnés après **Q19** à la section 2.10. Elle est cependant toujours vérifiée par les nombres qui nous intéressent ici, à savoir les *entiers naturels*.

ensuite à du calcul — à une simple manipulation mécanique de ces étapes évidentes. Mais l'argumentation gödelienne (§2.5) montre que cela est impossible. Il n'existe aucun moyen de se passer de *nouvelles* compréhensions « évidentes ». Ainsi, la compréhension mathématique ne peut se réduire à un calcul aveugle.

1.20 Visualisation mentale et réalité virtuelle

Les intuitions mathématiques évoquées à la section 1.19 avaient un caractère plutôt géométrique. Les raisonnements mathématiques peuvent s'appuyer sur de nombreux autres types d'intuitions — qui ne sont pas nécessairement géométriques. Toutefois, les intuitions qui *sont* géométriques s'avèrent souvent particulièrement précieuses pour la compréhension mathématique. Il n'est donc pas inintéressant de se demander quel type d'activité physique intervient dans le cerveau lorsque nous visualisons géométriquement un objet. Aucune exigence logique n'impose que cette activité fournisse une « image à l'identique » de l'objet visualisé. Nous allons le voir, elle peut engendrer quelque chose qui est tout à fait différent.

On peut s'aider pour cela d'une comparaison avec ce que l'on appelle la « réalité virtuelle », une discipline que l'on dit pertinente pour les questions de « visualisation ». La réalité virtuelle[27] permet par exemple de simuler sur ordinateur une structure non existante — telle le projet architectural d'un bâtiment — de sorte qu'un sujet humain observant cette simulation perçoit la structure comme si elle était « réelle ». Grâce à des mouvements des yeux ou de la tête, ou parfois même des jambes — comme s'il se promenait dans le bâtiment —, le sujet voit la structure sous des angles différents, comme si elle était effectivement réelle (Fig. 1.8). Certains[28] pensent que les processus de visualisation consciente se déroulant dans notre cerveau sont très semblables aux calculs intervenant dans l'élaboration d'une telle simulation. En fait, lorsque nous regardons avec notre propre « œil mental » une structure immobile *réelle*, il semble que nous en construisions un modèle mental fixe qui n'est pas affecté par l'instabilité des images rétiniennes due aux incessants mouvements de notre tête, de nos yeux ou de notre corps. Cette correction des mouvements corporels constituant un élément très important de la procédure générant une réalité virtuelle, on a suggéré que des processus très analogues seraient à l'œuvre lors de la construction des « modèles mentaux » constituant l'acte même de visualisation. De tels calculs, bien entendu, n'ont aucun lien géométrique réel avec la structure modélisée — ils ne sont en rien son « image à l'identique ». Les partisans du point de vue $\mathcal{A}$ considèrent naturellement que nos visualisations conscientes sont le résultat, à l'intérieur de notre cerveau, d'une telle simulation numérique du monde extérieur. Pour ma part cependant, il me semble que la *compréhension* intervenant lorsque nous percevons

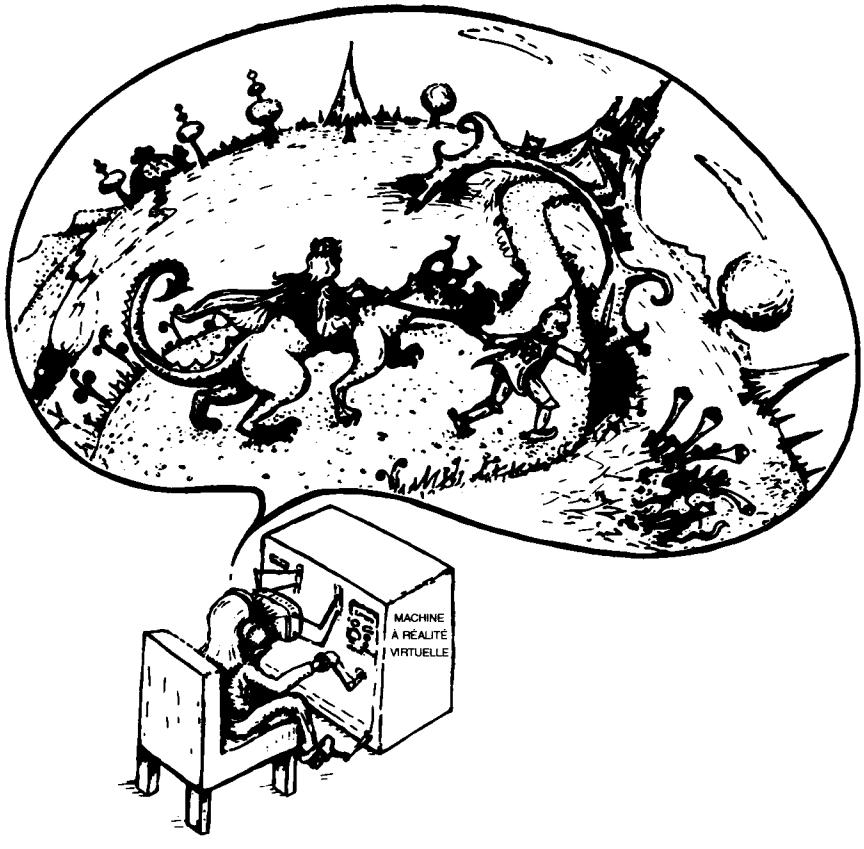


Figure 1.8. La réalité virtuelle. On peut faire apparaître sur ordinateur un monde virtuel tridimensionnel réagissant de manière cohérente aux mouvements de la tête et du corps.

de manière consciente une scène visuelle est d'une nature très différente de la modélisation du monde par une telle simulation numérique.

On peut aussi imaginer que notre cerveau contient un élément agissant davantage comme un « calculateur analogique » et que la modélisation du monde extérieur s'effectue non pas à l'aide de calculs numériques — comme c'est le cas avec les ordinateurs électroniques —, mais à l'aide d'une structure physique interne dont le comportement reflète celui du système extérieur modélisé. Une manière tout à fait directe d'obtenir un dispositif analogique simulant les déplacements d'un corps rigide consisterait à avoir, intérieurement, un petit corps physique réel ayant la forme (mais non la taille) de l'objet extérieur modélisé — quoique je ne veuille nullement suggérer ici que *ce* modèle particulier ait un rapport direct avec ce qui se passe à l'intérieur de notre cerveau ! Les mouvements de ce corps interne perçus sous des angles différents donneraient des effets externes très similaires à ceux d'un calcul numérique. Un tel dispositif pourrait également être intégré dans un système de

« réalité virtuelle » qui, au lieu d'offrir un modèle purement numérique de la structure concernée, en donnerait un modèle physique réel dont la taille serait la seule différence d'avec la « réalité » simulée. D'une manière générale, la simulation analogique ne serait pas aussi directe ni aussi triviale. On pourrait utiliser, à la place de la distance physique réelle, un paramètre — par exemple le potentiel électrique. Il faudrait simplement s'assurer que les lois physiques gouvernant la structure interne reflètent très précisément celles gouvernant la structure externe modélisée. Il ne serait pas nécessaire que la structure interne soit l'image fidèle de la structure externe.

Les dispositifs analogiques réaliseront-ils des simulations inaccessibles au pur calcul numérique ? Nous l'avons dit à la section 1.8, dans le cadre de la physique actuelle, rien ne permet de le penser. Ainsi, si nous parvenons à démontrer que notre imagination visuelle accomplit des choses non réductibles au calcul, ses fondements seront à rechercher hors du cadre de la physique actuelle.

1.21 L'imagination mathématique est-elle non algorithmique ?

Rien dans cette discussion ne dit explicitement que le processus mis en jeu lorsque nous visualisons un objet ne peut être simulé par une procédure algorithmique. Même si notre cerveau recourt à un processus analogique interne pour visualiser un objet, rien ne semble en effet interdire que nous puissions au moins *simuler* numériquement ce processus.

La « visualisation » dont j'ai parlé jusqu'ici a essentiellement concerné ce qui est « visuel » au sens littéral du terme, *i.e.* les images mentales qui semblent correspondre aux signaux transmis des yeux vers le cerveau. D'une manière plus générale cependant, nos images mentales ne sont pas toutes « visuelles ». C'est le cas, par exemple, lorsque nous comprenons le sens d'un mot abstrait ou que nous nous remémorons un morceau de musique — et les images mentales d'un aveugle de naissance n'ont certainement aucun lien direct avec les signaux émis par ses yeux. Ainsi, la « visualisation » à laquelle je fais référence concerne davantage le problème général de la « connaissance immédiate » que les phénomènes associés au système visuel. Cela dit, je ne connais aucun raisonnement portant directement sur la nature, algorithmique ou non, de notre faculté de visualisation au sens littéral du terme. Ma croyance en la nature non algorithmique du processus de visualisation repose sur le fait qu'il *semble* que l'on puisse démontrer que *d'autres* formes de connaissance immédiate humaine échappent à tout calcul. Si l'on voit difficilement comment démontrer directement la non-calculabilité propre à la visualisation géométrique, une démonstration que *certaines* formes de connaissance immédiate consciente ne sont pas algorithmiques suggérerait fortement que la connaissance immédiate

impliquée dans la visualisation géométrique ne l'est pas non plus. Il ne semble en effet y avoir aucune raison de croire à une frontière étanche entre les différentes manifestations de la compréhension consciente.

Soyons plus précis. La connaissance immédiate dont j'affirme que *l'on peut démontrer* qu'elle échappe à tout calcul est celle qui intervient dans notre compréhension des propriétés des entiers naturels 0, 1, 2, 3, 4, ... (On pourrait même dire que notre concept d'entier naturel *est*, en un sens, une forme de « visualisation » *non* géométrique.) Nous verrons à la section 2.5, en appliquant une version très accessible du théorème de Gödel (*cf.* la réponse à la question **Q16**), que cette compréhension ne peut s'exprimer par un ensemble fini de règles — d'où il découle qu'on ne peut la simuler par une procédure de calcul. De temps en temps, on entend dire que tel ordinateur a « été éduqué » pour « comprendre » le concept d'entier naturel[29]. Nous le verrons également, cela est totalement faux. C'est notre *connaissance immédiate* de la signification réelle d'un « nombre » qui nous permet de saisir le concept correct. Et c'est seulement lorsque nous disposons de ce concept correct que nous pouvons — du moins en principe — répondre correctement aux questions que l'on peut nous poser sur les nombres, ce qu'aucun ensemble fini de règles ne peut faire. Avec les seules règles et sans aucune connaissance immédiate directe, un robot commandé par ordinateur (comme *Deep Thought* ; *cf.* §1.15) est nécessairement limité sur des plans où nous-mêmes ne sommes pas limités. Bien sûr, si l'on donne au robot des règles suffisamment intelligentes, il peut accomplir des exploits prodigieux, dont certains dépassent de loin les capacités de l'homme dans des domaines spécifiques extrêmement délimités. Durant un temps, il peut même nous abuser en nous faisant croire qu'il possède également une connaissance immédiate.

Soulignons encore une fois (*cf.* §1.15) que lorsqu'un ordinateur numérique (ou analogique) *parvient* à une simulation parfaite d'un système extérieur, c'est presque toujours parce que l'on a pu tirer avantage d'une compréhension humaine appréciable des idées mathématiques sous-jacentes à cette simulation. Considérez par exemple la simulation numérique des déplacements géométriques d'un corps rigide. Les calculs nécessaires à cette simulation dépendent tout particulièrement des idées émises par certains grands penseurs du XVII^e^ siècle, tels les mathématiciens français Descartes, Fermat et Desargues, qui ont introduit les concepts de coordonnées et de géométrie projective. En ce qui concerne la simulation des mouvements d'une corde, les idées géométriques permettant de comprendre les contraintes affectant son mouvement — *i.e.* sa « nodosité » — sont très sophistiquées ; elles sont en outre remarquablement récentes, nombre de progrès fondamentaux n'ayant été accomplis qu'au cours du présent siècle. Tandis qu'il est parfois relativement facile en pratique de trouver, par un simple jeu de mains et en faisant marcher son bon sens, si une boucle fermée mais enchevêtrée contient ou non des nœuds, les algorithmes numériques qui permettent de résoudre ce problème sont étonnamment complexes, sophistiqués et inefficaces.

Les simulations numériques de ce type de problème ont jusqu'à présent surtout correspondu à des organisations descendantes, dépendant considérable-

ment de la compréhension et de l'intuition humaines. Or, il y a peu de chances que la visualisation d'un objet par un cerveau humain repose sur une procédure de ce type. Il vaudrait donc mieux élaborer des programmes comportant une forte dose de stratégie ascendante, de sorte que les « images visuelles » simulées surviendraient seulement après une « période d'apprentissage » considérable. J'ignore cependant s'il existe des approches ascendantes performantes (*e.g.* basées sur des réseaux de neurones formels) pour ce type de problème. Je pense d'ailleurs qu'une approche reposant *entièrement* sur une organisation ascendante donnerait de médiocres résultats. Je conçois en effet difficilement que l'on puisse obtenir une bonne simulation des déplacements géométriques d'un corps rigide ou des restrictions topologiques du mouvement d'une corde — *i.e.* de sa *nodosité* — sans faire intervenir une authentique compréhension de ce qui se passe réellement.

Quel est le processus physique responsable de notre connaissance immédiate — cette connaissance immédiate qui semble nécessaire à toute compréhension authentique ? Se dérobe-t-il à toute simulation numérique, comme le prétend le point de vue $\mathcal{C}$? Est-il accessible à notre compréhension, du moins en principe ? Je crois qu'il en est ainsi, et que le point de vue $\mathcal{C}$ est une véritable possibilité scientifique, pouvant toutefois entraîner des modifications subtiles mais importantes de nos critères et méthodes scientifiques. Nous devrons être à l'affût d'indices pouvant surgir de manière imprévue, et dans des domaines où la compréhension authentique semble *a priori* ne jouer aucun rôle. Pour les discussions qui vont suivre, je demanderai au lecteur de garder un esprit ouvert tout en prêtant soigneusement attention aux raisonnements et témoignages scientifiques proposés, même s'ils peuvent parfois paraître contredire le sens commun. Préparez-vous à méditer un peu sur les arguments que je vais m'efforcer de présenter aussi clairement que possible. Armons-nous donc de courage — et lançons-nous dans l'aventure.

Je vais maintenant laisser de côté la physique et les facteurs biologiques suggérant la non-calculabilité affirmée par le point de vue $\mathcal{C}$. Ils feront l'objet de la deuxième partie de ce livre. Je l'ai dit, je suis convaincu que le processus de compréhension consciente met en œuvre des actions échappant à tout calcul. Il est temps de justifier mon affirmation et de rechercher de telles actions. C'est la raison pour laquelle je dois à présent me tourner vers les mathématiques.

2
Le raisonnement gödelien

2.1 Théorème de Gödel et machines de Turing

C'est en mathématiques que la pensée s'exprime sous sa forme la plus pure. Si penser consistait simplement à accomplir un certain type de calcul, la pensée mathématique devrait l'illustrer de la manière la plus nette. Curieusement pourtant, il s'avère que c'est exactement le contraire. C'est en mathématiques que nous trouvons les témoignages les plus clairs indiquant que le processus de pensée consciente contient un élément irréductible aux calculs. Cela peut paraître paradoxal — mais il sera primordial, pour les raisonnements qui vont suivre, que nous y soyons préparés.

Avant de commencer, je voudrais inciter le lecteur à ne pas se laisser intimider par les mathématiques que nous allons rencontrer dans les prochaines sections (§ 2.2-§ 2.5), même si nous allons examiner certaines conséquences de ce qui n'est pas moins que le théorème de logique mathématique le plus important de tous les temps, le célèbre théorème de Gödel. Je ne présenterai qu'une version extrêmement simplifiée de ce théorème, inspirée notamment d'idées formulées par Alan Turing. Je n'aurai besoin d'aucun formalisme mathématique autre que celui de l'arithmétique la plus élémentaire. Certes le raisonnement que je vais développer sera par endroits déroutant, mais il sera *simplement* déroutant, et non « difficile » dans le sens où il n'exigera pas de connaissances mathématiques préalables. Suivez ce raisonnement à votre rythme et n'ayez aucune honte à le reprendre autant de fois que cela vous semblera nécessaire. J'explorerai plus loin (§2.6-§2.10) quelques idées plus ardues sous-jacentes au théorème de Gödel ; le lecteur non intéressé par ce sujet pourra sauter sans inconvénient cette partie du livre.

C'est en 1930, lors d'un colloque à Königsberg, que le jeune et brillant mathématicien Kurt Gödel stupéfia un groupe d'éminents spécialistes mondiaux des mathématiques et de la logique en présentant ce qui devait devenir son fameux théorème. Celui-ci fut rapidement reconnu comme une contribution essentielle aux fondements des mathématiques — probablement la plus fondamentale qui ait jamais été produite —, mais j'affirmerai qu'en posant son théorème, Gödel a également initié un progrès majeur dans le domaine de la philosophie de l'esprit.

Parmi les résultats irréfutables établis par Gödel figure le fait qu'aucun *système formel* de règles sûres de démonstration mathématique ne permet, même en principe, de vérifier la validité de toutes les propositions de l'arithmétique ordinaire. Cela est effectivement très remarquable. Mais un raisonnement solide permet en outre d'affirmer que ce résultat démontre plus, à savoir que la compréhension et l'intuition humaines sont irréductibles à un ensemble de règles de calcul. Car ce que Gödel a en effet démontré est qu'aucun système formel ne permet de démontrer même les propositions de l'arithmétique dont la vérité est accessible, en principe, à l'intuition et à la compréhension humaines — d'où il résulte que ces deux qualités sont irréductibles à un ensemble de règles. Je vais notamment tenter ici de convaincre le lecteur que le théorème de Gödel montre effectivement cela et que l'on peut en déduire que la pensée humaine recouvre probablement un champ d'activité bien plus vaste que celui de n'importe quel ordinateur, actuel ou à venir.

Je n'aurai pas besoin de donner la définition d'un « système formel » pour développer mon raisonnement (voyez cependant la section 2.7). Il me suffira d'exploiter les travaux fondamentaux réalisés par Turing (et quelques autres, essentiellement Church et Post) dans les années 1936-1937 et qui permirent de définir les processus que nous appelons aujourd'hui « calculs » ou « algorithmes ». Il existe en effet une équivalence entre ces processus et ce que peut accomplir un système formel mathématique, de sorte qu'il importera peu de savoir exactement ce qu'est un système formel pourvu que l'on ait une idée raisonnablement claire du concept de calcul ou d'algorithme. Je n'aurai d'ailleurs même pas besoin d'une définition précise de ce dernier concept.

Les lecteurs familiers de mon précédent livre *l'Esprit, l'ordinateur et les lois de la physique* (EOLP, *cf.* chapitre 2) savent qu'un algorithme est ce que peut effectuer un ordinateur mathématiquement idéalisé appelé *machine de Turing*. Cette machine travaille en suivant pas à pas une procédure dont chaque étape est totalement déterminée d'une part par la nature d'une marque figurant sur un « ruban » que la machine examine en permanence, d'autre part par l'« état interne » (défini de manière discrète) de la machine. Ces marques et ces états internes sont en nombre fini — mais le ruban lui-même a une longueur illimitée. La machine démarre dans un état particulier (désignons-le par « 0 ») et ses instructions sont inscrites sur le ruban, par exemple sous forme d'un nombre binaire (une suite de « `0` » et de « `1` »). Elle lit alors ces instructions en déplaçant le ruban (ou, ce qui est équivalent, en se déplaçant elle-même le long du ruban) dans un sens spécifié à chaque étape par son état interne et le chiffre particulier qu'elle rencontre sur le ruban. Elle efface les marques ou en inscrit

de nouvelles en fonction de ce que lui dicte la procédure, et continue ainsi jusqu'à ce qu'elle atteigne une instruction particulière : « **STOP** ». Alors (et alors seulement), elle affiche sur le ruban la réponse du calcul qu'elle a exécuté et arrête son activité. Elle est maintenant prête pour le calcul suivant.

Certaines machines de Turing, dites *universelles*, ont la capacité d'imiter toute autre machine de Turing. Ainsi, toute machine de Turing universelle peut effectuer *tout* calcul (ou algorithme) que l'on peut lui proposer. Bien que la structure interne d'un ordinateur soit très différente de celle d'une machine de Turing universelle (et que son « espace de travail » interne, bien que très grand, ne soit pas infini comme le ruban idéalisé d'une machine de Turing), tous les ordinateurs généralistes actuels sont en fait des machines de Turing universelles.

2.2 Calculs

Je vais examiner ici la notion de calculs. Par « calcul » (ou « algorithme »), j'entends l'action d'une machine de Turing : l'exécution par un ordinateur des instructions d'un programme informatique. Notez qu'un calcul ne se réduit pas uniquement à l'accomplissement d'opérations arithmétiques ordinaires telles que l'addition ou la multiplication de nombres, mais peut également comporter, entre autres, des *opérations logiques* bien définies. Considérons par exemple le problème suivant :

(A) Trouver un nombre qui ne soit pas la somme de trois carrés.

Par « nombre », j'entends ici un « entier naturel », *i.e.* un élément de la suite

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, \ldots$$

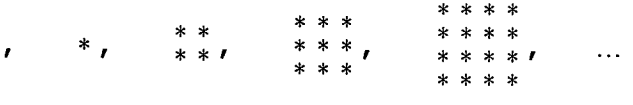
Un *carré* est le produit d'un entier naturel par lui-même, *i.e.* un élément de la suite

$$0, 1, 4, 9, 16, 25, 36, \ldots,$$

ceux-ci étant respectivement définis par

$$0 \times 0 = 0^2,\ 1 \times 1 = 1^2,\ 2 \times 2 = 2^2,\ 3 \times 3 = 3^2,\ 4 \times 4 = 4^2,$$
$$5 \times 5 = 5^2,\ 6 \times 6 = 6^2, \ldots$$

De tels nombres sont appelés « carrés » parce qu'on peut les représenter par des arrangements carrés (y compris l'arrangement vide pour représenter 0) :



Le calcul **(A)** peut alors procéder ainsi. On prend l'un après l'autre chaque entier naturel, en partant de 0, et on regarde s'il est ou non la somme de trois carrés. Il suffit de considérer les carrés qui ne sont pas supérieurs au nombre lui-même. Ainsi, pour chaque entier naturel, il n'y a qu'un nombre fini de carrés à essayer. Dès qu'il trouve trois carrés dont la somme est égale au nombre en question, le calcul passe à l'entier naturel suivant et cherche un nouveau triplet de carrés (chacun non supérieur au nombre lui-même) dont la somme soit égale à ce nombre. Le calcul ne s'arrête que lorsqu'il trouve un entier naturel qui n'est la somme d'aucun triplet de carrés. Pour voir comment cela fonctionne, partons de 0. On peut toujours écrire $0 = 0^2 + 0^2 + 0^2$, donc 0 est la somme de trois carrés. Essayons ensuite 1. $0^2 + 0^2 + 0^2$ ne marche pas, mais on a $0^2 + 0^2 + 1^2 = 1$. Le calcul passe alors à 2. On vérifie que $0^2 + 0^2 + 0^2$ et $0^2 + 0^2 + 1^2$ ne marchent pas, mais que l'on a $2 = 0^2 + 1^2 + 1^2$. On passe alors à 3 et on trouve $3 = 1^2 + 1^2 + 1^2$; pour 4, on trouve $4 = 0^2 + 0^2 + 2^2$; pour 5 on a $5 = 0^2 + 1^2 + 2^2$. Après avoir trouvé que $6 = 1^2 + 1^2 + 2^2$, on passe à 7 ; mais ici, aucun des triplets de carrés (dont chaque membre n'est pas supérieur à 7)

$$0^2 + 0^2 + 0^2, \quad 0^2 + 0^2 + 1^2, \quad 0^2 + 0^2 + 2^2, \quad 0^2 + 1^2 + 1^2, \quad 0^2 + 1^2 + 2^2,$$
$$0^2 + 2^2 + 2^2, \quad 1^2 + 1^2 + 1^2, \quad 1^2 + 1^2 + 2^2, \quad 1^2 + 2^2 + 2^2, \quad 2^2 + 2^2 + 2^2$$

ne donne une somme égale à 7. Le calcul s'arrête ici et nous sommes parvenus au résultat recherché : 7 est un nombre qui *n'est pas* la somme de trois carrés.

2.3 Des calculs qui ne s'arrêtent pas

Nous avons eu de la chance avec le problème **(A)**. Supposez en effet que nous nous soyons donné ce problème-ci :

(B) Trouver un nombre qui ne soit pas la somme de quatre carrés.

Dans ce cas, une fois arrivé à 7, on trouve qu'il *est* la somme de quatre carrés : $7 = 1^2 + 1^2 + 1^2 + 2^2$. Nous devons alors passer à 8 pour trouver $8 = 0^2 + 0^2 + 2^2 + 2^2$, puis à 9 pour trouver $9 = 0^2 + 0^2 + 0^2 + 3^2$, puis à 10 pour trouver $10 = 0^2 + 0^2 + 1^2 + 3^2$, etc. Le calcul continue : ..., $23 = 1^2 + 2^2 + 3^2 + 3^2$, $24 = 0^2 + 2^2 + 2^2 + 4^2$, ..., $359 = 1^2 + 3^2 + 5^2 + 18^2$, ..., et ainsi de suite. On peut penser que la solution du problème est fantastiquement grande et que notre ordinateur va devoir utiliser un temps et une mémoire énormes pour trouver la réponse. On peut même commencer à se demander si cette solution existe vraiment. Le calcul continue et semble ne jamais devoir s'arrêter. En fait, c'est bien le cas : il ne s'arrête jamais ! C'est là un célèbre théorème démontré pour la première fois en 1770 par le mathématicien français (d'origine italienne) Joseph Lagrange : *tout* nombre est la somme de quatre carrés. Ce n'est pas si facile à démontrer (même le grand

mathématicien suisse Leonhard Euler, contemporain de Lagrange et homme d'une originalité, d'une productivité et d'une intuition mathématique étonnantes, n'a pas réussi à en trouver une démonstration).

Je ne vais pas embêter le lecteur avec les détails de la démonstration de Lagrange. Examinons plutôt quelque chose de bien plus simple :

(C) Trouver un nombre impair qui soit la somme de deux nombres pairs.

J'espère qu'il est évident pour le lecteur que *ce* calcul ne s'arrête jamais ! La somme de nombres pairs, autrement dit de multiples de deux,

$$0, 2, 4, 6, 8, 10, 12, 14, 16, \ldots$$

donnant toujours un nombre pair, il n'existe aucun nombre impair, *i.e.* aucun nombre de la suite

$$1, 3, 5, 7, 9, 11, 13, 15, 17, \ldots$$

qui soit la somme de deux nombres pairs.

Je viens de donner deux exemples de calculs (**(B)** et **(C)**) qui ne s'arrêtent jamais. Pour le premier, ce fait, bien que vrai, n'est pas du tout facile à vérifier, tandis que pour l'autre, il est évident. Permettez-moi de donner un autre exemple :

(D) Trouver un nombre pair supérieur à 2 qui ne soit pas la somme de deux nombres premiers.

Rappelons qu'un nombre premier est un entier naturel (différent de 0 ou 1) dont les seuls diviseurs sont 1 et lui-même, autrement dit un nombre de la suite

$$2, 3, 5, 7, 11, 13, 17, 19, 23, \ldots.$$

Il est fort probable que le calcul **(D)**, lui aussi, ne se termine pas, mais personne n'en est sûr. Il dépend de la vérité de la célèbre « conjecture de Goldbach », formulée en 1742 par Christian Goldbach dans une lettre adressée à Euler, mais qui reste à ce jour non démontrée.

2.4 Comment décidons-nous que certains calculs ne s'arrêtent jamais ?

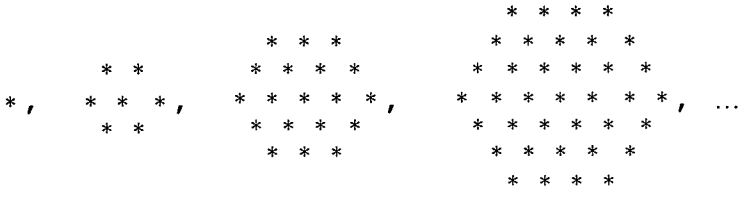
Nous savons maintenant que certains calculs peuvent ne pas s'arrêter. De plus, dans le cas où ils ne s'arrêtent effectivement pas, nous savons que cela peut être plus ou moins facile à déceler, voire même si difficile que personne n'est encore parvenu à démontrer qu'il est en bien ainsi. Quelles procédures permettent aux mathématiciens de se convaincre — ou de convaincre leurs confrères — que certains calculs ne se terminent pas ? Découlent-elles elles-

mêmes d'une procédure de calcul (d'une procédure algorithmique) ? Avant de tenter de répondre à cette question, considérons un autre exemple. Il sera légèrement plus difficile à voir que **(C)**, mais cependant bien plus facile que **(B)**. Il illustre la manière dont les mathématiciens aboutissent parfois à leurs conclusions.

Cet exemple se fonde sur les nombres dits *hexagonaux* :

$$1, 7, 19, 37, 61, 91, 127, \ldots,$$

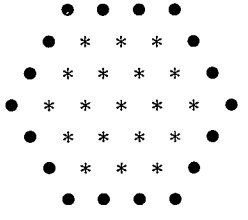
à savoir les nombres que l'on peut disposer selon des arrangements hexagonaux (*en excluant* cette fois l'arrangement vide) :



Ces nombres s'obtiennent en partant de 1 et en ajoutant les multiples successifs de 6 :

$$6, 12, 18, 24, 30, 36, \ldots,$$

comme on peut le voir en observant que chaque nombre hexagonal s'obtient à partir du précédent en l'entourant d'un anneau hexagonal



et en remarquant que le nombre de points composant cet anneau est un multiple de 6, le multiplicateur augmentant de 1 chaque fois que l'hexagone s'agrandit.

Additionnons maintenant, jusqu'à un certain point, les nombres hexagonaux successifs en partant de 1. Qu'obtient-on ?

$$1 = 1,\ 1 + 7 = 8,\ 1 + 7 + 19 = 27,\ 1 + 7 + 19 + 37 = 64,$$

$$1 + 7 + 19 + 37 + 61 = 125.$$

Quelle est la particularité des nombres 1, 8, 27, 64, 125 ? Ce sont tous des *cubes*. Un cube est le produit de trois fois le même nombre :

$$1 = 1^3 = 1 \times 1 \times 1,\ 8 = 2^3 = 2 \times 2 \times 2,\ 27 = 3^3 = 3 \times 3 \times 3,$$

$$64 = 4^3 = 4 \times 4 \times 4,\ 125 = 5^3 = 5 \times 5 \times 5, \ldots$$

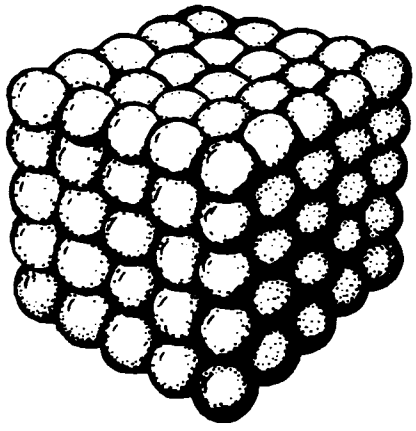


Figure 2.1. Entassement cubique de sphères.

Est-ce une propriété générale des nombres hexagonaux ? Si l'on regarde la somme suivante, on trouve effectivement

$$1 + 7 + 19 + 37 + 61 + 91 = 216 = 6 \times 6 \times 6 = 6^3.$$

Mais en est-il toujours ainsi ? Si oui, le calcul correspondant au problème suivant ne se termine jamais :

(E) Trouver une somme de nombres hexagonaux successifs, en partant de 1, qui ne soit pas un cube.

Je vais tenter de vous convaincre que ce calcul continue éternellement.

Premièrement, un cube s'appelle un cube parce qu'on peut le représenter par un arrangement de points analogue à celui de la figure 2.1. Je veux que vous essayiez d'imaginer cet arrangement construit étape par étape, en partant d'un coin auquel on ajoute une succession d'arrangements à trois faces consistant à chaque fois, comme le décrit la figure 2.2, d'un mur arrière, d'un mur latéral et d'un plafond.

Maintenant, placez-vous très loin, dans la direction du sommet commun aux trois faces, et regardez l'arrangement. Que voyez-vous ? Un *hexagone*, comme à la figure 2.3. Les points qui constituent ces hexagones de tailles successives croissantes, si on les réunit, reconstituent le cube entier. Cela démontre donc que l'addition de nombres hexagonaux successifs, en partant de 1, donne toujours un cube. Autrement dit, nous venons de démontrer que **(E)** ne s'arrête jamais.

Le lecteur pourra objecter que ce raisonnement est plutôt intuitif et n'est pas une démonstration mathématique formelle et rigoureuse. Il est en vérité parfaitement fondé, et mon objectif est notamment de montrer qu'il existe des raisonnements mathématiques légitimes qui ne sont pas « formalisés » selon des règles prédéterminées et reconnues. Un exemple bien plus élémentaire de raisonnement géométrique, utilisé pour obtenir une propriété générale des

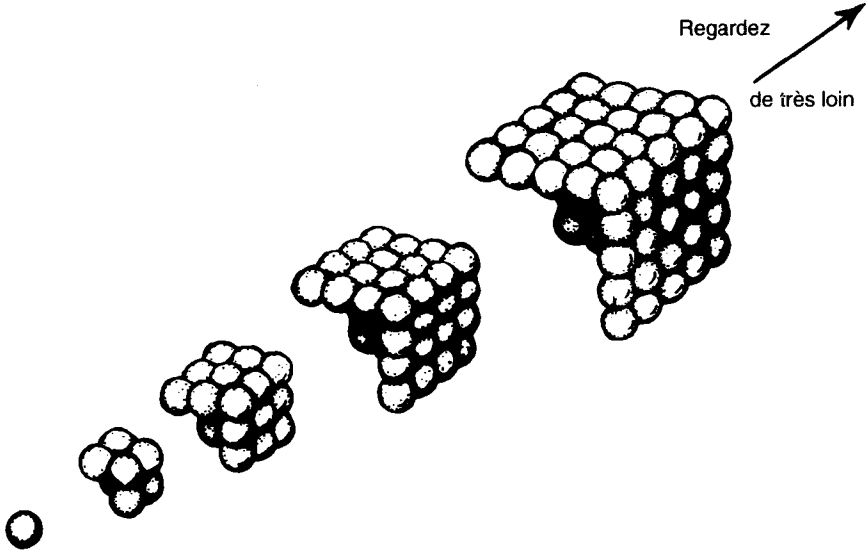


Figure 2.2. Chaque cube s'obtient en plaquant un mur arrière, un mur latéral et un plafond sur le cube précédent.

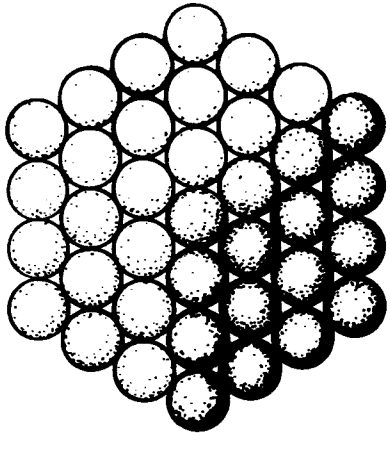


Figure 2.3. Vus de loin, le mur arrière, le mur latéral et le plafond forment un hexagone.

entiers naturels, est la démonstration de $a \times b = b \times a$ donnée à la section 1.19. Cela aussi est une « démonstration » parfaitement correcte, bien que non formelle.

On peut bien sûr, si on le désire, remplacer le précédent raisonnement sur l'addition des nombres hexagonaux successifs par une démonstration mathématique plus formelle. Une telle démonstration pourrait, par exemple, reposer sur le *principe de récurrence*, une procédure qui vérifie, à l'aide d'un seul calcul, la validité d'énoncés s'appliquant à *tous* les entiers naturels. Ce principe permet en substance de déduire qu'une proposition $P(n)$ dépendant d'un entier naturel donné n (telle « la somme des n premiers nombres hexagonaux est égale à n^3 ») vaut pour tout n, à condition que l'on puisse démontrer, premièrement, qu'elle vaut pour $n = 0$ (ou, ici, $n = 1$) et, deuxièmement, que la validité de $P(n)$ *entraîne* la validité de $P(n + 1)$. Je ne vais pas ennuyer le lecteur en lui démontrant par le détail, à l'aide du principe de récurrence, que **(E)** ne s'arrête jamais ; il pourra s'il le souhaite s'essayer à cet exercice.

Les règles précises, telles le principe de récurrence, suffisent-elles toujours pour démontrer que des calculs qui ne s'arrêtent pas ne s'arrêtent effectivement pas ? Curieusement, la réponse est « non ». Nous le verrons bientôt, c'est là une des conséquences du théorème de Gödel. Il sera donc important de tenter de comprendre ce théorème. Ce n'est pas seulement le principe de récurrence qui est insuffisant. *Tout* ensemble de règles, quel qu'il soit, est insuffisant, si, par « ensemble de règles », on entend un système de procédures formelles pour lesquelles on peut vérifier de manière purement algorithmique, quel que soit le cas particulier, si ces règles ont été ou non correctement appliquées. Cela peut paraître une conclusion pessimiste, car elle semble signifier qu'il existe des calculs qui ne s'arrêtent pas, mais dont on ne pourra jamais démontrer mathématiquement qu'ils ne s'arrêtent pas. Toutefois, ce n'est pas du tout ce que dit le théorème de Gödel. Ce qu'il dit *vraiment* peut être considéré sous un angle plus positif, à savoir que les intuitions accessibles aux mathématiciens humains — en fait, à quiconque réfléchit de manière logique en s'aidant de son intelligence et de son imagination — se dérobent à tout ce qui peut être formalisé par un ensemble de règles. Si les règles sont parfois un substitut partiel à la compréhension, elles ne peuvent jamais la remplacer totalement.

2.5 Familles de calculs ; la conclusion $\mathcal{G}$ de Gödel-Turing

Afin de voir comment le théorème de Gödel (dans la forme simplifiée que je vais lui donner en m'inspirant notamment des idées de Turing) démontre cela, nous allons généraliser légèrement la formulation des calculs que j'ai considérés jusqu'ici. Au lieu de demander si un seul calcul — tel **(A)**, **(B)**, **(C)**,

(D) ou **(E)** — se termine, nous allons considérer un calcul dépendant d'un *entier naturel n.* Si nous désignons ce calcul par $C(n)$, nous pouvons le considérer comme appartenant à une *famille* de calculs distincts $C(0)$, $C(1)$, $C(2)$, $C(3)$, $C(4)$, ..., correspondant respectivement à chacun des entiers naturels 0, 1, 2, 3, 4, ..., et dont la dépendance par rapport à n est elle-même donnée par un calcul.

En termes de machines de Turing, cela signifie que $C(n)$ est l'action d'une machine de Turing sur le nombre n. Autrement dit, le nombre n est une donnée de départ inscrite sur le ruban de la machine, et celle-ci effectue simplement le calcul associé à ce nombre. Si vous ne vous sentez pas à l'aise avec le concept de « machine de Turing », pensez simplement à un ordinateur standard et considérez n comme la « donnée » sur laquelle l'ordinateur va travailler. Ce qui nous intéresse ici est de savoir si, pour chaque valeur de n, l'action de l'ordinateur va ou non s'arrêter.

Pour éclairer la notion de calcul dépendant d'un entier naturel n, considérons les deux exemples suivants :

(F) Trouver un nombre qui ne soit pas la somme de n carrés

et

(G) Trouver un nombre impair qui soit la somme de n nombres pairs.

D'après ce que nous avons vu plus haut, il est clair que le calcul **(F)** s'arrêtera seulement lorsque $n = 0$, 1, 2 et 3 (pour donner respectivement les nombres 1, 2, 3 et 7), tandis que **(G)** ne s'arrêtera pour aucune valeur de n. Si nous entreprenons de démontrer que **(F)** ne s'arrête pas pour n supérieur ou égal à 4, nous devons mettre en branle une colossale machinerie mathématique (la démonstration de Lagrange) ; en revanche, le non-arrêt de **(G)** pour toute valeur de n est évident. Quelles sont les procédures générales dont disposent les mathématiciens pour démontrer qu'un calcul ne s'arrête pas ? Peut-on les formuler elles-mêmes en termes de calcul ?

Supposons que nous ayons une procédure de calcul A dont l'arrêt* fournit une démonstration du non-arrêt d'un tel calcul $C(n)$. Supposons en outre que A englobe *toutes* les procédures dont les mathématiciens humains disposent pour démontrer de manière convaincante qu'un calcul ne s'arrête pas. Si donc, lors d'un calcul particulier, A s'arrête, cela démontrera aux mathématiciens humains que ce calcul *ne s'arrête jamais.* Cette caractéristique supplémentaire de A n'interviendra pas au niveau purement mathématique de la démonstration qui va suivre, mais sera déterminante pour la formulation de la conclusion $\mathscr{G}$.

Je ne demanderai certainement pas que A puisse toujours démontrer que $C(n)$ ne s'arrête pas, mais j'exige qu'elle ne donne jamais de réponse erronée :

---

* Pour les besoins de mon raisonnement, je considère que si la procédure A s'arrête, cela signifie qu'elle a démontré avec succès que $C(n)$ ne s'arrête pas. Si A « se bloque » pour toute autre raison que le « succès », je dirai alors qu'elle a échoué : elle n'a pas démontré que $C(n)$ ne s'arrête pas. Voir plus loin les questions **Q3** et **Q4**, ainsi que l'appendice A.

si elle conclut que $C(n)$ ne s'arrête pas, alors $C(n)$ ne s'arrête effectivement pas. Si A ne donne jamais de réponses erronées, nous dirons qu'elle est *sûre**.

Remarquez que si A n'était pas sûre, il serait alors en principe possible de démontrer cela à l'aide d'un calcul direct : une procédure A qui n'est pas sûre est falsifiable. Car si A affirmait par erreur que le calcul $C(n)$ ne s'arrête jamais alors qu'en vérité il s'arrête, l'exécution même du calcul $C(n)$ conduirait finalement à une réfutation de A. (La faisabilité concrète d'une telle réfutation est un autre problème ; il sera examiné sous **Q8**.)

Nous allons maintenant coder les différents calculs $C(n)$ de sorte que A puisse utiliser ce codage pour agir sur eux. Pour cela, dressons une liste de tous les différents calculs C possibles, par exemple,

$$C_0,\ C_1,\ C_2,\ C_3,\ C_4,\ C_5,\ \ldots,$$

où C_q est le $q^{\text{ième}}$ calcul. Pour exprimer que ces calculs sont appliqués à un nombre donné n, nous écrirons respectivement

$$C_0(n),\ C_1(n),\ C_2(n),\ C_3(n),\ C_4(n),\ C_5(n),\ \ldots$$

On peut considérer que cet ordre est, par exemple, calqué sur un ordre numérique attribué aux programmes informatiques. (D'une manière explicite, nous pouvons, si nous le désirons, considérer que cet ordre est celui de la numérotation des machines de Turing donnée dans EOLP, de sorte que le calcul $C_q(n)$ correspond à l'action de la $q^{\text{ième}}$ machine de Turing T_q opérant sur n.) Un point techniquement important ici est que la constitution de cette liste est un processus *calculable*, *i.e.* il existe un calcul unique** $C_\bullet$ qui donne C_q lorsqu'on l'applique à q, ou, plus précisément encore, le calcul $C_\bullet$ agit sur le *couple* (q, n) — *i.e.* sur q puis sur n — pour donner $C_q(n)$.

On peut maintenant traiter la procédure A comme un calcul particulier qui, appliqué au couple (q, n), cherche à établir que le calcul $C_q(n)$ ne s'arrête jamais. Ainsi, si A *s'arrête*, cela démontre que $C_q(n)$ *ne s'arrête pas*. Le calcul A dépendant des deux nombres q et n, on peut le désigner par $A(q, n)$. On a alors :

(H) Si $A(q, n)$ s'arrête, alors $C_q(n)$ ne s'arrête pas.

Considérons maintenant les énoncés **(H)** pour lesquels q est égal à n. Cela peut sembler une idée curieuse, mais elle est parfaitement légitime. (Elle correspond en fait à la première étape de la puissante procédure de « diagonalisation », découverte au XIX^{e} siècle par le très original et très influent mathématicien danois/russe/allemand Georg Cantor, qui constitue la clé de voûte des raisonnements de Gödel et de Turing.) Pour q égal à n, **(H)** devient :

(I) Si $A(n, n)$ s'arrête, alors $C_n(n)$ ne s'arrête pas.

---

* Certains informaticiens théoriciens disent aussi *correcte* ou *adéquate*. (N.d.T.)

** Cette liste s'obtient en fait par l'action d'une machine de Turing universelle sur le couple (q, n) ; voir l'appendice A et EOLP, p. 56-63.

Remarquons que $A(n, n)$ ne dépend que d'*un seul* nombre, n — et non plus de deux. Elle est alors nécessairement l'un des calculs C_0, C_1, C_2, C_3, C_4, C_5, ... (appliqué à n), car nous avons supposé que cette liste comprenait *tous* les calculs pouvant être effectués sur un entier naturel n. Supposons que $A(n, n)$ soit en fait C_k. On a alors :

(J) $A(n, n) = C_k(n)$.

Examinons maintenant la valeur particulière $n = k$. (C'est la deuxième étape de la procédure diagonale de Cantor !) On a, d'après **(J)**,

(K) $A(k, k) = C_k(k)$

et, d'après **(I)**, avec $n = k$:

(L) Si $A(k, k)$ s'arrête, alors $C_k(k)$ ne s'arrête pas.

Substituant **(K)** dans **(L)**, on trouve :

(M) Si $C_k(k)$ s'arrête, alors $C_k(k)$ ne s'arrête pas.

Il en résulte que le calcul $C_k(k)$ *ne s'arrête pas.* (S'il s'arrêtait, on en déduirait, en vertu de **(M)**, qu'il ne s'arrête pas !) Mais $A(k, k)$ ne peut elle non plus s'arrêter, car en vertu de **(K)**, elle est *identique* à $C_k(k)$. Ainsi, notre procédure A est incapable de vérifier que ce calcul particulier $C_k(k)$ ne s'arrête pas, alors même qu'il ne s'arrête pas.

En outre, si nous *savons* que A est sûre, nous *savons* alors que $C_k(k)$ ne s'arrête pas. Nous savons donc quelque chose que A est incapable de vérifier. Il en résulte que A *ne peut* englober notre compréhension.

Je conseillerais ici au lecteur de relire entièrement le raisonnement que je viens d'exposer, simplement pour être sûr que je ne me suis pas livré quelque part à un petit tour de passe-passe ! Je le reconnais, ce raisonnement paraît tenir de la prestidigitation, mais il est parfaitement légitime et devient d'autant plus convaincant qu'on l'examine plus attentivement. Nous avons trouvé un calcul $C_k(k)$ dont nous savons qu'il ne s'arrête pas ; pourtant, la procédure de calcul A n'a pas le pouvoir de vérifier cela. C'est là le théorème de Gödel (-Turing) — dans la version dont j'ai besoin. Il s'applique à toute procédure de calcul A conçue pour vérifier que des calculs ne s'arrêtent pas et *dont nous savons qu'elle est sûre.* Il en résulte qu'aucun ensemble de règles de calcul (telle A) que l'on sait sûr n'est suffisamment puissant pour vérifier le non-arrêt de tous les calculs qui ne s'arrêtent pas : il existe en effet des calculs (tel $C_k(k)$) qui ne s'arrêtent pas et dont le non-arrêt ne peut être détecté par ces règles. En outre, si nous connaissons la procédure A et si nous savons qu'elle est sûre, nous pouvons construire de toutes pièces un calcul $C_k(k)$ dont nous pouvons *voir* qu'il ne s'arrête pas. Nous en déduisons que, quoi que puisse être A, elle *ne peut pas être* une formalisation des procédures dont disposent les mathématiciens pour vérifier que les calculs ne s'arrêtent pas. Ainsi :

$\mathcal{G}$ Ce n'est pas en utilisant un algorithme qu'ils savent sûr que les mathématiciens humains établissent la vérité mathématique.

Cette conclusion m'apparaît incontournable. Nombre de personnes toutefois la contestent — en lui opposant des objections analogues à celles résumées par les questions **Q1**-**Q20** des sections 2.6 et 2.10 — et contestent certainement la conclusion plus forte selon laquelle le processus de pensée contient un élément fondamentalement non algorithmique. Le lecteur peut d'ailleurs légitimement se demander en quoi un raisonnement mathématique comme celui-ci, sur la nature abstraite du calcul, nous renseigne sur les mécanismes de l'esprit humain. Y a-t-il vraiment un rapport entre tout cela et le problème de la connaissance immédiate consciente ? La réponse à cette question est que ce raisonnement révèle non seulement un aspect très important de la qualité mentale appelée *compréhension* — son rôle dans le problème général du calcul —, mais aussi, ainsi qu'il a été dit à la section 1.12, que la compréhension dépend de la connaissance immédiate consciente. Certes la démonstration précédente est essentiellement un raisonnement mathématique, mais le point capital est que l'algorithme A y participe à deux niveaux tout à fait distincts. D'une part il y est traité comme un simple algorithme possédant certaines propriétés, d'autre part il y est considéré comme étant *réellement* « l'algorithme que nous utilisons nous-mêmes » pour nous convaincre qu'un calcul ne s'arrête pas. Ce raisonnement *n'est donc pas* purement mathématique. Il concerne également la manière dont nous utilisons notre compréhension consciente pour déduire la validité d'un énoncé mathématique — en l'occurrence, le fait que $C_k(k)$ ne s'arrête pas. Et c'est l'interaction entre les deux niveaux auxquels on considère cet algorithme — en tant que candidat à la représentation de l'activité consciente et en tant que calcul en soi — qui nous permet de conclure à l'existence d'une opposition fondamentale entre une telle activité consciente et un simple calcul.

La conclusion $\mathscr{G}$ présente il est vrai plusieurs points faibles. Je vais donc consacrer le reste de ce chapitre à examiner *tous* les contre-arguments pertinents dont j'ai eu connaissance — cela fera l'objet des questions **Q1**-**Q20** des sections 2.6 et 2.10, questions qui incluent également quelques contre-arguments de mon propre cru. J'y répondrai aussi précisément que possible. Nous verrons que $\mathscr{G}$ sortira pratiquement indemne de ces critiques. Puis, au chapitre 3, j'examinerai les conséquences propres de $\mathscr{G}$. Nous découvrirons alors que cette conclusion nous donne de très bonnes raisons de croire que la compréhension mathématique consciente ne peut *aucunement* être simulée à l'aide d'algorithmes, qu'ils soient descendants, ascendants ou mixtes. Nombre de personnes trouveront peut-être cette conclusion alarmante dans la mesure où elle ne nous laisse pas d'autres possibilités. Dans la deuxième partie, je présenterai toutefois un point de vue plus positif. J'exposerai ce que je pense être un argument scientifique plausible étayant mes propres spéculations sur les processus physiques probablement à l'œuvre dans l'activité cérébrale — tels ceux mis en jeu lorsque nous suivons jusqu'au bout un raisonnement de ce type, et expliquant pourquoi cette activité cérébrale pourrait effectivement se dérober à toute description algorithmique.

2.6 Les objections techniques que l'on peut opposer à $\mathscr{G}$

La conclusion $\mathscr{G}$ peut paraître très surprenante, eu égard notamment à la simplicité du raisonnement qui a permis de l'établir. Avant d'envisager son incidence sur la construction d'un robot intelligent commandé par ordinateur et capable de faire des mathématiques — ce sera l'objet du chapitre 3 —, je vais examiner très soigneusement un certain nombre de points techniques concernant ce raisonnement. Si cette discussion ne vous intéresse pas et si vous acceptez la conclusion $\mathscr{G}$, laissez-la de côté (du moins pour l'instant) et passez directement au chapitre 3 ; en outre, si vous acceptez la conclusion plus forte selon laquelle il n'existe *aucune* explication algorithmique à notre compréhension — mathématique ou autre —, vous pouvez passer directement à la deuxième partie — en vous arrêtant peut-être sur le dialogue imaginaire de la section 3.23 (il résume les arguments essentiels du chapitre 3) et sur les conclusions de la section 3.28.

Plusieurs points mathématiques chiffonnent généralement les gens à propos du raisonnement gödelien donné à la section 2.5. Essayons de faire le tri.

Q1. J'ai considéré *une seule* procédure A, alors que nos démonstrations mathématiques recourent indéniablement à de nombreux types de raisonnement. N'aurais-je pas dû envisager la liste complète de tous les « A » possibles ?

En fait, il n'y a aucune perte de généralité à présenter les choses comme que je l'ai fait. Toute liste finie $A_1, A_2, A_3, \ldots, A_i, \ldots, A_r$ de procédures algorithmiques peut toujours être reformulée à l'aide d'un seul algorithme A ayant la propriété de ne pas s'arrêter si *aucun* des algorithmes individuels $A_1, \ldots, A_r$, ne s'arrête. (Cette procédure A pourrait grossièrement fonctionner ainsi : « Exécute les dix premiers pas de A_1 ; rappelle-toi le résultat ; exécute les dix premiers pas de A_2 ; rappelle-toi le résultat ; exécute les dix premiers pas de A_3 ; rappelle-toi le résultat ; et ainsi de suite jusqu'à A_r ; reviens alors à A_1 et exécute ses dix pas suivants ; rappelle-toi le résultat ; et ainsi de suite ; puis de nouveau les dix pas suivants de A_1, etc. Arrête-toi dès que l'un des A_i s'arrête. ») Si maintenant la liste des A_i est infinie, on peut pareillement reformuler l'ensemble entier $A_1, A_2, A_3, \ldots$ à l'aide d'une seule procédure algorithmique A. Ce A résultant agit comme suit :

« dix premiers pas de A_1 ;
dix pas suivants de A_1, dix premiers pas de A_2 ;
dix pas suivants de A_1, dix pas suivants de A_2, dix premiers pas de A_3 ;
etc. »

A s'arrête uniquement si l'un des A_i s'arrête.

On pourrait également imaginer que la liste infinie $A_1, A_2, A_3, \ldots$, n'est pas donnée à l'avance, même en principe, et qu'elle résulte de l'intégration, au fil du temps, de procédures algorithmiques successives. Toutefois, en l'absence

de procédure algorithmique préalablement définie pour générer cette liste, la procédure résultante n'est pas vraiment complète.

Q2. Pourquoi avoir choisi un algorithme *A* fixe ? L'être humain ayant la capacité d'apprendre, l'algorithme qu'il utilise est certainement soumis à de continuelles variations.

Un algorithme variable exige que soient définies les règles auxquelles obéissent ses changements. Si ces règles sont purement algorithmiques, elles figurent déjà au nombre de celles qui spécifient « *A* » ; un « algorithme variable » *de ce type* n'est donc qu'une autre forme d'algorithme fixe, et l'argumentation gödelienne s'applique sans autres modifications. Qu'en est-il maintenant si les variations de l'algorithme *ne sont pas* de nature algorithmique ? Ces variations peuvent correspondre par exemple à l'intégration d'éléments aléatoires ou à une interaction avec l'environnement. Je reviendrai plus loin sur le statut « non algorithmique » des moyens permettant de modifier un algorithme (*cf.* §3.9, §3.10) ; voir aussi la discussion de la section 1.9 affirmant qu'aucun de ces moyens ne fournit d'échappatoire plausible à l'algorithmisme* (et donc ne rejoint le point de vue $\mathcal{G}$). Notre raisonnement étant ici purement mathématique, nous envisagerons uniquement le cas de variations algorithmiques. Mais si l'on accepte que ces variations *ne peuvent* être algorithmiques, on ne peut que reconnaître la validité de la conclusion $\mathcal{G}$.

Peut-être devrais-je être davantage explicite sur ce que l'on peut entendre par algorithme *A* « algorithmiquement variable ». On peut supposer que *A* dépend non seulement de *q* et de *n*, mais aussi d'un autre paramètre *t* représentant par exemple le « temps » ou simplement le décompte du nombre d'activations subies antérieurement par l'algorithme. Quelle que soit sa définition, on peut supposer également que ce paramètre *t* est un entier naturel, de sorte que la liste de nos algorithmes s'écrit maintenant

$$A_0(q, n), A_1(q, n), A_2(q, n), A_3(q, n), \ldots,$$

où chaque $A_t(q, n)$ est censé être une procédure sûre vérifiant qu'un calcul $C_q(n)$ ne s'arrête pas et dont l'ensemble croît en efficacité à mesure que *t* augmente, le moyen permettant d'améliorer l'efficacité de ces $A_t(q, n)$ étant supposé algorithmique. Ce « moyen algorithmique » peut éventuellement dépendre des « expériences » acquises par les $A_t(q, n)$ antérieurs, mais ces « expériences » sont elles aussi censées être algorithmiquement générées (sinon, on se retrouve en accord avec $\mathcal{G}$), de sorte que l'on peut les inclure — elles ou les moyens qui les ont générées — dans ce qui constitue l'algorithme suivant (*i.e. dans* $A_t(q, n)$ lui-même). On aboutit ainsi à *un seul* algorithme ($A_t(q, n)$) dépendant algorithmiquement des *trois* paramètres *t*, *q*, *n*. On peut alors construire un algorithme A^* qui est aussi efficace que la liste entière des $A_t(q, n)$, mais qui dépend seulement des deux entiers naturels *q* et *n*. Pour

* Ce terme d'« algorithmisme », qui désigne (essentiellement) mon « point de vue $\mathcal{A}$ », a été forgé par Hao Wang (1993).

construire cet algorithme $A^*(q, n)$, il suffit, comme plus haut, d'exécuter les 10 premiers pas de $A_0(q, n)$ en se rappelant le résultat ; puis les 10 premiers pas de $A_1(q, n)$ et les 10 pas suivants de $A_0(q, n)$ en se rappelant les résultats ; puis les 10 premiers pas de $A_2(q, n)$, les 10 pas suivants de $A_1(q, n)$ et les 10 pas suivants de $A_0(q, n)$ en se rappelant les résultats ; et ainsi de suite, en se rappelant à chaque fois les résultats. L'*arrêt* de $A^*(q, n)$ intervient finalement lorsque l'un *quelconque* de ces algorithmes s'arrête. Si alors on remplace A par A^* dans le raisonnement gödelien, on aboutit à la même conclusion $\mathscr{G}$.

Q3. N'ai-je pas imposé une restriction injustifiée en exigeant que A continue de calculer indéfiniment dans les cas où il apparaît en fait clairement que $C_q(n)$ *s'arrête* ? Si l'on autorise A à *s'arrêter* lorsque $C_q(n)$ s'arrête, l'argumentation gödelienne s'écroule. Il arrive somme toute que l'intuition permette à un mathématicien humain de conclure à l'arrêt de certains calculs. Or je semble avoir ignoré ce point. Cela ne signifie-t-il pas que j'ai envisagé une situation manquant de généralité ?

Absolument pas. L'argumentation ne doit être appliquée qu'aux intuitions qui nous permettent de conclure qu'un calcul *ne s'arrête pas*, et non à celles qui nous permettent de conclure le contraire. Le succès de l'algorithme hypothétique A ne consiste pas à conclure qu'un calcul *s'arrête*. Ce n'est pas son travail.

Si cela vous pose problème, pensez à ceci : incluez ces *deux* types d'intuition dans A en lui faisant toutefois délibérément décrire une boucle (*i.e.* en lui faisant simplement répéter indéfiniment une même opération quelconque) lorsque la conclusion est que le calcul $C_q(n)$ s'arrête. Bien sûr, ce n'est pas ainsi qu'agirait un mathématicien, mais là n'est pas la question. L'argumentation gödelienne est un *raisonnement par l'absurde* : elle part de l'hypothèse que nous utilisons un algorithme — que nous savons sûr — pour vérifier une affirmation mathématique, puis elle en déduit une contradiction. Il n'est pas nécessaire, dans ce raisonnement, que l'algorithme *soit* réellement cet algorithme hypothétique ; ce peut être un algorithme construit à partir de lui, comme le A dont nous venons de parler à l'instant.

Ce commentaire s'appliquerait pareillement à toute autre objection soulevée contre le raisonnement de la section 2.5 et qui serait de la forme : « Il se pourrait que A s'arrête pour diverses fausses raisons sans pour autant démontrer que $C_q(n)$ ne s'arrête pas. » Si on se donne un « A » qui se comporte ainsi, il suffit d'appliquer le raisonnement de la section 2.5 à un A légèrement différent, à savoir celui qui décrit une boucle chaque fois que le « A » original s'arrête pour l'une de ces fausses raisons.

Q4. En établissant la liste C_0, C_1, C_2, ..., je semble avoir supposé que chaque C_q désigne un calcul bien défini, alors que ce n'est certainement pas le cas de tout classement rudimentaire de programmes informatiques, qu'il soit numérique ou lexicographique.

On pourrait effectivement difficilement assurer que pour tout entier naturel q, chaque C_q de notre liste soit un calcul qui fonctionne. Par exemple, la

classification des machines de Turing donnée dans EOLP n'offre pas une telle garantie ; *cf.* EOLP, p.58-59. Avec les règles de spécifications utilisées dans cet ouvrage, une machine de Turing T_q correspondant à un entier q donné est qualifiée de « ratée » si elle est affectée de l'un des quatre handicaps suivants : soit elle fonctionne indéfiniment sans jamais s'arrêter ; soit elle n'est pas « correctement spécifiée » parce que le nombre q conduit à un développement binaire comprenant trop de 1 successifs (cinq ou plus) et n'admettant donc aucune interprétation dans la spécification considérée ; soit elle rencontre une instruction lui disant de se mettre dans un état interne non existant ; soit enfin elle produit, lorsqu'elle s'arrête, un simple ruban vierge, ce qui n'admet aucune interprétation numérique. (Voir aussi l'appendice A.) Pour les besoins du raisonnement de Gödel-Turing que je viens de donner, il suffit de regrouper tous ces handicaps sous la rubrique « ne marche pas ». En particulier, quand je dis que la procédure de calcul A « s'arrête » (*cf.* la note p. 66), cela signifie qu'elle « s'arrête » dans le sens mentionné plus haut (et donc qu'elle ne contient aucune suite de 1 dépourvue d'interprétation ni ne produit de ruban vierge) — *i.e.* « s'arrête » signifie que le calcul est un calcul correctement spécifié et qui fonctionne. De même, « $C_q(n)$ s'arrête » signifie qu'il s'arrête correctement dans ce même sens. Si l'on adopte cette interprétation, mon raisonnement gödelien n'est pas affecté par les considérations **Q4**.

Q5. N'ai-je pas simplement démontré que l'on peut mettre en échec une procédure algorithmique *particulière* A à l'aide du calcul $C_q(n)$? Pourquoi cela montre-t-il que je peux faire mieux que n'importe quel A ?

L'argumentation gödelienne montre *indéniablement* que nous pouvons mieux faire que *n'importe quel* algorithme. C'est là tout l'intérêt d'utiliser, comme je l'ai fait, un raisonnement par l'absurde. Pour mieux comprendre cela, aidons-nous d'une analogie. Certains lecteurs connaissent probablement l'argument avancé par Euclide pour montrer qu'il n'existe pas de plus grand nombre premier. Cet argument est lui aussi un raisonnement par l'absurde. Le voici : supposons, au contraire, qu'il existe un nombre premier p supérieur à tous les autres. Considérons alors le produit de tous les nombres premiers jusqu'à p et ajoutons-lui 1. On obtient un nombre N donné par :

$$N = 2 \times 3 \times 5 \times \ldots \times p + 1.$$

N est manifestement plus grand que p, mais n'est divisible par aucun des nombres premiers 2, 3, 5, ..., p (car chacune de ces divisions donne un reste égal à 1) ; ainsi, soit N est le nombre premier recherché, soit il n'est pas premier — auquel cas il est divisible par un nombre premier supérieur à p. Dans les deux cas, il y a un nombre premier supérieur à p, ce qui contredit l'hypothèse initiale selon laquelle p est le plus grand des nombres premiers. Ainsi, il n'existe pas de plus grand nombre premier.

Ce raisonnement par l'absurde ne montre pas simplement que l'on peut dépasser un nombre premier *donné* en en trouvant un plus grand ; il montre également qu'*il n'existe pas* de nombre premier supérieur à tous les autres. De même, l'argumentation de Gödel-Turing montre non seulement que l'on

peut mettre en échec un algorithme *donné*, mais aussi qu'il n'existe *aucun* algorithme (sûr) qui soit équivalent aux intuitions que nous utilisons pour vérifier que certains calculs ne s'arrêtent pas.

Q6. On pourrait programmer un ordinateur pour qu'il suive exactement l'argumentation que je viens de développer. Ne pourrait-il, ainsi, parvenir lui-même à toute conclusion à laquelle j'ai moi-même abouti ?

Il est clair que la détermination du calcul particulier $C_k(k)$ associé à l'algorithme A est un processus algorithmique. Cela peut d'ailleurs être démontré de manière explicite*. Cela signifie-t-il pour autant que l'intuition mathématique soi-disant non algorithmique — cette intuition qui nous permet de voir que $C_k(k)$ ne s'arrête jamais — est, en définitive, algorithmique ?

Je pense que cette question mérite d'être examinée en détail, car elle résume l'un des malentendus les plus répandus à propos du raisonnement gödelien. Il devrait être clair ensuite qu'elle n'infirme *rien* de ce qui a été dit auparavant. Bien que la procédure qui permet d'obtenir $C_k(k)$ à partir de A puisse s'exprimer par un calcul, ce calcul ne fait pas partie des procédures contenues dans A. La raison en est que A est incapable de vérifier que $C_k(k)$ ne s'arrête pas, tandis que ce nouveau calcul (associé à A) en est capable. Ainsi, bien que ce nouveau calcul conduise effectivement à $C_k(k)$, il ne fait pas partie du club des « juges officiels de la vérité ».

Permettez-moi d'exposer les choses autrement. Imaginez un robot, commandé par ordinateur, qui serait capable de vérifier des vérités mathématiques à l'aide des procédures algorithmiques contenues dans A. Pour rendre ma description plus vivante, je vais utiliser un langage anthropomorphique et dire que le robot « connaît » ces vérités mathématiques — en l'occurrence le non-arrêt de calculs — qu'il parvient à déduire en utilisant A. Or si A est tout ce que « connaît » notre robot, il *ne « saura » pas* que $C_k(k)$ ne s'arrête pas, même si la procédure pour obtenir $C_k(k)$ à partir de A est parfaitement algorithmique. Bien sûr, nous pourrions *dire* au robot qu'en réalité $C_k(k)$ ne s'arrête pas (en utilisant notre propre intuition à cet effet), mais si le robot acceptait ce fait, il devrait alors modifier ses propres règles en ajoutant cette nouvelle vérité à celles qu'il « connaît » déjà. On pourrait imaginer plus et faire savoir à notre robot, par des moyens appropriés, que la procédure de calcul générale pour obtenir $C_k(k)$ à partir de A est également un moyen qu'il devrait « connaître » pour obtenir de nouvelles vérités à partir des anciennes. Le « savoir » du robot pourrait ainsi s'accroître grâce à toutes sortes de procédures bien définies et algorithmiques. Mais nous serions alors en présence d'un *nouvel* « A », et c'est à ce nouvel « A », et non plus à l'ancien, que s'appliquerait alors le raisonnement gödelien. Autrement dit, nous aurions dû utiliser ce nouvel « A » — et

---

* Pour souligner la valeur que j'accorde à cette objection, je renvoie le lecteur à l'appendice A où est explicitement présentée une procédure de calcul (en recourant aux règles données en détail au chapitre 2 de EOLP) permettant d'obtenir l'action d'une machine de Turing $C_k(k)$ à partir de l'algorithme A. A y est supposé donné par la spécification d'une machine de Turing T_a ; le verdict de T_a sur $C_q(n)$ est défini par le codage de son action sur q puis sur n.

non l'ancien — dès le début de notre démonstration, car changer de « A » en cours de route équivaut à tricher. Ainsi, on constate que le contresens contenu dans **Q6** est très semblable à celui de **Q5**. Notre raisonnement par l'absurde suppose que A — qui est censé être une procédure, connue et sûre, pour vérifier que des calculs ne s'arrêtent pas — représente réellement la *totalité* des procédures de ce type dont disposent les mathématiciens, puis en déduit une contradiction. Introduire une autre procédure de calcul — juge de la vérité — non contenue dans A *après* avoir décidé que A représente cette totalité est une pure tricherie.

Le problème pour notre pauvre robot est qu'en l'absence de toute *compréhension* du raisonnement de Gödel, il ne dispose d'aucun moyen indépendant et fiable pour découvrir par lui-même la vérité, sinon de la connaître par notre intermédiaire. (C'est là un problème distinct des aspects algorithmiques de l'argumentation gödelienne.) Pour pouvoir faire plus que cela, il doit, comme nous, comprendre le sens des opérations qu'on lui a demandé d'effectuer. Sans compréhension, il pourrait aussi bien « savoir » (à tort) que $C_k(k)$ *s'arrête* alors qu'il ne s'arrête pas. Tout comme la déduction (correcte) du non-arrêt de $C_k(k)$, la déduction (erronée) de l'arrêt de $C_k(k)$ est une affaire d'algorithme. Ainsi, la nature algorithmique de ces opérations n'est pas en cause ; ce qui est en cause ici c'est que pour distinguer les algorithmes qui le font aboutir à un jugement correct de ceux qui le font aboutir à un jugement erroné, notre robot a besoin de disposer de *jugements de vérité* fiables. À ce stade de l'argumentation, rien n'interdit cependant que la « compréhension » corresponde à un autre type d'activité algorithmique, correctement défini et sûr, mais absent de toutes les procédures telles que A. Par exemple, la compréhension pourrait être donnée par un algorithme inconnaissable ou non sûr. Lors d'une discussion ultérieure (au chapitre 3), je tenterai de convaincre le lecteur que la compréhension est une activité qui n'a en fait rien d'algorithmique. Mais pour l'instant, puisque nous nous intéressons uniquement aux conséquences rigoureuses de l'argumentation de Gödel-Turing, le fait que l'on puisse obtenir algorithmiquement $C_k(k)$ à partir de A ne nous concerne pas.

Q7. Tous les résultats de tous les mathématiciens qui ont vécu jusqu'ici, plus tous les résultats de tous les mathématiciens humains qui vivront (disons) au cours du prochain millier d'années forment un ensemble fini que l'on pourrait stocker dans la mémoire d'un ordinateur approprié. Cet ordinateur *pourrait* donc certainement simuler ces résultats et se comporter (extérieurement) comme un mathématicien humain — même si l'argumentation gödelienne semble nous dire le contraire.

Si cette affirmation est apparemment vraie, elle ignore le point essentiel, à savoir comment nous (ou les ordinateurs) savons que tel énoncé mathématique est vrai et que tel autre est faux. (Quoi qu'il en soit, le simple *stockage* d'énoncés mathématiques pourrait être accompli par un dispositif bien moins sophistiqué qu'un ordinateur, *e.g.* par enregistrement photographique.) Le recours à l'ordinateur dans **Q7** dénote d'ailleurs une méconnaissance totale du problème crucial du *jugement de vérité*. On pourrait également envisager des

ordinateurs contenant uniquement des listes de « théorèmes » mathématiques totalement faux, ou des mélanges aléatoires de vérités et d'erreurs. À quel ordinateur ferions-nous alors confiance ? Je n'affirme pas ici l'impossibilité d'une simulation parfaite du produit de l'activité humaine consciente (en l'occurrence, des mathématiques) : il se peut en effet que par pur hasard l'ordinateur accomplisse correctement cette simulation — même sans disposer de la moindre parcelle de compréhension. Mais cette possibilité est fabuleusement faible, et **Q7** ne s'attaque pas aux problèmes que j'examine dans ce livre, à savoir comment on décide de la vérité ou de la fausseté des énoncés mathématiques.

En revanche, **Q7** pose involontairement une question plus sérieuse : les discussions sur les structures infinies (*e.g. tous* les entiers naturels ou *tous* les algorithmes) ont-elles vraiment un rapport avec nos présentes considérations dans la mesure où les résultats des humains et des ordinateurs sont *finis* ? Examinons maintenant cet important problème.

Q8. Les calculs qui ne se terminent pas sont des constructions mathématiques idéalisées associées au concept d'infini. Ont-ils dès lors vraiment un rapport avec les discussions sur les objets physiques finis tels que les cerveaux et les ordinateurs ?

C'est vrai. Nos discussions théoriques sur les machines de Turing, sur les calculs sans fin, etc., mettent en jeu des processus (potentiellement) infinis, tandis que les êtres humains ou les ordinateurs sont des systèmes *finis*. Il importe donc d'évaluer la portée de ces discussions théoriques lorsqu'on les applique à des objets physiques réels et finis. Il s'avère cependant que la prise en compte de la finitude n'affecte pas substantiellement l'argumentation de Gödel-Turing. Rien en effet n'interdit de *raisonner* sur des calculs idéalisés, puis de déduire mathématiquement les limites théoriques de ces calculs. Nous pouvons par exemple nous demander, en termes parfaitement finis, s'il existe un nombre impair qui soit la somme de deux nombres pairs ou s'il existe un entier naturel qui ne soit pas la somme de quatre carrés (comme dans (**B**) et (**C**) à la section 2.3), même si, en traitant ces problèmes, nous considérons implicitement l'ensemble infini de *tous* les entiers naturels. Nous pouvons tout à fait légitimement raisonner sur des calculs qui ne se terminent pas ou, d'une manière générale, sur des machines de Turing en tant que constructions *mathématiques*, même s'il est impossible, concrètement, de construire une machine de Turing fonctionnant indéfiniment. (Remarquez notamment que l'action d'une machine de Turing qui chercherait un nombre impair somme de deux nombres pairs est, strictement parlant, physiquement irréalisable : au bout d'un certain temps, l'usure des pièces de la machine provoquerait son arrêt.) La spécification de n'importe quel calcul (*i.e.* l'action d'une machine de Turing) est une opération parfaitement finie, et le problème de savoir si ce calcul se termine ou non est parfaitement bien défini. Mais une fois achevés nos raisonnements sur ces calculs idéalisés, nous devons regarder dans quelle mesure ils s'appliquent à des systèmes finis tels qu'un ordinateur réel ou un être humain.

Les limitations dues à la finitude surviennent soit (i) parce que la spécification du calcul considéré est démesurément grande (*i.e.* le nombre n dans C_n,

ou celui associé à la paire de nombres q, n dans $C_q(n)$, est trop grand pour être spécifié par un ordinateur réel ou un être humain), soit (ii), dans le cas où la spécification est acceptable, parce que l'exécution du calcul dure trop longtemps, de sorte qu'elle semble ne devoir jamais s'arrêter, même si, théoriquement, elle finit par s'arrêter. Nous le verrons, il s'avère que seul (i) affecte vraiment notre discussion — mais seulement dans une certaine mesure. L'absence d'influence de (ii) est peut-être surprenante. Il existe en effet nombre de calculs relativement simples qui finissent par s'arrêter, mais qu'aucun ordinateur concevable ne pourrait cependant mener à bien. Pensez par exemple au calcul suivant : « Imprimez une succession de $2^{2^{65\,536}}$ "1" puis arrêtez-vous. » (Quelques exemples mathématiquement bien plus intéressants seront donnés à la section 3.26.) Le recours à l'exécution directe pour déterminer si un calcul s'arrête ou non est souvent très inefficace.

Pour voir comment les limitations (i) ou (ii) peuvent affecter l'argumentation gödelienne, revenons sur les points essentiels de cette dernière. Pour satisfaire (i), notre liste de calculs doit être non pas infinie, mais *finie*. Écrivons-la sous la forme :

$$C_0, C_1, C_2, C_3, \ldots, C_Q,$$

où Q spécifie le plus grand calcul que notre ordinateur, ou notre être humain, est capable d'effectuer. Dans le cas d'un être humain, on peut objecter que la définition de Q contient un certain flou. Mais peu importe pour l'instant que Q ait ou non une valeur précise. (Le problème des capacités de calcul de l'être humain sera examiné à la section 2.10 en réponse à **Q13**.) Supposons en outre que lorsqu'on applique ces calculs à un entier naturel donné n, la valeur de n soit inférieure ou égale à un nombre fixe N parce que notre ordinateur (ou notre être humain) n'est pas conçu pour traiter des nombres supérieurs. (Strictement parlant, N pourrait être un nombre variable dépendant du calcul C_q envisagé — *i.e.* N pourrait dépendre de q —, mais cette nuance n'influe pas de manière notable sur notre raisonnement.)

Comme plus haut, nous considérons un algorithme sûr $A(q, n)$ dont l'arrêt démontre que le calcul $C_q(n)$ ne s'arrête pas. Lorsque nous disons « sûr », bien que selon (i) il suffise de considérer les valeurs de q qui ne sont pas supérieures à Q et les valeurs de n qui ne sont pas supérieures à N, nous entendons en fait que A est sûr pour *toutes* les valeurs de q et de n, aussi grandes soient-elles. (Ainsi, les règles contenues dans A sont des règles *mathématiques* précises, et non des règles approximatives liées à des contraintes d'ordre pratique déterminant quels calculs peuvent « réellement » être effectués.) De plus, lorsque nous disons que « $C_q(n)$ ne se termine pas », nous voulons dire qu'il ne se termine *réellement* pas, et non, au sens de (ii), que ce calcul est simplement trop long pour être exécuté par notre ordinateur ou notre être humain.

Rappelons que **(H)** dit :

Si $A(q, n)$ s'arrête, alors $C_q(n)$ ne s'arrête pas.

Étant donné la limitation (ii), il semble que si l'algorithme A comprend plus d'étapes que notre ordinateur ou notre être humain ne peuvent en traiter, il

ne nous sera guère utile pour décider si un autre calcul s'arrête ou non. Il s'avère toutefois que cela n'a aucune importance pour notre raisonnement. Nous allons en effet trouver un algorithme $A(k, k)$ qui, lui, ne s'arrête pas. Peu nous importera si, dans certains cas, ceux dans lesquels A s'arrête *réellement*, nous devrions attendre trop longtemps pour constater qu'effectivement il s'arrête.

De **(H)**, nous aboutissons ensuite à **(J)** qui nous dit qu'il existe un entier naturel k pour lequel l'algorithme $A(n, n)$ est identique à $C_k(n)$ quel que soit n :

$$A(n, n) = C_k(n).$$

Nous devons toutefois maintenant tenir compte de l'éventualité où, en vertu de (i), ce k serait supérieur à Q. Pour un A horriblement compliqué, cela peut effectivement être le cas, mais seulement si A est déjà proche de la taille limite supérieure (telle que l'exprime le nombre de chiffres binaires correspondant à la spécification de A en tant que machine de Turing) pouvant être traitée par notre ordinateur ou notre être humain. Cela tient au fait que le calcul qui donne la valeur k à partir de la spécification de A (par exemple, en termes de machine de Turing) est extrêmement simple et peut être donné explicitement (ainsi que nous l'avons déjà remarqué dans la réponse à **Q6**).

Le véritable calcul dont nous avons besoin pour prendre A en défaut est $C_k(k)$. En posant $n = k$ dans **(H)**, on obtient :

(L) Si $A(k, k)$ s'arrête, alors $C_k(k)$ ne s'arrête pas.

Puisque $A(k, k)$ est identique à $C_k(k)$, notre raisonnenment montre que ce calcul particulier $C_k(k)$ ne s'arrête pas et que A ne peut vérifier cela, même s'il pouvait tourner bien plus longtemps que toute limite imposée par (ii). La spécification de $C_k(k)$ s'exprime à l'aide du k ci-dessus, et pourvu que k ne soit pas supérieur à Q ou à N, $C_k(k)$ est un calcul que pourrait effectivement exécuter notre ordinateur ou notre être humain — dans le sens où ce calcul pourrait *démarrer*. Il ne pourrait, en tout état de cause, être mené à son terme puisqu'il ne s'arrête en fait jamais !

Se peut-il maintenant que k soit réellement supérieur à Q ou à N ? Cela est possible uniquement si la spécification de A exige un nombre de chiffres si important qu'un accroissement modeste de ce nombre entraînerait un dépassement de la capacité de calcul de notre ordinateur ou de notre être humain. Il découle cependant du fait que A est sûr que nous *savons* que ce $C_k(k)$ ne s'arrête pas, même si nous pouvons rencontrer des difficultés pour exécuter ce calcul. La limitation (i) ne nous permet pas d'écarter l'éventualité que A soit d'une complexité telle que sa spécification le situerait près de la limite des calculs à la portée d'un être humain, mais que l'accroissement comparativement faible du nombre de chiffres considéré donne un calcul hors de portée d'un cerveau humain. Selon moi, quoi que nous puissions penser d'une telle éventualité, un ensemble aussi énorme de règles de calcul contenu dans cet A hypothétique serait assurément si complexe que nous ne pourrions probablement *savoir* si ce A est ou non *sûr*, même si nous pouvions en connaître les règles

précises. Ainsi, notre conclusion reste inchangée : nous *ne pouvons* vérifier un énoncé mathématique à l'aide d'un ensemble de règles algorithmiques *que nous savons sûres.*

Examinons maintenant d'un peu plus près l'accroissement de complexité comparativement faible intervenant dans le passage de A à $C_k(k)$. Cet accroissement revêtira une importance particulière aux sections 3.19 et 3.20. L'appendice A donne une expression explicite de $C_k(k)$ à l'aide des instructions de machine de Turing présentées au chapitre 2 de EOLP. Selon ces instructions, T_m désigne la « $m^{\text{ième}}$ machine de Turing ». Pour plus de clarté, nous utiliserons ici cette notation plutôt que « C_m », notamment pour définir le *degré de complexité* d'une procédure de calcul ou d'un calcul individuel. Je poserai ainsi que le degré de complexité μ de la machine de Turing T_m est égal au nombre de chiffres binaires contenus dans la spécification de m développé sous forme binaire (*cf.* EOLP, p. 58) ; le degré de complexité d'un calcul particulier $T_m(n)$ est alors défini par le plus grand des deux nombres μ et ν, ν étant le nombre de chiffres binaires contenus dans la spécification de n. Considérons maintenant les instructions — données explicitement dans l'appendice A — permettant d'obtenir le calcul $C_k(k)$ à partir de A. En désignant par α le degré de complexité de A, on trouve que le degré de complexité du calcul explicite $C_k(k)$ est inférieur à $\alpha + 210 \log_2(\alpha + 336)$, nombre qui, pour α très grand, ne dépasse α que d'une quantité relativement infime.

Ce raisonnement contient cependant un point qui peut troubler certains lecteurs : n'est-il pas absurde de considérer un calcul qui serait trop complexe pour être écrit ou qui, une fois écrit, nécessiterait un temps très largement supérieur à l'âge de l'Univers pour être exécuté, même si l'exécution de chacune de ses étapes durait la plus petite des fractions de seconde durant laquelle peuvent raisonnablement se dérouler des processus physiques ? Le calcul mentionné plus haut — qui imprime une succession de $2^{2^{65\,536}}$ « 1 » et ne s'arrête qu'une fois cette tâche achevée — en est un exemple, et ce serait adopter une attitude mathématique extrêmement anticonformiste que d'affirmer qu'il ne se termine pas. Il existe toutefois quelques points de vue, pas anticonformistes au point de dire qu'un tel calcul ne s'arrête pas — mais malgré tout résolument anticonformistes — selon lesquels on pourrait douter de la vérité absolue des énoncés mathématiques idéalisés. C'est ce type de points de vue que nous allons maintenant passer brièvement en revue.

Q9. Selon l'*intuitionnisme* — et d'autres points de vue « constructivistes » et « finitistes » —, on ne peut déduire qu'un calcul se termine du simple fait que s'il continuait indéfiniment, on aboutirait à une contradiction. Cela met-il en question le raisonnement gödelien ?

Dans mon raisonnement gödelien, j'ai utilisé, en **(M)**, un argument de la forme : « L'hypothèse "X est faux" conduit à une contradiction, donc X est vrai. » Dans **(M)**, « X » était l'énoncé « $C_k(k)$ ne s'arrête pas ». C'est là ce que l'on appelle un *raisonnement par l'absurde* — et de fait, l'argumentation gödelienne tout entière repose sur ce type de raisonnement. Le point de vue mathématique appelé « intuitionnisme » (introduit vers 1912 par le mathématicien

hollandais L. E. J. Brouwer ; *cf.* Kleene 1952 et EOLP, p. 120-124) nie que l'on puisse légitimement raisonner en se fondant sur un raisonnement par l'absurde. L'intuitionnisme a vu le jour en réaction à certaines tendances mathématiques de la fin du XIX^e^ siècle et du début du XX^e^ selon lesquelles on pouvait affirmer l'« existence » d'un objet mathématique même s'il n'y avait aucun moyen de construire réellement cet objet. Une trop libre utilisation d'un nébuleux concept d'existence mathématique conduisait effectivement parfois à une contradiction. L'exemple le plus célèbre en est l'ensemble paradoxal de Bertrand Russell, à savoir « l'ensemble de tous les ensembles qui n'appartiennent pas à eux-mêmes ». (Si l'ensemble de Russel appartient à lui-même, il n'appartient alors pas à lui-même ; et s'il lui appartient, il ne lui appartient pas ! Voir la section 3.4 et EOLP, p. 106, pour plus de détails.) Afin de contrer cette tendance générale qui autorisait l'« existence » d'objets mathématiques très librement définis, le point de vue intuitionniste rejeta la validité des raisonnements concluant à l'existence d'un objet mathématique à partir de la simple nature contradictoire de sa non-existence. Un tel raisonnement par l'absurde ne fournit pas une construction réelle de l'objet en question.

En quoi le rejet de cette utilisation du raisonnement par l'absurde affecte-t-il notre argumentation gödelienne ? En rien. Simplement parce que le raisonnement par l'absurde y est utilisé de manière inverse, à savoir que la contradiction se déduit d'une hypothèse d'*existence*, et non d'une hypothèse de non-existence. Selon l'intuitionnisme, il est parfaitement légitime de déduire la *non-existence* d'un objet à partir de la contradiction entraînée par l'hypothèse de son existence. L'argumentation gödelienne telle que je l'ai donnée ici est donc parfaitement acceptable d'un point de vue intuitionniste. (Voir Kleene 1952, p. 492.)

Des remarques analogues valent également pour tous les autres points de vue « constructivistes » ou « finitistes » que je connais. La réponse à **Q8** montre que même le point de vue suggéré par cette question, à savoir que l'on ne peut « réellement » considérer que les entiers naturels continuent indéfiniment, ne permet pas d'échapper à la conclusion $\mathcal{G}$: ce n'est pas un algorithme sûr que l'on utilise pour établir une vérité mathématique.

2.7 Quelques considérations mathématiques plus profondes

Pour mieux apprécier la portée du raisonnement gödelien, nous allons revenir sur les motivations qui lui ont donné naissance. Au début du XX^e^ siècle, les spécialistes des fondements des mathématiques se heurtèrent à un certain nombre de graves difficultés. Vers la fin du XIX^e^ — grâce en grande partie aux travaux profondément originaux de Georg Cantor (dont nous avons déjà

rencontré la méthode de « diagonalisation ») —, les mathématiciens avaient découvert, en recourant aux propriétés des *ensembles infinis*, des méthodes puissantes permettant d'établir certains de leurs plus profonds résultats. Toutefois, si l'on faisait un usage trop libre du concept d'ensemble infini, les avantages procurés par ces méthodes s'accompagnaient de difficultés fondamentales. En particulier, le paradoxe de Russell (que j'ai brièvement évoqué dans la réponse à **Q9** — *cf.* aussi §3.4 — et qui avait également été remarqué par Cantor) mettait en lumière certains des obstacles surgissant lorsqu'on raisonnait d'une manière trop cavalière sur les ensembles infinis. Il apparut cependant qu'en prenant certaines précautions de raisonnement, on pouvait effectivement obtenir de puissants résultats mathématiques. Le problème sembla ainsi se réduire à définir avec une *précision* absolue ce que l'on devait entendre par « précaution de raisonnement ».

Un mouvement, baptisé *formalisme*, vit alors le jour, qui se fixa pour objectif de garantir cette précision et dont l'un des principaux animateurs fut le grand mathématicien David Hilbert. Selon le formalisme, toutes les formes de raisonnement — y compris les raisonnements sur les ensembles infinis — autorisées dans tout domaine précis des mathématiques devaient être fixées une fois pour toutes. Ces formes de raisonnement se présentent comme un ensemble de règles et d'énoncés mathématiques appelé *système formel*. Une fois déterminées les règles d'un système formel $\mathbb{F}$, il ne s'agit plus ensuite que de vérifier mécaniquement si ces règles — nécessairement en nombre fini* — ont été correctement appliquées. Bien entendu, ces règles devaient être considérées comme des formes valides de raisonnement mathématique, de sorte que tout résultat déduit de leur utilisation pouvait être admis comme *vrai*. Toutefois, lorsque certaines de ces règles concernaient la manipulation d'ensembles infinis, la frontière séparant les formes légitimes de raisonnement de celles qui ne le sont absolument pas pouvait varier d'un mathématicien à l'autre. Tous les doutes pouvaient en effet être raisonnablement permis sur ce plan, eu égard aux incohérences survenant lorsqu'on allait jusqu'à s'autoriser à utiliser par exemple l'ensemble paradoxal de Bertrand Russell, l'« ensemble de tous les ensembles qui ne sont pas membres d'eux-mêmes ». Les règles de $\mathbb{F}$ devaient certes interdire l'existence de cet « ensemble » de Russell, mais où devaient-elles s'arrêter ? Interdire l'utilisation de tout ensemble infini représentait une contrainte trop grande (par exemple, l'espace euclidien ordinaire comprend un ensemble infini de points, et même l'ensemble des entiers naturels est infini) ; en outre, il était clair qu'existaient en fait divers types de systèmes formels parfaitement satisfaisants (ne permettant pas, par exemple, une

* Certains systèmes formels sont présentés comme ayant un nombre *infini* d'axiomes — décrits en termes de structures appelées « schémas axiomatiques ». Mais pour avoir véritablement droit au titre de « système formel » tel que je l'entends ici, ces systèmes doivent pouvoir être exprimés en termes finis, autrement dit être engendrés par un ensemble fini de règles de calcul. Cela est effectivement possible pour les systèmes formels standard utilisés dans les démonstrations mathématiques — tels le classique système formel de « Zermelo-Fraenkel » $\mathbb{ZF}$ décrivant la théorie des ensembles.

description explicite de l'« ensemble » de Russell) à partir desquels on pouvait obtenir la plupart des résultats mathématiques recherchés. Comment distinguer alors les systèmes formels fiables de ceux qui ne l'étaient pas ?

Considérons l'un de ces systèmes formels F et qualifions de VRAIS les énoncés mathématiques pouvant être obtenus à l'aide des règles de F, et de FAUX ceux dont on peut, par le même moyen, obtenir la *négation* (*i.e.* « non » l'énoncé en question). Qualifions ensuite d'INDÉCIDABLE tout énoncé qui peut être formulé dans F mais qui n'est ni VRAI ni FAUX. Certains estimèrent que puisque les ensembles infinis eux-mêmes étaient peut-être réellement « dépourvus de signification », il n'y avait aucun sens à parler de vérité ou de fausseté absolues à leur propos (cela valait du moins pour certains types d'ensembles infinis). Ainsi, selon ce point de vue, peu importait réellement de savoir quels énoncés concernant certains ensembles infinis étaient VRAIS et quels autres étaient FAUX, à condition qu'aucun énoncé ne puisse être *en même temps* VRAI *et* FAUX — ce qui revient à exiger que le système F soit *consistant* (ou *non contradictoire*). Pour ces *formalistes* de stricte obédience, les seules questions revêtant une importance primordiale au niveau d'un système formel F étaient de savoir (a) si ce système était ou non *consistant* et (b) s'il était ou non *complet*. Le système F est dit *complet* si tout énoncé mathématique correctement formulé dans F s'avère toujours être soit VRAI soit FAUX (de sorte que F ne contient aucun énoncé INDÉCIDABLE).

Pour un pur formaliste, la question de savoir si un énoncé sur les ensembles infinis est *réellement vrai* n'a pas forcément de sens et n'a assurément aucun rapport avec les procédures des mathématiques formelles. Ainsi, à la recherche de la vérité mathématique absolue des énoncés portant sur ces quantités infinies se substitua la recherche d'une démonstration de la consistance et de la complétude de systèmes formels appropriés. Sur quelles règles mathématiques une telle démonstration pouvait-elle reposer ? Ces règles devaient elles-mêmes être fiables et ne recourir à aucun raisonnement douteux sur des ensembles infinis non rigoureusement définis (tels l'ensemble de Russell). On espéra trouver, au sein de certains systèmes formels manifestement sûrs et comparativement simples (tel le système relativement élémentaire connu sous le nom d'*arithmétique de Peano*), des procédures logiques qui suffiraient à démontrer la consistance d'autres systèmes formels plus sophistiqués, permettant de raisonner de manière formelle sur des ensembles infinis très « grands » et dont la consistance pouvait ne pas être immédiatement évidente. D'un point de vue formaliste, cette démonstration de la consistance d'un système formel F justifiait du moins le recours aux raisonnements autorisés par F. Les démonstrations des théorèmes mathématiques pourraient alors s'obtenir en utilisant de manière cohérente des ensembles infinis, et l'on pourrait peut-être se dispenser de s'interroger sur la « signification » réelle de tels ensembles. En outre, si l'on parvenait également à montrer qu'un tel F était complet, on pourrait alors raisonnablement penser qu'il englobait réellement *toutes* les procédures mathématiques permises ; ainsi, en un sens, on pourrait considérer que F représentait *réellement* la formulation complète des mathématiques du domaine en question.

Mais en 1930, Gödel mit un terme au rêve formaliste ! Il démontra qu'aucun système formel $\mathbb{F}$ — suffisamment puissant pour contenir une formulation des énoncés de l'arithmétique ordinaire et de la logique standard — ne peut être à la fois consistant (dans un certain sens « fort » que je décrirai à la prochaine section) et complet. Ainsi, le théorème de Gödel s'appliquait aux systèmes $\mathbb{F}$ dans lesquels des assertions arithmétiques telles que le théorème de Lagrange et la conjecture de Goldbach (*cf.* §2.3) admettaient une formulation sous forme d'énoncés mathématiques.

Dans les discussions qui vont suivre, je m'intéresserai uniquement aux systèmes formels suffisamment vastes pour contenir les opérations arithmétiques nécessaires à la formulation du théorème de Gödel (et, si besoin, pour contenir également les opérations de toute machine de Turing ; voir plus bas). Lorsque je parlerai d'un système formel $\mathbb{F}$, je supposerai implicitement qu'il est suffisamment vaste pour contenir ces opérations. Cette hypothèse ne limitera pas de manière essentielle la validité de mes discussions. (Toutefois, par souci de clarté, il m'arrivera, en fonction du contexte, de rappeler cette hypothèse en ajoutant l'expression « suffisamment vaste » — ou une expression analogue.)

2.8 La condition d'ω-consistance

La forme la plus familière du théorème de Gödel affirme qu'un système formel $\mathbb{F}$ suffisamment vaste ne peut être en même temps complet et consistant. Ce n'est pas tout à fait le fameux « théorème d'incomplétude » originellement présenté par Gödel lors du colloque de Königsberg (*cf.* §2.1 et §2.7), mais une version légèrement plus forte obtenue quelques années plus tard par le logicien américain J. Barkley Rosser (1936). La version initialement présentée par Gödel revenait à montrer que $\mathbb{F}$ ne peut être en même temps complet et *ω-consistant*. La condition d'ω-consistance est légèrement plus forte que celle de consistance ordinaire. Pour voir ce qu'elle recouvre, nous avons besoin d'introduire quelques notations supplémentaires. Un système formel $\mathbb{F}$ contient certains symboles désignant des opérations logiques. Il y a un symbole indiquant la *négation*, *i.e.* « non », que l'on peut désigner par « ~ ». Ainsi, si Q est une proposition exprimable dans $\mathbb{F}$, alors $\sim Q$ désigne « non Q ». Il y a également un symbole qui dit « pour tout [entier naturel] », appelé le *quantificateur universel* et que l'on peut désigner par « $\forall$ ». Si $P(n)$ est une proposition qui dépend de l'entier naturel n (P est alors appelée *fonction propositionnelle*), la suite de symboles $\forall n[P(n)]$ traduit l'énoncé « pour tout entier naturel n, $P(n)$ est vrai ». Un exemple de $P(n)$ est « n peut s'exprimer comme une somme de trois carrés » ; $\forall n[P(n)]$ signifie alors « tout entier naturel est la somme de trois carrés » — ce qui, en l'occurrence, est faux (mais vrai si « trois » est remplacé par « quatre »). Il y a de nombreuses façons d'associer ces symboles ; en particulier, la chaîne de symboles

$$\sim\forall n[P(n)]$$

exprime la *négation* de la validité de $P(n)$ pour tout entier naturel n.

L'ω-consistance affirme que si $\sim\forall n[P(n)]$ est démontrable par les méthodes de $\mathbb{F}$, alors les énoncés

$$P(0),\ P(1),\ P(2),\ P(3),\ P(4),\ \ldots$$

ne sont pas tous démontrables dans $\mathbb{F}$. Il en résulte que si $\mathbb{F}$ n'était pas ω-consistant, on pourrait avoir cette situation anormale où, pour un certain P, on pourrait démontrer chacun des $P(0)$, $P(1)$, $P(2)$, $P(3)$, $P(4)$, ..., bien que l'affirmation selon laquelle *certaines* de ces fonctions propositionnelles sont fausses serait *également* démontrable ! Aucun système formel fiable ne peut admettre une telle situation. Si $\mathbb{F}$ est *sûr*, il est certainement ω-consistant.

Dans ce livre, j'utiliserai les notations « $G(\mathbb{F})$ » et « $\Omega(\mathbb{F})$ » pour désigner respectivement les assertions « le système formel $\mathbb{F}$ est consistant » et « le système formel $\mathbb{F}$ est ω-consistant ». En fait, si $\mathbb{F}$ est suffisamment vaste, $G(\mathbb{F})$ et $\Omega(\mathbb{F})$ peuvent elles-mêmes être formulées à l'aide des opérations de $\mathbb{F}$. Le fameux théorème d'incomplétude de Gödel dit que $G(\mathbb{F})$ *n'est pas un théorème* de $\mathbb{F}$ (*i.e.* n'est pas démontrable à l'aide des procédures autorisées par $\mathbb{F}$), et $\Omega(\mathbb{F})$ non plus — à condition que $\mathbb{F}$ soit effectivement ω-consistant ! La version légèrement plus forte du théorème de Gödel obtenue par Rosser dit que si $\mathbb{F}$ est consistant, alors $\sim G(\mathbb{F})$ n'est pas non plus un théorème de $\mathbb{F}$. Dans le reste de ce chapitre, je formulerai généralement mes arguments en termes de l'expression plus familière $G(\mathbb{F})$ plutôt qu'en termes de $\Omega(\mathbb{F})$, bien que ces arguments restent essentiellement inchangés que l'on prenne l'une ou l'autre de ces deux assertions. (Pour les arguments plus explicites du chapitre 3, il sera cependant parfois plus approprié d'utiliser $\Omega(\mathbb{F})$.)

Dans la plupart de mes discussions, je ne me soucierai pas d'opérer une distinction bien nette entre la consistance et l'ω-consistance, mais la version du théorème de Gödel que j'ai présentée à la section 2.5 est essentiellement celle qui affirme que si $\mathbb{F}$ est consistant, il ne peut alors être complet, car il ne peut établir $G(\mathbb{F})$ en tant que théorème. Je ne tenterai pas de démontrer cela ici (mais voyez Kleene 1952). En fait, pour que cette forme du théorème de Gödel soit réductible au raisonnement que j'ai formulé, $\mathbb{F}$ doit contenir un peu plus que « l'arithmétique et la logique ordinaire ». Il doit être suffisamment vaste pour inclure les actions de toute *machine de Turing*. Ainsi, parmi les énoncés que l'on peut formuler correctement en utilisant les symboles du système $\mathbb{F}$ doivent figurer ceux du type « lorsqu'elle agit sur l'entier naturel n, telle machine de Turing donne l'entier naturel p ». En fait, il existe un théorème (*cf.* Kleene 1952, chapitres 11 et 13) qui affirme que c'est automatiquement le cas si, outre les opérations ordinaires de l'arithmétique, $\mathbb{F}$ contient également l'opération (appelée μ-opération) « trouver le plus petit entier naturel possédant telle propriété arithmétique ». Rappelons que dans notre premier exemple de calcul, **(A)**, notre procédure trouvait effectivement le *plus petit* entier naturel qui n'est pas la somme de trois carrés. D'une manière générale, il faut que les algorithmes puissent accomplir des choses de ce genre.

De fait, c'est *cela* qui conduit également à la possibilité de rencontrer des calculs qui ne se terminent pas, tel par exemple **(B)** où l'on tente de trouver le plus petit entier naturel qui ne soit pas la somme de *quatre* carrés, alors que ce nombre n'existe pas.

2.9 Systèmes formels et démonstration algorithmique

Dans l'argumentation de Gödel-Turing exposée à la section 2.5, je n'ai parlé que de « calculs », sans faire référence à des « systèmes formels ». Il existe cependant un lien très étroit entre ces deux concepts. L'une des propriétés essentielles d'un système formel $\mathbb{F}$ est qu'il contient une procédure algorithmique (à base de calcul) F permettant de *vérifier* si les règles de $\mathbb{F}$ ont été correctement appliquées. Si une proposition P est VRAIE selon les règles de $\mathbb{F}$, l'algorithme F peut également vérifier qu'elle est vraie (par exemple en examinant toutes les suites possibles de chaînes de symboles appartenant à l'« alphabet » du système $\mathbb{F}$, et en s'arrêtant lorsqu'elle rencontre la proposition P en question, toutes les étapes de cet examen se déroulant selon les règles du système $\mathbb{F}$).

Réciproquement, si E est une procédure de calcul *donnée*, destinée à vérifier certains énoncés mathématiques, on peut alors construire un système formel $\mathbb{E}$ dans lequel toutes les vérités que l'on peut obtenir à partir de la procédure E sont effectivement VRAIES. Cette construction exige cependant une petite précaution : un système formel contient normalement les opérations logiques standard, tandis que la procédure E peut ne pas être suffisamment vaste pour intégrer directement ces opérations. Si tel est le cas, il convient alors d'adjoindre à E ces opérations logiques lors de la construction de $\mathbb{E}$, de sorte que les propositions VRAIES de $\mathbb{E}$ soient non seulement les énoncés que l'on peut directement obtenir par la procédure E, mais aussi ceux qui sont des conséquences logiques élémentaires de ces mêmes énoncés. $\mathbb{E}$ n'est alors pas strictement équivalent à E, mais est légèrement plus puissant.

(Ces opérations logiques sont simplement des choses comme « si P & Q, alors P » ; « si P et $P \Rightarrow Q$, alors Q » ; « si $\forall x[P(x)]$, alors $P(n)$ » ; « si $\sim\forall x[P(x)]$, alors $\exists x[\sim P(x)]$ » ; etc. Ici, les symboles « & », « $\Rightarrow$ », « $\forall$ », « $\exists$ », « $\sim$ » signifient respectivement « et », « entraîne », « pour tout [entier naturel] », « il existe un [entier naturel] », « non ». Il peut également y avoir d'autres symboles de ce type.)

Pour construire $\mathbb{E}$ à partir de E, on peut partir d'un système formel $\mathbb{L}$ très fondamental (et manifestement consistant) traduisant simplement ces règles primitives d'inférence logique — tel le système connu sous le nom de *calcul des prédicats* (Kleene 1952) qui fait justement cela. On construit alors $\mathbb{E}$ en adjoignant E à $\mathbb{L}$ sous forme d'axiomes et de règles de procédures additionnels, puis en considérant comme VRAIE toute proposition P obtenue par la

procédure *E*. Cela n'est cependant pas facile à réaliser en pratique. Si *E* est simplement une spécification de machine de Turing, on peut avoir à adjoindre à $\mathbb{L}$, dans le cadre de son alphabet et de ses règles de procédure, toutes les notations et opérations de machine de Turing avant de pouvoir adjoindre *E* lui-même sous forme, en fait, d'axiome additionnel. (Voir la fin de la section 2.8 ; pour plus de détails, voir Kleene 1952.)

Peu importe en fait pour notre objectif que le système $\mathbb{E}$ ainsi construit contienne parfois des propositions VRAIES autres que celles directement accessibles par *E* (les règles logiques primitives de $\mathbb{L}$ n'étant elles-mêmes pas nécessairement représentées dans le cadre de la procédure *E*). À la section 2.5, nous avons considéré un algorithme hypothétique *A* censé englober toutes les procédures (connues ou connaissables) dont disposent les mathématiciens pour vérifier que des calculs ne s'arrêtent pas. Tout algorithme de ce type *doit* certainement contenir, entre autres choses, toutes les opérations fondamentales de l'inférence logique élémentaire. Dans les discussions qui vont suivre, je supposerai donc que *A* contient effectivement toutes ces opérations.

Pour les besoins de mon raisonnement, les algorithmes (*i.e.* les procédures de calcul) et les systèmes formels sont donc fondamentalement *équivalents* en tant que procédures permettant de juger de la vérité des énoncés mathématiques. Ainsi, bien que l'argumentation que j'ai donnée à la section 2.5 soit uniquement formulée en termes de calculs, elle s'applique également aux systèmes formels en général. Rappelons que ce raisonnement envisage une liste de tous les calculs (de toutes les actions de machines de Turing) $C_q(n)$. Pour qu'il puisse s'appliquer en détail à un système formel $\mathbb{F}$, il faut donc que $\mathbb{F}$ soit suffisamment vaste pour englober les actions de toutes les machines de Turing. La procédure algorithmique *A* qui permet d'établir que certains calculs ne s'arrêtent pas peut alors être intégrée dans les règles de $\mathbb{F}$, de sorte que les calculs pour lesquels on peut établir, à l'aide de $\mathbb{F}$, qu'ils ne s'arrêtent pas sont les mêmes que ceux pour lesquels cette même propriété peut s'établir à l'aide de *A*.

Quel lien unit le raisonnement originel présenté par Gödel à Königsberg et celui que j'ai exposé à la section 2.5 ? Voici, sans entrer dans les détails, en quoi il consiste. Ma procédure algorithmique *A* joue le rôle du système formel $\mathbb{F}$ considéré dans le théorème originel de Gödel :

$$\text{algorithme } A \leftrightarrow \text{règles de } \mathbb{F}.$$

La proposition « $C_k(k)$ ne s'arrête pas » obtenue à la section 2.5, qui est inaccessible par la procédure *A* mais dont on reconnaît la vérité dès que l'on pense que *A* est sûre, joue le rôle de la proposition $G(\mathbb{F})$ présentée par Gödel à Königsberg et qui affirme que $\mathbb{F}$ est consistant :

$$\text{l'énoncé « } C_k(k) \text{ ne s'arrête pas » } \leftrightarrow \text{ l'affirmation « } \mathbb{F} \text{ est consistant »}$$

Cela aide peut-être à comprendre comment la croyance dans la sûreté d'une procédure — telle que *A* — peut nous conduire à une autre procédure qui est hors de portée de la procédure de départ mais dont nous avons la conviction qu'elle est, *elle aussi*, sûre. Car si nous croyons que les procédures d'un système

formel $\mathbb{F}$ sont sûres — *i.e.* qu'elles nous permettent d'obtenir uniquement des vérités (et non des faussetés) mathématiques, de sorte que si l'on déduit qu'une proposition *P* est VRAIE, elle *est* nécessairement réellement *vraie* —, nous sommes alors également conduits à croire que $\mathbb{F}$ est ω-consistant. Si « VRAI » implique « vrai » et « FAUX » implique « faux » — comme c'est probablement le cas pour tout système formel $\mathbb{F}$ sûr —, on a alors certainement :

si « *P*(*n*) est vraie pour tous les entiers naturels *n* » est FAUX,
alors *P*(0), *P*(1), *P*(2), *P*(3), *P*(4), ... ne sont pas toutes VRAIES

ce qui est justement ce qu'affirme l'ω-consitance.

La croyance dans la sûreté de $\mathbb{F}$ entraîne la croyance non seulement en son ω-consistance, mais aussi en sa consistance. Car si « VRAI » implique « vrai » et « FAUX » implique « faux », on a alors :

aucune *P* ne peut être *à la fois* VRAIE *et* FAUSSE,

ce qui est précisément ce qu'affirme la consistance. En fait, pour nombre de systèmes, la distinction entre ω-consistance et consistance disparaît. Par souci de simplicité, je ne ferai généralement pas, dans la suite de ce chapitre, de distinction entre ces deux types de consistance et je parlerai simplement de « consistance ». Ce que Gödel et Rosser ont démontré est que la consistance d'un système formel (suffisamment vaste) échappe au pouvoir démonstratif du système formel lui-même. Si le premier théorème de Gödel (exposé à Königsberg) s'appuyait sur l'ω-consistance, la version plus familière qu'il livra par la suite invoque seulement la consistance ordinaire.

L'intérêt du théorème de Gödel pour notre argumentation est qu'il montre comment on sort du cadre d'un ensemble donné de règles de calcul, que l'on croit sûres, pour obtenir une nouvelle règle, non incluse dans les premières et qui nous apparaît elle aussi nécessairement sûre, à savoir celle qui affirme la *consistance* des règles initiales. Le point essentiel pour notre raisonnement est celui-ci :

la croyance dans la *sûreté* entraîne la croyance dans la *consistance.*

Nous n'avons pas le droit d'utiliser les règles d'un système formel $\mathbb{F}$, et de croire que les résultats que nous dérivons de ce système sont réellement *vrais,* si nous ne croyons pas également en la consistance de ce système formel. (Par exemple, si $\mathbb{F}$ était inconsistant, on pourrait alors démontrer que la proposition « 1 = 2 » est VRAIE, ce qui est loin d'être vrai !) Ainsi, si nous sommes convaincus que nous faisons réellement des mathématiques lorsque nous utilisons un système formel $\mathbb{F}$, nous devons alors accepter également l'éventualité d'un raisonnement sortant des limites du système $\mathbb{F}$, quoi que puisse être ce système.

2.10 Quelques autres objections techniques soulevées contre $\mathcal{C}$

Poursuivons maintenant notre examen des diverses objections mathématiques que l'on oppose à mon utilisation de l'argumentation de Gödel-Turing. Nombre de ces objections sont intimement liées entre elles, mais par souci de clarté, je les formulerai séparément.

Q10. La vérité mathématique est-elle absolue ? Nous avons déjà noté l'absence de consensus sur la vérité absolue des énoncés liés aux ensembles infinis. Pouvons-nous dès lors accepter des raisonnements reposant sur un vague concept de « vérité mathématique » qui semble contraster avec celui clairement défini de VÉRITÉ formelle ?

Dans le cas d'un système formel $\mathbb{F}$ lié à la théorie générale des ensembles, la « vérité » ou la « fausseté » d'un énoncé n'a effectivement pas toujours clairement un sens absolu — ce qui peut mettre en question le concept même de « sûreté » d'un système formel tel que $\mathbb{F}$. Un résultat démontré par Gödel (1940) et Cohen (1966) offre un exemple célèbre permettant de comprendre ce genre de problème. Ce résultat dit que les énoncés mathématiques connus sous le nom d'*hypothèse du continu* de Cantor et d'*axiome du choix* sont indépendants des axiomes de *Zermelo-Fraenkel* de la théorie des ensembles — axiomes constituant un système formel standard que je désignerai par $\mathbb{ZF}$. (L'axiome du choix affirme que pour toute collection d'ensembles non vides, il existe un autre ensemble qui contient exactement un élément de chaque ensemble de la collection[1]. L'hypothèse du continu de Cantor affirme que le nombre de sous-ensembles de l'ensemble des entiers naturels — qui est égal au nombre de nombres *réels* — est l'infini immédiatement supérieur au nombre d'entiers naturels lui-même[2]. Il n'est pas nécessaire que le lecteur se pénètre du sens de ces deux affirmations. Pour ma part, je n'aurai pas besoin ici d'exposer en détail les axiomes et règles de procédure de $\mathbb{ZF}$.) Certains mathématiciens soutiennent que $\mathbb{ZF}$ englobe toutes les formes de raisonnement mathématique nécessaires aux mathématiques ordinaires. Certains avancent même qu'un raisonnement mathématique acceptable est précisément un raisonnement qui peut, en principe, être formulé et démontré dans $\mathbb{ZF}$. (Voir plus loin, sous **Q14**, l'impact de ce point de vue sur l'argumentation gödelienne.) Ces mathématiciens affirment donc que les énoncés mathématiques qui sont, respectivement, VRAIS, FAUX et INDÉCIDABLES dans $\mathbb{ZF}$ sont précisément les énoncés dont on peut, en principe, établir mathématiquement la vérité, la fausseté et l'indécidabilité. Selon ces mêmes mathématiciens, l'axiome du choix et l'hypothèse du continu sont mathématiquement indécidables (ainsi que le montre, affirment-ils, le résultat de Gödel-Cohen) ; ils vont même jusqu'à dire que la vérité ou la fausseté de ces deux énoncés mathématiques est une pure affaire de convention.

Il se trouve que ces apparentes incertitudes sur le caractère absolu de la vérité mathématique n'affectent en rien les conclusions que nous avons déduites de

l'argumentation de Gödel-Turing. Cette dernière porte en effet sur une classe de problèmes mathématiques dont la portée est bien plus restreinte que celle de problèmes qui, tels l'axiome du choix ou l'hypothèse du continu, concernent des ensembles infinis non constructibles. Seuls nous intéresseront ici des énoncés de la forme :

« tel calcul ne se termine pas »,

où le calcul en question peut être précisément spécifié en termes d'actions de machines de Turing. Les logiciens désignent ce genre d'énoncé du nom technique d'*énoncé* Π_1 (ou plus exactement d'énoncé Π_1^0). Pour tout système formel $\mathbb{F}$, $G(\mathbb{F})$ est un énoncé Π_1, mais $\Omega(\mathbb{F})$ n'en est pas un (*cf.* §2.8). Il semble peu raisonnable de douter que la vérité ou la fausseté d'un énoncé Π_1 quelconque ait un caractère *absolu*, indépendant de la position que l'on choisit d'adopter à l'égard des questions liées aux ensembles infinis non constructibles — intervenant par exemple dans l'axiome du choix et l'hypothèse du continu. (Nous verrons toutefois dans un instant que le type de raisonnement dont on reconnaît qu'il fournit des *démonstrations* convaincantes d'énoncés Π_1 peut quant à lui dépendre du point de vue adopté à l'égard des ensembles infinis non constructibles ; *cf.* plus loin **Q11**.) Il semble clair qu'à l'exception de la position extrême de certains intuitionnistes (*cf.* réponse à **Q9**), le seul problème que puisse raisonnablement poser le caractère absolu de la vérité de tels énoncés soit que l'exécution de certains calculs qui se terminent soit si lente qu'elle ne pourrait être menée à son terme en un temps correspondant, par exemple, à la durée de vie de l'Univers ; ou encore que la spécification du calcul lui-même (bien que finie) mette en jeu un si grand nombre de symboles qu'elle ne pourrait jamais être entièrement écrite. Ces questions ont toutefois été pleinement examinées dans la réponse à **Q8**, et nous avons vu qu'elles n'affectent pas la conclusion $\mathscr{G}$. Rappelons également, comme il a été dit dans la réponse à **Q9**, que la position intuitionniste ne permet pas non plus d'échapper à cette conclusion.

En outre, le concept (très limité) de vérité mathématique que j'utilise dans l'argumentation de Gödel-Turing est en fait tout aussi bien défini que les concepts VRAI, FAUX et INDÉCIDABLE associés à tout système formel $\mathbb{F}$. Nous l'avons dit à la section 2.9, il existe un *algorithme* F qui est équivalent à $\mathbb{F}$. Si F opère sur une proposition P (formulable dans $\mathbb{F}$), cet algorithme s'arrête exactement lorsque P est démontrable à l'aide des règles de $\mathbb{F}$, autrement dit lorsque P est VRAIE. De même, P est FAUX uniquement si F s'arrête lorsqu'on l'applique à $\sim P$; et P est INDÉCIDABLE si aucun de ces deux calculs ne se termine. Le problème de la VÉRITÉ, de la FAUSSETÉ ou de l'INDÉCIDABILITÉ d'un énoncé mathématique P est par nature identique à celui de la vérité de l'arrêt ou du non-arrêt de certains calculs — *i.e.* de la fausseté ou de la vérité de certains énoncés Π_1 —, et c'est là tout ce dont nous avons besoin dans notre argumentation de Gödel-Turing.

Q11. Il existe certains énoncés Π_1, démontrables à l'aide de la théorie des ensembles infinis, mais dont on ne connaît aucune démonstration recourant aux méthodes « finies » standard. Cela ne trahit-il pas la

subjectivité avec laquelle les mathématiciens décident de la vérité ou de la fausseté d'énoncés même aussi bien définis ? Des mathématiciens professant des opinions différentes envers la théorie des ensembles ne se fonderaient-ils pas alors sur des critères non équivalents pour apprécier la vérité mathématique d'énoncés Π_1 ?

Cela pourrait effectivement être une objection de taille à l'égard des conclusions que je tire du raisonnement de Gödel(-Turing), et peut-être ne lui ai-je pas accordé une attention suffisante lors de la brève discussion que je lui ai consacrée dans EOLP. Curieusement toutefois, il semble que je sois le seul à l'avoir considérée — du moins, personne ne me l'a signalée ! Ici comme dans EOLP (p. 453-455), j'ai formulé l'argumentation de Gödel(-Turing) par rapport à ce que les « mathématiciens » ou la « communauté mathématique » peuvent apprécier à l'aide de la raison et de l'intuition. L'avantage d'une telle formulation — par rapport à une formulation fondée sur ce qu'un individu *particulier* peut apprécier en se servant de sa raison ou de son intuition — est qu'elle permet d'échapper à certaines critiques fréquemment soulevées contre la version de l'argumentation gödelienne donnée par Lucas (1961). Selon diverses personnes[3], « Lucas lui-même », par exemple, ne pourrait connaître son propre algorithme. (Certaines personnes ont même opposé une objection identique à ma propre présentation[4] — oubliant apparemment que mon argumentation n'a pas le moins du monde cette formulation « personnelle » !) L'avantage d'une formulation basée sur le raisonnement et l'intuition des « mathématiciens » ou de la « communauté mathématique » est qu'elle n'est pas concernée par la suggestion selon laquelle des individus différents auraient, en fonction chacun de leur propre algorithme inconnaissable, des idées différentes sur la notion de vérité mathématique. Il est bien plus difficile d'accepter que la compréhension sur laquelle s'accordent l'ensemble des mathématiciens soit fondée sur un quelconque algorithme inconnaissable, que d'accepter qu'un algorithme inconnaissable personnel préside à la compréhension de chaque mathématicien individuel. La question soulevée par **Q11** est que cette compréhension commune ne serait pas si universelle et impersonnelle que je l'ai supposé.

Il est vrai qu'il *existe* des énoncés Π_1 dont les seules démonstrations connues dépendent, comme le remarque **Q11**, d'une utilisation appropriée de la théorie des ensembles infinis. Ces énoncés surviennent par exemple dans le codage arithmétique d'un énoncé tel que « les axiomes de $\mathbb{F}$ sont consistants », où $\mathbb{F}$ est un système formel mettant en jeu la manipulation d'ensembles infinis dont l'existence même pourrait être controversée. Un mathématicien qui croit en l'*existence* réelle d'un ensemble **S** infini non constructible conclura que $\mathbb{F}$ est en fait consistant, tandis qu'un autre mathématicien qui ne croit pas en **S** n'aura pas nécessairement une telle foi en $\mathbb{F}$. Ainsi, même en nous restreignant au problème de l'arrêt de l'action des machines de Turing (*i.e.* à la fausseté ou à la vérité des énoncés Π_1), nous ne pouvons ignorer le problème de la subjectivité des *convictions* concernant, par exemple, l'existence d'un grand ensemble infini **S** non constructible. Si, pour

vérifier la vérité de certains énoncés Π_1, différents mathématiciens emploient des « algorithmes personnels » *non équivalents*, il est alors peut-être malhonnête de ma part de me référer simplement aux « mathématiciens » ou à la « communauté mathématique ».

J'imagine que, strictement parlant, cela peut paraître effectivement légèrement malhonnête ; aussi le lecteur qui partage ce sentiment préférera-t-il peut-être à la place de $\mathscr{G}$ la conclusion suivante :

$\mathscr{G}^*$ Aucun mathématicien individuel ne vérifie une vérité mathématique en recourant uniquement à un algorithme qu'il sait sûr.

Si l'on présente les choses ainsi, mes arguments restent valides — mais certains des derniers perdent, je pense, une grande partie de leur force. De plus, avec la version $\mathscr{G}^*$, ils prennent à mon avis une orientation qui ne présente guère d'intérêt dans la mesure où ils portent davantage sur les mécanismes particuliers gouvernant les actions d'individus particuliers que sur les principes sous-tendant les actions de l'ensemble de la communauté humaine. Je m'intéresse moins, à ce stade, aux diverses approches suivies par différents mathématiciens qu'au caractère *universel* de notre compréhension et de notre perception des mathématiques.

Sommes-nous *réellement* obligés d'adopter la version $\mathscr{G}^*$? Les jugements des mathématiciens sont-ils effectivement subjectifs au point de pouvoir, *en principe*, diverger sur la vérité d'un énoncé Π_1 ? (Bien entendu, le raisonnement établissant cet énoncé peut s'avérer trop long ou trop complexe pour être suivi par l'un ou l'autre des mathématiciens — *cf.* **Q12** plus loin —, de sorte que leurs jugements peuvent éventuellement différer *en pratique*. Mais là n'est pas la question. Nous nous intéressons uniquement ici aux problèmes de *principe*.) En fait, une démonstration mathématique n'est pas aussi subjective qu'il y paraît. En dépit du fait que différents mathématiciens puissent avoir des points de vue passablement différents sur ce qui, au niveau des fondements mathématiques, possède un indéniable caractère de vérité, ils s'accordent généralement sur la valeur des démonstrations ou réfutations d'énoncés Π_1 clairement définis. Un énoncé Π_1 qui, par exemple, affirme la consistance d'un système $\mathbb{F}$ n'est pas normalement considéré comme convenablement démontré si tout ce sur quoi l'on puisse se fonder est l'existence d'un ensemble infini **S** controversé. Une formulation plus acceptable de ce qui a en fait été démontré serait : « Si **S** existe, alors $\mathbb{F}$ est consistant, et dans ce cas, cet énoncé Π_1 est vrai. »

Il peut cependant y avoir des exceptions, par exemple lorsqu'un mathématicien considère comme « évidente » l'existence d'un ensemble infini **S** non constructible — ou du moins considère que l'hypothèse de cette existence ne débouche sur aucune contradiction — tandis qu'un autre ne partage pas une telle conviction. En matière de *fondements*, les mathématiciens semblent effectivement parfois s'enfermer dans des débats sans fin. En principe, cela pourrait les amener à être dans l'impossibilité de communiquer de manière convaincante leurs démonstrations, même celles concernant les énoncés Π_1. Peut-être chaque mathématicien a-t-il sa propre perception de la vérité des énoncés sur

les ensembles infinis non constructibles. Il est vrai qu'ils *affirment* souvent un tel individualisme. Mais je pense que ces différences sont fondamentalement semblables à celles qui peuvent marquer leurs *intuitions* sur l'éventuelle vérité d'une proposition mathématique ordinaire. Ces intuitions correspondent simplement à des opinions provisoires. Tant qu'ils ne disposent pas d'une démonstration — ou d'une réfutation — convaincante, ils peuvent s'opposer sur ce qu'ils estiment être la vérité ; mais une fois que l'un d'entre eux parvient à une telle démonstration, elle convainc (en principe) les autres mathématiciens. En ce qui concerne les problèmes portant sur les fondements des mathématiques, on manque effectivement de démonstrations convaincantes. Peut-être n'en trouvera-t-on jamais. Peut-être aussi *ne peut-on* en trouver parce qu'elles n'existent pas et que l'on doit admettre qu'il existe simplement sur ce plan plusieurs points de vue également valides.

Il faut cependant souligner que l'éventualité d'un point de vue *erroné* sur les énoncés Π_1 — j'entends par là un point de vue qui entraîne des conclusions incorrectes sur la validité de certains de ces énoncés — *ne nous concerne pas* ici. Il arrive probablement qu'en s'en tenant aux faits, les mathématiciens utilisent de mauvaises « intuitions » — en particulier des *algorithmes non sûrs* —, mais cette situation n'affecte pas la présente discussion dans la mesure où elle est *en accord* avec $\mathcal{G}$. Cette possibilité sera examinée en détail à la section 3.4. Le problème qui nous préoccupe ici n'est donc pas de savoir si les mathématiciens peuvent avoir des points de vue *divergents*, mais s'il pourrait, en principe, exister un point de vue qui serait *plus pertinent* que les autres. Chaque point de vue serait parfaitement fondé et entraînerait des conséquences parfaitement cohérentes au niveau de la vérité des énoncés Π_1, mais l'un d'entre eux, contrairement aux autres, permettrait à ses adeptes de vérifier que certains calculs ne se terminent pas. Il y aurait donc alors divers degrés d'intuition mathématique.

Je ne pense pas que cette possibilité constitue une menace réelle pour la formulation $\mathcal{G}$. Si, à l'égard des ensembles infinis, les mathématiciens peuvent raisonnablement adopter des points de vue différents, ceux-ci ne sont cependant pas *véritablement* nombreux — il n'y en a probablement pas plus de quatre ou cinq. Les seules différences importantes portent par exemple sur l'axiome du choix (évoqué sous **Q10**), que nombre de mathématiciens considèrent comme « évident » tandis que d'autres n'acceptent pas la non-constructibilité qui lui est associée. Curieusement, les diverses attitudes envers l'axiome du choix *ne débouchent pas* directement sur un énoncé Π_1 à la validité contestée. En effet, qu'on le considère ou non comme « vrai », cet axiome n'entraîne aucune contradiction avec les axiomes $\mathbb{ZF}$, ainsi que le montre le théorème de Gödel-Cohen (mentionné sous **Q10**). Il pourrait y avoir cependant *d'autres* axiomes contestés pour lesquels on ne connaîtrait aucun théorème analogue. Mais habituellement, lorsqu'il s'agit d'accepter — ou de rejeter — un axiome de la théorie des ensembles — appelons-le Q — les énoncés mathématiques prennent généralement la forme : « Si l'on admet l'axiome Q, il s'ensuit que... » Il n'y a aucun lieu de débattre sur ces énoncés. L'axiome du choix semble faire exception dans la mesure où il est fréquemment admis sans être

explicitement mentionné ; mais il ne constitue apparemment pas une objection contre la formulation générale et impersonnelle que j'ai donnée de $\mathscr{G}$, à condition de limiter $\mathscr{G}$ aux énoncés Π_1 et donc d'écrire :

$\mathscr{G}^{**}$ Ce n'est pas en utilisant un algorithme qu'ils savent sûr que les mathématiciens humains établissent la vérité des énoncés Π_1,

ce qui est tout ce dont nous avons besoin.

Y a-t-il d'autres axiomes contestés — considérés comme « évidents » par certains mais mis en question par d'autres ? Je pense qu'à propos des hypothèses de la théorie des ensembles, il serait très exagéré de dire qu'il existe jusqu'à dix points de vue fondamentalement différents qui ne soient pas explicitement considérés comme des hypothèses. Mais acceptons ce nombre 10 et examinons les conséquences qu'il entraîne. Il signifierait qu'il y aurait essentiellement dix catégories de mathématiciens, classés selon les types de raisonnement — sur les ensembles infinis — qu'ils reconnaissent comme « manifestement » valides. On pourrait appeler ces diverses catégories de mathématiciens des mathématiciens d'échelon n, n prenant seulement quelques valeurs — pas plus d'une dizaine. (Plus l'échelon est élevé, plus le point de vue du mathématicien est puissant.) À la place de $\mathscr{G}^{**}$, nous aurions alors :

$\mathscr{G}^{***}$ Pour chaque n (n prenant un nombre très limité de valeurs), les mathématiciens humains d'échelon n ne vérifient pas la validité des énoncés Π_1 en s'appuyant uniquement sur un algorithme qu'ils savent sûr.

Cette conclusion résulte du fait que l'argumentation de Gödel(-Turing) s'applique séparément à chaque échelon. (Qu'il soit bien clair que le raisonnement de Gödel n'est par lui-même l'objet d'aucune controverse parmi les mathématiciens, de sorte que, quel que soit n, si les mathématiciens d'échelon n considèrent comme sûr l'hypothétique algorithme d'échelon n, l'argumentation de Gödel(-Turing) conduit à une contradiction.) Ainsi, pas plus qu'avec $\mathscr{G}$, le problème de la multiplicité des algorithmes que l'on considère comme sûrs — chacun de ces algorithmes étant propre à chaque mathématicien — ne se pose pas. En revanche, nous avons exclu qu'il puisse y avoir un très petit nombre d'algorithmes non équivalents, dont on ne pourrait prouver qu'ils sont sûrs, classés en fonction de leur fécondité, et donnant naissance à différentes « écoles de pensée ». Dans les discussions qui vont suivre, le rôle de la version $\mathscr{G}^{***}$ ne différera pas grandement de celui de $\mathscr{G}$ ou de $\mathscr{G}^{**}$; par souci de simplicité, je ne ferai aucune distinction entre elles et les désignerai collectivement par la seule lettre $\mathscr{G}$.

Q12. Quoi qu'ils puissent ou non *en principe* adopter comme points de vue, les mathématiciens ont, *en pratique*, des aptitudes très diverses à suivre un raisonnement. De même, ils s'aident manifestement d'intuitions tout aussi diverses pour faire des découvertes mathématiques.

Certes, c'est vrai, mais cela ne nous concerne pas vraiment ici. Je ne m'intéresse pas au détail des raisonnements particuliers qu'un mathématicien peut faire *en pratique*. Et je me préoccupe encore moins des raisonnements qu'un

mathématicien peut, en pratique, *découvrir*, ou des intuitions et des inspirations qui peuvent le conduire à ces découvertes. Le problème qui m'intéresse ici est de savoir quel type de raisonnement peut en principe être considéré comme valide par les mathématiciens.

Je tiens à préciser que dans les discussions précédentes, l'expression « en principe » a été utilisée intentionnellement. Supposons qu'un mathématicien dispose d'une démonstration — ou d'une réfutation — d'un énoncé Π_1. Les désaccords avec d'autres mathématiciens quant à la validité de cette démonstration ne peuvent se régler que si ceux-ci ont le temps, la patience, l'ouverture d'esprit, l'aptitude et la détermination nécessaires pour suivre entièrement, et comprendre avec précision, un raisonnement qui peut s'avérer long et subtil. En pratique, ces mathématiciens risquent fort de baisser les bras avant d'arriver à la fin du raisonnement. Mais ce genre de situation ne nous concerne pas ici. Car il semble évident que ce qu'un mathématicien peut *en principe* comprendre est identique (hormis dans le contexte des considérations développées sous **Q11**) à ce que peut comprendre un autre mathématicien — ou en fait toute autre personne qui pense. Le raisonnement peut être très long, les concepts mis en jeu obscurs ou subtils, mais il existe des raisons suffisamment convaincantes pour croire que rien de ce qui est accessible à l'entendement d'une personne n'est en principe inaccessible à l'entendement d'une autre. Cela s'applique également au cas où l'on recourt à un ordinateur pour suivre tous les détails de la partie purement opératoire d'une démonstration. Bien qu'on ne puisse parfois pas exiger d'un mathématicien humain qu'il parvienne à suivre tous les détails des calculs intervenant dans un raisonnement, il ne fait cependant aucun doute que ce même mathématicien humain est capable de comprendre et d'approuver les *étapes* de ce raisonnement.

En disant cela, je ne considère que la complexité d'un raisonnement mathématique, et non les questions fondamentales de principe qui pourraient opposer deux mathématiciens sur les types de raisonnement qu'ils sont prêts à accepter. Certes j'ai rencontré des mathématiciens affirmant que certains raisonnements dépassent leurs capacités intellectuelles : « Je sais, disent-ils, que je ne pourrai jamais comprendre cette chose, ou Untel, quels que soient mes efforts ; ce genre de raisonnement m'échappe totalement. » Face à une telle affirmation, il faut se demander si l'on est effectivement en présence d'un raisonnement qui est *en principe* inaccessible à la compréhension du mathématicien — *i.e.* analogue à ceux discutés sous **Q11** — ou si ce mathématicien *pourrait*, en faisant un effort suffisamment important et prolongé, comprendre les principes sous-jacents à ce raisonnement. Très souvent, on est dans ce deuxième cas. La situation la plus courante est en fait celle où c'est le style obscur de « Untel » qui est cause du désespoir de notre mathématicien, et non un quelconque principe fondamental échappant à son entendement ! Un bon exposé d'un sujet apparemment obscur peut faire des miracles.

Pour rendre ce point encore plus clair, je confesse qu'il m'arrive souvent d'assister à des séminaires de mathématiques au cours desquels je ne suis pas (ou n'essaie même pas de suivre) en détail les raisonnements présentés. Peut-être ai-je le sentiment que si je les étudiais en long et en large, je serais effecti-

vement capable de les comprendre — bien qu'il me faudrait sans doute m'aider de quelques lectures et de quelques explications orales pour combler certaines lacunes affectant ma formation et probablement aussi l'exposé lui-même. Mais je sais que je ne le ferai pas. Cela me demanderait trop de temps, de concentration et d'enthousiasme. Pourtant, il m'arrive d'accepter le résultat présenté pour toutes sortes de raisons « irrationnelles », telles que le fait qu'il me « semble » plausible, ou que le conférencier est connu pour son sérieux, ou encore que certaines personnes dans l'auditoire, dont je sais qu'elles sont bien plus expertes que moi sur le sujet, n'ont pas contesté ce résultat. Je peux bien sûr me tromper sur toute la ligne et le résultat peut se révéler inexact, ou être correct mais ne pas découler des arguments qui ont été produits. Ce sont là des questions de détail qui n'affectent pas le problème de principe que je pose ici. Soit le résultat est exact et correctement démontré, auquel cas je pourrais *en principe* suivre le raisonnement, soit ce raisonnement est erroné, mais cette situation — je l'ai dit plus haut — ne nous concerne pas ici (*cf.* §3.2 et §3.4). Les seules exceptions possibles seraient que la conférence traitât de certains aspects controversés de la théorie des ensembles infinis ou qu'elle dépendît d'une forme de raisonnement insolite et critiquable de certains points de vue mathématiques (ce qui, en soi, pourrait m'intriguer suffisamment pour que j'examine par la suite en détail le raisonnement en question). Ces situations exceptionnelles sont justement celles prises en compte plus haut sous **Q11**.

Dans la pratique, nombre de mathématiciens n'ont en fait pas de position claire à l'égard des principes mathématiques fondamentaux auxquels ils adhèrent. Mais comme il a été dit sous **Q11**, un mathématicien prudent qui ne sait s'il doit ou non accepter, disons, l'« axiome Q », énonce toujours ses résultats dépendant de Q sous la forme : « Si l'on admet l'axiome Q, il s'ensuit que... » Bien qu'appartenant à une race notoirement pédante, les mathématiciens oublient parfois ce type de précautions. Il leur arrive même de temps en temps de commettre des erreurs manifestes. Mais ces erreurs, si elles sont surtout des lapsus et n'ont rien à voir avec des principes intangibles, sont *rectifiables.* (Nous l'avons dit, l'utilisation par les mathématiciens d'un algorithme non sûr comme point de départ de leurs décisions est une éventualité qui sera examinée en détail aux sections 3.2 et 3.4. Cette possibilité étant *prise en compte* par $\mathcal{G}$, elle n'est pas au menu de la présente discussion.) Nous ne nous attarderons pas sur les erreurs rectifiables, car elles n'interviennent pas dans les situations de principe envisagées ici. En revanche, nous devrons examiner plus avant les éventuelles incertitudes affectant les convictions des mathématiciens.

Q13. Les mathématiciens n'ont pas d'opinions *définitivement* arrêtées sur la sûreté ou la consistance des systèmes formels qu'ils utilisent — quand ils n'en sont pas à se demander *sur quels* systèmes formels particuliers on devrait considérer qu'ils fondent leurs travaux. Leurs convictions ne s'estompent-elles simplement pas à mesure que diminue le lien entre ces systèmes formels et leurs intuitions et expériences immédiates ?

De fait, lors des discussions sur les fondements des mathématiques, il est rare de rencontrer un mathématicien aux opinions tranchées et absolument

cohérentes. En outre, avec l'expérience, la position des mathématiciens à l'égard de ce qu'ils considèrent comme des vérités inébranlables peut varier — si tant est qu'ils considèrent jamais *un* énoncé mathématique comme indiscutablement vrai. Peut-on être totalement certain, par exemple, que 1 est différent de 2 ? Existe-t-il vraiment une certitude humaine *absolue* ? Il faut bien cependant adopter une position. Une position raisonnable consisterait à décider qu'un *certain* ensemble de convictions et de principes est inattaquable puis à raisonner à partir de cet ensemble. Il se peut bien entendu que nombre de mathématiciens n'aient même pas une idée claire de ce qui leur semble incontestablement vrai. Dans ce cas, je leur demanderai d'adopter néanmoins une position, même si par la suite ils peuvent être amenés à la modifier. L'argumentation gödelienne montre que *quelle que soit* la position adoptée, on ne peut savoir si elle s'intègre dans les règles d'un système formel connaissable. Cela ne résulte pas du fait que cette position subit de continuelles modifications ; le corps de croyances suscité par *tout* système formel $\mathbb{F}$ (suffisamment vaste) s'étend nécessairement au-delà de ce que $\mathbb{F}$ peut accomplir. Toute position qui inclut la sûreté de $\mathbb{F}$ au nombre de ses convictions inattaquables doit également admettre la vérité de la proposition gödelienne* $G(\mathbb{F})$. La croyance dans $G(\mathbb{F})$ ne représente pas un changement de position ; cette croyance est déjà implicitement contenue dans la position originelle qui a reconnu la validité de $\mathbb{F}$ — même si le fait que l'on doive également accepter $G(\mathbb{F})$ a pu ne pas apparaître immédiatement.

Bien sûr, il y a toujours l'éventualité qu'une erreur ait pu se glisser dans les déductions que l'on tire des prémisses d'une position particulière. La simple *possibilité* d'avoir commis une telle erreur quelque part — même si en réalité on ne l'a pas commise — peut conduire à une perte de confiance envers ses propres conclusions. Mais ce phénomène-là ne nous concerne pas ici. Comme les erreurs réelles, cette erreur est « rectifiable ». En outre, plus on examine un raisonnement correctement mené, plus on se persuade de la validité de ses conclusions. Cette forme d'érosion de confiance se situant au niveau des sentiments qu'un mathématicien peut éprouver *en pratique* — et non en principe —, elle nous ramène à la discussion de **Q12**.

Le problème est maintenant de savoir si une telle érosion de confiance peut survenir *en principe*, de sorte qu'un mathématicien pourrait décider, par exemple, que la sûreté d'un système formel $\mathbb{F}$ est inattaquable, et considérer qu'un système formel plus puissant $\mathbb{F}^*$ ne serait que « presque certainement » sûr. Quoi que soit $\mathbb{F}$, il doit impérativement contenir les règles ordinaires de la logique et des opérations arithmétiques. Notre mathématicien, qui croit que $\mathbb{F}$ est sûr, doit alors également croire que $\mathbb{F}$ est consistant, et donc que la proposition de Gödel $G(\mathbb{F})$ est vraie. Ainsi, les déductions tirées du seul $\mathbb{F}$ ne peuvent représenter la totalité des convictions mathématiques du mathématicien, *quoi que puisse être* $\mathbb{F}$.

---

* Voir la section 2.8 pour la notation utilisée ici. Nous le verrons à la fin de cette discussion, $G(\mathbb{F})$ aurait pu être partout remplacé par $\Omega(\mathbb{F})$.

Mais peut-on considérer que la vérité de $G(\mathbb{F})$ est *inattaquable* dès lors que $\mathbb{F}$ est incontestablement sûr ? Je pense qu'on ne peut guère douter qu'il en est ainsi. C'est certainement le cas si l'on adhère au point de vue que nous avons adopté jusqu'ici sur la possibilité de suivre « en principe » un raisonnement mathématique. La seule véritable question concerne les détails du codage concret de l'affirmation « $\mathbb{F}$ est consistant » en un énoncé mathématique (un énoncé Π_1). L'*idée* sous-jacente est manifestement incontestable : si $\mathbb{F}$ est sûr, il est alors certainement consistant. (Car s'il ne l'était pas, il contiendrait par exemple l'assertion « 1 = 2 » et ne serait donc pas sûr.) En ce qui concerne le codage, il faut ici aussi faire une distinction entre sa faisabilité « en pratique » et sa faisabilité « en principe ». S'il n'est pas trop difficile de se convaincre que ce codage est possible en principe (même s'il faut parfois un certain temps pour s'assurer que le raisonnement ne contient pas d'« entourloupette »), il n'en va pas de même pour se persuader que le codage *concret* a été correctement effectué. Les détails de ce codage sont généralement quelque peu arbitraires et peuvent différer grandement d'une présentation à l'autre. Et il peut s'y glisser une erreur ou une coquille mineures qui, techniquement parlant, invalident la proposition particulière censée exprimer numériquement « $G(\mathbb{F})$ ».

J'espère qu'il est clair que l'éventualité de telles erreurs n'est pas fondamentalement ce qui interdit d'accepter que $G(\mathbb{F})$ est incontestablement vraie — j'entends ici la *vraie* proposition $G(\mathbb{F})$, et non celle que l'on aurait malencontreusement formulée à la suite d'erreurs ou de coquilles accidentelles. Cela me rappelle une anecdote sur le grand physicien américain Richard Feynman. Feynman expliquait une idée à un étudiant, mais ne parvenait pas à la formuler correctement. Lorsque l'étudiant lui exprima son étonnement, Feynman répondit : « N'écoutez pas ce que je dis, écoutez ce que je *veux dire* ! »*

Un codage possible consisterait à utiliser les spécifications des machines de Turing que j'ai données dans EOLP et à suivre exactement l'argumentation gödelienne présentée à la section 2.5 (dont l'appendice A fournit le codage explicite). Cela, toutefois, ne constituerait pas encore le codage explicite complet, car il faut également coder les règles de $\mathbb{F}$ dans le langage d'une machine de Turing — appelons-la $T_{\mathbb{F}}$. (Si une proposition P, formulable dans le langage de $\mathbb{F}$, est repérée par le nombre p, $T_{\mathbb{F}}$ vérifiera par exemple $T_{\mathbb{F}}(p) = 1$ chaque fois que P est un théorème, et ne s'arrêtera pas dans le cas contraire.) Bien sûr, tout cela laisse une large place à de potentielles erreurs techniques. Outre les difficultés pouvant survenir dans la construction de $T_{\mathbb{F}}$ à partir de $\mathbb{F}$ et dans la détermination de p à partir de P, je peux avoir moi-même commis une erreur dans ma spécification des machines de Turing — voire, si on utilise cette spécification pour calculer $C_k(k)$, dans le code donné à l'appendice A. Je ne pense pas qu'il y ait une erreur, mais la confiance que je place en moi-même n'est pas aussi grande que celle que j'accorde aux

* Je n'ai pu retrouver la source de cette anecdote. Mais comme me l'a fait remarquer Richard Jozsa, peu importe que je l'aie mal rapportée, puisque je peux lui appliquer son propre message !

spécifications originelles (mais plus complexes) données par Gödel lui-même. J'espère cependant qu'il est clair maintenant que l'éventualité de telles erreurs n'est pas une objection fondamentale. N'oubliez pas la réponse de Feynman !

Il me faut cependant mentionner un autre point technique à propos de mes propres spécifications. À la section 2.5, ma version de l'argumentation de Gödel(-Turing) invoque non pas la consitance de $\mathbb{F}$, mais le fait que l'algorithme A permettant de vérifier le non-arrêt de certains calculs (donc de vérifier la vérité d'énoncés Π_1) est sûr. Cela ne change toutefois rien, car nous avons vu que la sûreté de A entraîne la vérité de l'assertion selon laquelle $C_k(k)$ ne s'arrête pas —, de sorte que l'on peut utiliser cette assertion explicite — qui est aussi un énoncé Π_1 — à la place de $G(\mathbb{F})$. En outre, comme nous l'avons remarqué plus haut (*cf.* §2.8), l'argument dépend en fait de l'ω-consistance de $\mathbb{F}$, et non de sa consistance. Si $\mathbb{F}$ est sûr, il est alors manifestement non seulement ω-consistant, mais aussi consistant. Si $\mathbb{F}$ est sûr, ni $\Omega(\mathbb{F})$ ni $G(\mathbb{F})$ ne découlent des règles de $\mathbb{F}$ (cf. § 2.8), bien que ces énoncés Π_1 soient tous deux vrais.

En résumé, je pense qu'il est clair que quelle que soit l'« érosion » de confiance que peuvent subir les convictions d'un mathématicien passant de la croyance qu'un système formel $\mathbb{F}$ est sûr à la croyance que la proposition $G(\mathbb{F})$ (ou $\Omega(\mathbb{F})$) est vraie, elle résulte entièrement de l'éventualité qu'une erreur ait pu se glisser dans la formulation précise de la proposition « $G(\mathbb{F})$ ». (La même chose vaut pour $\Omega(\mathbb{F})$.) Ce point n'est pas véritablement pertinent pour la présente discussion, et il ne devrait y avoir aucune érosion de la croyance dans la version *théorique* de $G(\mathbb{F})$. Si $\mathbb{F}$ est indiscutablement sûr, *cette* proposition $G(\mathbb{F})$ est incontestablement vraie. Si « vrai » signifie « incontestablement vrai », les diverses formulations de la conclusion $\mathcal{G}$ (ou $\mathcal{G}^{**}$, ou $\mathcal{G}^{***}$) sont valables.

Q14. Le système $\mathbb{ZF}$ — ou une modification standard de $\mathbb{ZF}$ (appelons-la $\mathbb{ZF}^*$) — suffit certainement pour faire sérieusement des mathématiques. Pourquoi alors ne pas se borner à $\mathbb{ZF}$, accepter que sa consistance n'est pas démontrable et faire simplement des mathématiques avec ce système ?

Je pense que c'est là un point de vue très répandu parmi les mathématiciens — notamment ceux que n'intéressent pas particulièrement les fondements ou la philosophie de leur discipline. Ce n'est pas un point de vue déraisonnable pour les mathématiciens dont la préoccupation première est de faire efficacement des mathématiques (bien que ces mêmes mathématiciens expriment *de fait* très rarement leurs résultats dans le cadre des règles strictes d'un système tel que $\mathbb{ZF}$). Selon ce point de vue, on ne doit s'intéresser qu'à ce qu'on peut démontrer ou réfuter à l'intérieur d'un système formel donné (tel $\mathbb{ZF}$ ou l'une de ses modifications, $\mathbb{ZF}^*$). Les mathématiques deviennent alors une sorte de « jeu ». Appelons *jeu* $\mathbb{ZF}$ (ou jeu $\mathbb{ZF}^*$) le jeu qui consiste à jouer selon les règles propres du système $\mathbb{ZF}$ (ou $\mathbb{ZF}^*$). Ce point de vue est celui du *formaliste* pur,

dont l'intérêt se limite strictement à ce qui est VRAI et à ce qui est FAUX, et non nécessairement à ce qui est vrai et à ce qui est faux. Si le système formel est sûr, tout ce qui est VRAI est également vrai et tout ce qui est FAUX est également faux. Mais il existe alors certains énoncés, formulables au sein du système, qui sont vrais sans être VRAIS, et d'autres qui sont faux sans être FAUX, autrement dit des énoncés qui sont, dans les deux cas, INDÉCIDABLES. Dans ce jeu ZF, si ZF est consistant, l'énoncé gödelien* G(ZF) et sa négation $\sim G$(ZF) appartiennent respectivement à ces deux catégories. (En fait, si ZF était *in*consistant, G(ZF) et sa négation $\sim G$(ZF) seraient *tous deux* VRAIS — et également FAUX !)

Le jeu ZF correspond certainement à une attitude parfaitement raisonnable quand il s'agit d'accomplir la plupart des travaux dignes d'intérêt en mathématiques courantes. Toutefois, pour les raisons que j'ai données plus haut, il ne me semble pas traduire d'authentiques *convictions* mathématiques. Car si l'on croit que les mathématiques que l'on fait proviennent de vérités mathématiques réelles — par exemple, d'énoncés Π_1 —, on est alors obligé de reconnaître que le système que l'on utilise est *sûr* ; et si l'on croit qu'il est sûr, on doit alors également admettre qu'il est *consistant* et donc reconnaître que l'énoncé Π_1 qui énonce G(F) est *réellement* vrai — bien qu'INDÉCIDABLE. Ainsi, les convictions mathématiques d'un mathématicien doivent aller au-delà de ce qu'il peut déduire dans le cadre du jeu ZF. Si en revanche on ne croit pas que ZF est sûr, on ne peut alors considérer que les résultats VRAIS obtenus au jeu ZF sont réellement vrais. Dans les deux cas, le jeu ZF en lui-même ne représente pas une attitude satisfaisante envers la vérité mathématique. (La même chose vaut pour tout ZF*.)

Q15. Le système formel F que nous choisissons d'utiliser pourrait *ne pas être* consistant — du moins, nous ne pouvons être *sûrs* qu'il le soit. De quel droit alors affirmons-nous que G(F) est « manifestement » vrai ?

Bien que ce problème ait été amplement traité lors des discussions précédentes, je pense qu'il n'est pas inutile de revenir sur ses aspects essentiels dans la mesure où ce sont des arguments analogues à **Q15** qui représentent les attaques les plus courantes contre l'utilisation du théorème de Gödel faite par Lucas et par moi-même. Le point important est que nous n'affirmons pas que G(F) est nécessairement vrai quel que soit F, mais que l'on est obligé de conclure que G(F) est une vérité tout aussi fiable que toute autre vérité déductible des règles de F. (En fait, G(F) est *plus* fiable que tout énoncé déduit *de ces règles*, car F peut *être* consistant sans être sûr !) Si l'on admet un énoncé P déduit des seules règles de F, on doit alors admettre G(F) avec au moins le même crédit que l'on accorde à P. Ainsi, aucun système formel connaissable F — ou son algorithme équivalent F — ne peut constituer le fondement entier de nos convictions mathématiques vraies. Nous l'avons dit dans les

* Comme plus haut, on peut remplacer sans inconvénient G(F) par Ω(F). La même remarque vaut pour les questions **Q15** à **Q20**.

réponses à **Q5** et **Q6**, l'argument se présente comme un raisonnement par l'absurde : on suppose que $\mathbb{F}$ constitue la totalité de ce fondement, puis on montre que cela conduit à une contradiction et donc que $\mathbb{F}$ ne constitue pas la totalité de ce fondement.

Comme dans **Q14**, nous pouvons bien entendu utiliser par commodité un certain système $\mathbb{F}$ même si nous ne sommes pas certains qu'il soit sûr, et donc consistant. Mais si nous avons un doute authentique sur $\mathbb{F}$, nous devons alors énoncer tout résultat P obtenu à l'aide de $\mathbb{F}$ sous la forme

« P est déductible à l'intérieur de $\mathbb{F}$ »

(ou « P est VRAI ») et non affirmer simplement « P est vrai ». C'est là un énoncé mathématique parfaitement correct, qui peut se révéler soit réellement vrai soit réellement faux. Il serait tout à fait légitime de se restreindre à des énoncés mathématiques de cette forme, mais cela reviendrait en définitive à formuler encore des énoncés sur des vérités mathématiques absolues. De temps en temps, on pourrait croire que l'on a démontré qu'un énoncé donné sous la forme ci-dessus est lui-même faux, autrement dit que l'on a démontré que

« P n'est pas déductible à l'intérieur de $\mathbb{F}$ ».

Les énoncés de ce type sont de la forme : « Tel calcul ne se termine pas » (en fait : « F appliqué à P ne se termine pas »), ce qui est précisément la forme des énoncés Π_1 que j'ai considérés. Le problème est alors le suivant : quels sont les moyens autorisés pour déduire des énoncés de ce type ? Quelles sont, en fait, les procédures mathématiques auxquelles on *croit* lorsqu'on établit des vérités mathématiques ? La confiance que l'on a en ces procédures, si elle est raisonnable, ne peut être *équivalente* à la confiance que l'on accorde à un système formel, quoi que puisse être ce système formel.

Q16. La vérité de la proposition $G(\mathbb{F})$, à propos d'un système formel consistant $\mathbb{F}$, suppose que les symboles de $\mathbb{F}$ censés représenter les entiers naturels représentent *effectivement* les entiers naturels. Rien n'interdit que pour d'autres types de nombres — appelons-les nombres « surnaturels » —, $G(\mathbb{F})$ puisse se révéler fausse. Comment savons-nous que les symboles de notre système $\mathbb{F}$ représentent vraiment les entiers naturels, et non des nombres surnaturels ?

Il est vrai qu'aucun système axiomatique fini ne garantit que les « nombres » que nous utilisons sont réellement, comme nous le pensons, les *entiers naturels* et non un certain type de nombres « surnaturels »[5]. Mais en un sens, c'est justement ce que dit l'argumentation gödelienne. Quel que soit le système axiomatique $\mathbb{F}$ que nous prenions pour caractériser les entiers naturels, les règles de $\mathbb{F}$ ne peuvent nous dire par elles-mêmes si $G(\mathbb{F})$ est effectivement vrai ou faux. Si l'on suppose que $\mathbb{F}$ est consistant, on sait alors que le sens *que l'on attribue* à l'énoncé $G(\mathbb{F})$ est effectivement vrai — et non faux. Mais il faut cependant pour cela que les symboles constituant l'expression formelle

désignée par « $G(\mathbb{F})$ » aient bien le sens qu'on leur attribue. Si on réinterprète ces symboles dans un sens totalement différent, on peut alors aboutir à une interprétation totalement erronée de « $G(\mathbb{F})$ ».

Pour voir comment surviennent ces ambiguïtés, considérons deux nouveaux systèmes formels $\mathbb{F}^*$ et $\mathbb{F}^{**}$ obtenus respectivement en ajoutant $G(\mathbb{F})$ et $\sim G(\mathbb{F})$ aux axiomes de $\mathbb{F}$. Si $\mathbb{F}$ est sûr, $\mathbb{F}^*$ et $\mathbb{F}^{**}$ sont tous deux consistants (car $G(\mathbb{F})$ est vrai et $\sim G(\mathbb{F})$ n'est pas déductible des règles de $\mathbb{F}$). Mais dans l'interprétation attribuée aux symboles de $\mathbb{F}$ (appelée interprétation *standard*), si $\mathbb{F}$ est sûr, $\mathbb{F}^*$ est lui aussi sûr tandis que $\mathbb{F}^{**}$ *ne l'est pas.* Toutefois, l'une des propriétés des systèmes formels consistants est que l'on peut trouver des réinterprétations de leurs symboles, dites *non standard*, telles que les propositions qui sont fausses dans l'interprétation standard se révèlent vraies dans une interprétation non standard ; dans une telle interprétation non standard, ce sont alors $\mathbb{F}$ et $\mathbb{F}^{**}$ qui sont sûrs, tandis que $\mathbb{F}^*$ ne l'est plus. Cette réinterprétation peut certes éventuellement affecter la signification des symboles logiques (tels « $\sim$ » et « & » qui, dans l'interprétation standard, signifient respectivement « non » et « et »), mais nous nous intéressons ici uniquement aux symboles représentant des nombres indéterminés (« x », « y », « z », « x' », « x'' », etc.) et à la signification des quantificateurs logiques ($\forall$, $\exists$) utilisés avec ces nombres. Tandis que dans l'interprétation standard, « $\forall x$ » et « $\exists x$ » signifient respectivement « pour tout entier naturel x » et « il existe un entier naturel x tel que », dans une interprétation *non* standard, ces symboles s'appliquent non à des entiers naturels, mais à un autre type de nombres, possédant une structure d'ordre différente (et que nous pouvons effectivement appeler « surnaturels » selon la terminologie de Hofstadter (1979)).

Le fait est cependant que nous *savons* ce que sont les entiers naturels et que nous n'éprouvons aucune difficulté à les distinguer des étranges nombres surnaturels. Les entiers naturels sont des choses que l'on désigne ordinairement par les symboles 0, 1, 2, 3, 4, 5, 6, ... Ce sont là des concepts qui nous sont familiers dès l'enfance (*cf.* §1.21). Il y a cependant quelque chose de mystérieux dans le fait que *nous semblons* en avoir une connaissance instinctive. Car dans l'enfance (voire même à l'âge adulte), on ne nous donne qu'un nombre comparativement petit de descriptions de la signification de « zéro », « un », « deux », « trois », etc. (à travers des expressions comme « trois oranges », « une banane », etc.) ; pourtant, malgré l'insuffisance de ces descriptions, nous parvenons à saisir le concept d'entier naturel dans sa totalité. En un sens platonicien, les entiers naturels semblent être des entités ayant une existence conceptuelle absolue, indépendante de nous. Malgré cette indépendance à notre égard, ces descriptions vagues et apparemment imparfaites nous permettent, intellectuellement, d'entrer en contact avec le concept même d'entier naturel. En revanche, aucun ensemble fini d'*axiomes* ne parvient à distinguer complètement les entiers naturels de ces éventuelles alternatives dites « surnaturelles ».

En outre, nous percevons directement le caractère *infini* de la totalité des entiers naturels, tandis qu'un système contraint d'opérer selon un ensemble fini de règles précises ne parvient pas à distinguer l'infinitude propre aux

entiers naturels des autres possibilités (« surnaturelles »). Même si on la représente simplement par des points « ... » comme dans

« 0, 1, 2, 3, 4, 5, 6, ... »

ou par « etc. » dans

« zéro, un, deux, trois, etc. »,

nous comprenons ce que signifie l'infinitude des entiers naturels. Nous n'avons pas besoin qu'on nous explique, à l'aide de règles précises, ce qu'est un entier naturel. Heureusement d'ailleurs, car c'est impossible. Il suffit qu'on nous mette un peu sur la bonne voie pour que, d'une manière ou d'une autre, nous découvrions que nous *savons* ce qu'est un entier naturel !

Certains lecteurs familiers des *axiomes de Peano* pour l'arithmétique des entiers naturels (brièvement évoquée à la section 2.7) se demanderont peut-être pourquoi ces axiomes ne définissent pas correctement ces nombres. La définition de Peano part d'un symbole **0** et considère un « opérateur de succession » noté **S** que l'on peut interpréter simplement comme additionnant 1 au nombre sur lequel il agit, de sorte que l'on peut *définir* **1** comme étant **S0**, **2** comme étant **S1** = **SS0**, etc. Deux règles disent ensuite que si **Sa** = **Sb**, alors **a** = **b**, et qu'il n'existe pas de **x** tel que **0** soit de la forme **Sx**, cette dernière propriété caractérisant **0**. Il y a également le « principe de récurrence » selon lequel une propriété P est vraie pour *tous* les entiers naturels **n** si elle vérifie : (i) si P(**n**) est vraie, alors P(**Sn**) est également vraie pour tout **n** ; (ii) P(**0**) est vraie. Le problème vient des opérations logiques. Dans l'interprétation standard, les symboles $\forall$ et $\exists$ signifient respectivement « pour tout *entier naturel*... » et « il existe un *entier naturel*... tel que ». Dans une interprétation non standard, la signification de ces symboles se transforme en conséquence, de sorte qu'ils quantifient un autre type de « nombre ». S'il est vrai que les spécifications mathématiques données par Peano pour l'opérateur de succession **S** caractérisent bien la relation d'ordre qui distingue les entiers naturels de toute espèce de nombres « surnaturels », ces spécifications ne peuvent s'exprimer à l'aide des règles formelles vérifiées par ces quantificateurs $\forall$ et $\exists$. Pour bien rendre le sens des spécifications mathématiques de Peano, on doit passer à ce que l'on appelle une « logique d'ordre deux » dans laquelle apparaissent certes des quantificateurs tels que $\forall$ et $\exists$, mais où ils s'appliquent maintenant non à des entiers naturels individuels, mais à des *ensembles* (infinis) d'entiers naturels. Dans la « logique d'ordre un » de l'arithmétique de Peano, ces quantificateurs opèrent sur de simples nombres, et l'on a un système formel au sens ordinaire du terme. La logique d'ordre deux, en revanche, n'engendre pas de système formel. Un système formel au sens strict doit en effet procéder de manière purement *mécanique* (*i.e.* algorithmique) lorsqu'il vérifie si ses propres règles ont été correctement appliquées — c'est d'ailleurs la principale raison pour laquelle nous considérons des systèmes formels dans ce livre. Cette propriété est absente de la logique d'ordre deux.

Dans le même esprit que l'idée exprimée par **Q16**, on croit souvent à tort que le théorème de Gödel démontre qu'il existe de nombreux types

d'arithmétique, tous également valides. Ainsi, l'arithmétique particulière avec laquelle nous décidons de travailler serait simplement définie par un système formel arbitrairement choisi. Le théorème de Gödel montre qu'aucun de ces systèmes formels, s'il est consistant, ne peut être complet ; dès lors — affirme-t-on —, nous pouvons, selon notre fantaisie, leur ajouter de nouveaux axiomes pour obtenir toutes sortes de systèmes formels consistants avec lesquels nous pouvons choisir de travailler. On compare parfois cette situation à celle qui a prévalu avec la géométrie euclidienne. Durant quelque vingt et un siècles, on a pensé que la géométrie euclidienne était la seule géométrie possible. Puis, lorsqu'au XVIII[e] siècle Gauss, Bolyai et Lobatchevski montrèrent qu'elle admet des alternatives tout aussi justifiées, la géométrie perdit apparemment son statut absolu pour acquérir un caractère arbitraire. De même, on affirme souvent que Gödel a montré que l'arithmétique, elle aussi, est soumise à un choix arbitraire, chaque ensemble d'axiomes consistant étant aussi valide qu'un autre.

C'est cependant une interprétation totalement erronée de la démonstration de Gödel. La leçon du théorème de Gödel est que la notion même de système axiomatique formel échoue à rendre compte même du plus fondamental des concepts mathématiques. Lorsque nous employons le mot « arithmétique » sans autre précision, nous entendons en fait l'arithmétique ordinaire qui opère sur les entiers naturels ordinaires 0, 1, 2, 3, 4, ... (et éventuellement leurs homologues négatifs), non sur de quelconques nombres « surnaturels ». On peut, si on le désire, explorer les propriétés des systèmes formels, et c'est assurément une entreprise mathématiquement digne d'intérêt. Mais elle a peu à voir avec l'exploration des propriétés ordinaires des entiers naturels ordinaires. Cette situation n'est à certains égards peut-être pas très différente de celle qui est survenue en géométrie. L'étude des géométries non euclidiennes est une entreprise mathématiquement intéressante, qui connaît d'importantes applications (par exemple en physique ; voir EOLP, chapitre 5, en particulier les figures 5.1 et 5.2, ainsi que la section 4.4 du présent livre), mais lorsqu'on emploie le mot « géométrie » dans le langage quotidien (*i.e.* distinct de celui parlé par un mathématicien ou un physicien théoricien dans leurs activités professionnelles), on entend en fait la géométrie euclidienne. Il existe toutefois une différence : ce qu'un logicien peut entendre par « géométrie euclidienne » peut (sous certaines réserves[6]) être défini à l'aide d'un système formel particulier, tandis que, comme l'a montré Gödel, on ne peut définir pareillement l'« arithmétique » ordinaire.

Loin de montrer que les mathématiques (et plus particulièrement l'arithmétique) constituent une quête arbitraire engagée dans des voies dépendant de la fantaisie humaine, ce que Gödel a démontré, c'est qu'elles sont une entité absolue, que l'on doit découvrir et non inventer (*cf.* §1.17). Nous découvrons pour nous-mêmes ce que sont les entiers naturels, et nous n'avons aucune difficulté à les distinguer de n'importe quel type de nombres surnaturels. Gödel a montré qu'aucun système de règles « fabriquées » ne peut, par lui-même, accomplir cela à notre place. Cette conception platonicienne des mathématiques fut importante pour la pensée de Gödel ; elle le sera également pour nous lors des considérations que je développerai à la fin de livre (§8.7).

Q17. Supposons que le système formel $\mathbb{F}$ représente les vérités mathématiques qui sont en principe accessibles à l'esprit. Ne pouvons-nous pas contourner le problème de notre incapacité d'intégrer formellement dans $\mathbb{F}$ la proposition gödelienne $G(\mathbb{F})$ en intégrant à sa place une proposition qui aurait le même *sens* que $G(\mathbb{F})$, et en réinterprétant la signification des symboles de $\mathbb{F}$?

Il existe effectivement des moyens de représenter l'argument gödelien appliqué à $\mathbb{F}$ au sein d'un système formel $\mathbb{F}$ (suffisamment vaste). Il suffit pour cela de disposer d'une réinterprétation conférant aux symboles de $\mathbb{F}$ un sens différent de celui qui leur était initialement attribué. Toutefois, interpréter $\mathbb{F}$ en disant qu'il *est* la procédure par laquelle l'esprit parvient à ses conclusions mathématiques équivaut à tricher. Si l'on interprète l'activité mentale à l'aide des seuls symboles de $\mathbb{F}$, on ne peut modifier en cours de route le sens de ces symboles. Si l'on admet que cette activité mentale contient un élément irréductible aux opérations de $\mathbb{F}$, à savoir le *sens* mouvant de ces symboles, on doit alors connaître en détail les règles qui gouvernent ce changement de sens. Soit ces règles sont non algorithmiques, et donc $\mathcal{G}$ est établi, soit il existe une procédure algorithmique spécifique responsable de cette variation de sens, auquel cas nous aurions dû l'incorporer dans notre « $\mathbb{F}$ » de départ — appelons-le $\mathbb{F}^\dagger$ — afin qu'il représente la totalité de nos intuitions, et rien alors ne justifierait ce changement de sens. Si c'est cette seconde situation qui correspond à la réalité, la proposition gödelienne $G(\mathbb{F}^\dagger)$ prend la place de $G(\mathbb{F})$ dans la discussion précédente, et nous n'avons rien gagné.

Q18. On peut formuler, même à l'intérieur d'un système aussi simple que l'arithmétique de Peano, un théorème dont l'interprétation contienne l'implication

« $\mathbb{F}$ sûr » implique « $G(\mathbb{F})$ ».

N'est-ce pas là tout ce que nous demandons au théorème de Gödel ? Cela nous permettrait certainement, si nous acceptions ne serait-ce que l'arithmétique de Peano, de passer de la croyance pour tout système formel $\mathbb{F}$ qu'il est sûr, à la croyance que sa proposition gödelienne est vraie.

Il est vrai que l'on peut formuler un tel théorème[7] dans le cadre de l'arithmétique de Peano. Plus précisément (puisque l'on ne peut intégrer correctement la notion de « sûreté » ou de « vérité » dans aucun système formel — cela résulte d'un célèbre théorème de Tarski), on peut en fait énoncer un résultat plus fort, à savoir,

« $\mathbb{F}$ consistant » implique « $G(\mathbb{F})$ »,

ou encore,

« $\mathbb{F}$ ω-consistant » implique « $\Omega(\mathbb{F})$ ».

Ces deux énoncés contiennent l'implication demandée par **Q18**, car si $\mathbb{F}$ est sûr, il est alors certainement consistant ou ω-consistant — selon le cas. Si

nous comprenons le *sens* des symboles utilisés, nous pouvons effectivement passer d'une croyance que $\mathbb{F}$ est sûr à une croyance que $G(\mathbb{F})$ est vraie. Mais cela est déjà acquis : si nous comprenons le sens de ces symboles, nous pouvons passer de $\mathbb{F}$ à $G(\mathbb{F})$. Le problème apparaît lorsque nous voulons nous dispenser de ces interprétations et passer *automatiquement* de $\mathbb{F}$ à $G(\mathbb{F})$. Si cela était possible, nous pourrions automatiser le raisonnement gödelien et construire un dispositif algorithmique englobant tout ce que nous exigeons du théorème de Gödel. Mais en fait, cela est impossible. Car si nous ajoutions cette procédure algorithmique hypothétique au système formel $\mathbb{F}$ — quel qu'il soit — initialement choisi, nous obtiendrions un *nouveau* système formel $\mathbb{F}^{\#}$ dont la proposition gödelienne $G(\mathbb{F}^{\#})$ ne serait, *elle*, pas démontrable dans $\mathbb{F}^{\#}$. Quelle que soit la partie du théorème de Gödel que l'on parvient à intégrer dans une procédure formalisée ou algorithmique, il en reste toujours un *certain* aspect irréductible à toute manipulation de ce type. Cette « intuition gödelienne » exige une référence permanente à la signification réelle des symboles du système — quel qu'il soit — auquel est appliquée l'argumentation gödelienne. À cet égard, les problèmes soulevés par **Q17** et **Q18** sont très semblables. L'impossibilité d'automatiser la « gödelisation » est intimement liée aux arguments développés dans les discussions de **Q6** et de **Q19** (voir plus loin).

Un autre aspect de **Q18** mérite d'être considéré. Soit $\mathbb{H}$ un système formel sûr contenant l'arithmétique de Peano. Le théorème mentionné en **Q18** figure alors parmi les implications de $\mathbb{H}$, et la formulation de ce théorème lorsqu'il s'applique au $\mathbb{F}$ particulier qu'est $\mathbb{H}$ est un théorème de $\mathbb{H}$. Ainsi, on peut dire que l'une des implications de $\mathbb{H}$ est

« $\mathbb{H}$ sûr » implique « $G(\mathbb{H})$ »,

ou encore

« $\mathbb{H}$ consistant » implique « $G(\mathbb{H})$ ».

Puisque toute déduction faite à l'intérieur du système $\mathbb{H}$ repose sur l'hypothèse que ce dernier est sûr, les assertions ci-dessus sont équivalentes à l'assertion « $G(\mathbb{H})$ est vraie » ; si donc $\mathbb{H}$ affirme un énoncé qui dépend explicitement du fait qu'il est sûr, il peut aussi bien affirmer directement cet énoncé. (L'affirmation : « Si vous m'en croyez, X est vrai » entraîne l'affirmation plus simple, énoncée par le même locuteur : « X est vrai. ») Or un système formel sûr $\mathbb{H}$ *ne peut* affirmer $G(\mathbb{H})$, *i.e.* il est incapable d'affirmer lui-même qu'il est sûr. En outre, on voit qu'il ne peut de fait contenir l'interprétation des symboles avec lesquels il opère. La seconde des assertions ci-dessus conduit à une situation identique, avec en plus cette ironie : alors que $\mathbb{H}$ est incapable d'affirmer sa propre consistance quand il est en fait consistant, il ne souffre pas d'une telle inhibition quand il est *in*consistant. Un $\mathbb{H}$ inconsistant peut affirmer, en tant que « théorème », absolument tout ce qu'il est capable de formuler ! Il s'avère qu'il peut même formuler « $\mathbb{H}$ est consistant ». Un système formel (suffisamment vaste) peut affirmer sa propre consistance si et seulement s'il est *in*consistant !

Q19. Pourquoi ne pas simplement procéder à une intégration itérée de la proposition gödelienne $G(\mathbb{F})$ à tout système $\mathbb{F}$ que nous déciderions d'adopter, en autorisant cette procédure à se poursuivre *indéfiniment* ?

Si l'on se donne un système formel $\mathbb{F}$, suffisamment vaste et que l'on pense sûr, on peut imaginer de lui ajouter $G(\mathbb{F})$ en tant que nouvel axiome pour obtenir un nouveau système $\mathbb{F}_1$ que l'on pense également sûr. (Pour garder une notation cohérente dans ce qui va suivre, on peut également écrire $\mathbb{F}_0$ à la place de $\mathbb{F}$.) On imagine ensuite que l'on ajoute $G(\mathbb{F}_1)$ à $\mathbb{F}_1$ pour obtenir un nouveau système $\mathbb{F}_2$, que l'on pense lui aussi sûr. Itérant le procédé, on ajoute $G(\mathbb{F}_2)$ à $\mathbb{F}_2$ pour obtenir un nouveau système $\mathbb{F}_3$, et ainsi de suite. Un petit effort de réflexion permet de se convaincre que l'on aboutit ainsi à un autre ensemble formel $\mathbb{F}_\omega$ dont les axiomes permettent d'intégrer la *totalité* de l'ensemble infini $\{G(\mathbb{F}_0), G(\mathbb{F}_1), G(\mathbb{F}_2), G(\mathbb{F}_3), \ldots\}$ en tant qu'axiomes supplémentaires de $\mathbb{F}$. Ce système $\mathbb{F}_\omega$ est bien sûr lui aussi sûr. On peut poursuivre le procédé et ajouter $G(\mathbb{F}_\omega)$ à $\mathbb{F}_\omega$ pour obtenir $\mathbb{F}_{\omega+1}$, puis $G(\mathbb{F}_{\omega+1})$ à $\mathbb{F}_{\omega+1}$ pour obtenir $\mathbb{F}_{\omega+2}$, $\mathbb{F}_{\omega+3}$, etc. Puis, comme auparavant, on peut intégrer la *totalité* de ce nouvel ensemble infini d'axiomes pour obtenir $\mathbb{F}_{\omega 2}$ $(= \mathbb{F}_{\omega+\omega})$ qui est lui aussi évidemment sûr. En ajoutant $G(\mathbb{F}_{\omega 2})$ à $\mathbb{F}_{\omega 2}$, on obtient $\mathbb{F}_{\omega 2+1}$, puis $\mathbb{F}_{\omega 2+2}$, $\mathbb{F}_{\omega 2+3}$, etc. En répétant tout le processus jusqu'à ce point, on obtient $\mathbb{F}_{\omega 3}$ $(= \mathbb{F}_{\omega 2+\omega})$, puis, en recommençant de nouveau, $\mathbb{F}_{\omega 4}$, $\mathbb{F}_{\omega 5}$, $\mathbb{F}_{\omega 6}$, etc. Un petit effort de réflexion supplémentaire permet de se convaincre que l'on peut intégrer le nouvel ensemble d'axiomes $\{G(\mathbb{F}_\omega), G(\mathbb{F}_{\omega 2}), G(\mathbb{F}_{\omega 3}), G(\mathbb{F}_{\omega 4}), \ldots\}$ pour former un nouveau système $\mathbb{F}_{\omega^2}$ $(= \mathbb{F}_{\omega\omega})$. En répétant tout le procédé, on obtient un nouveau système $\mathbb{F}_{\omega^2+\omega^2}$, puis $\mathbb{F}_{\omega^2+\omega^2+\omega^2}$, etc., qui, une fois que l'on voit comment combiner *toutes* ces choses (grâce de nouveau à un petit effort de réflexion), conduit à un système $\mathbb{F}_{\omega^3}$ encore plus vaste, qui est certainement lui aussi sûr.

Les lecteurs familiers de la notation de Cantor pour les nombres *ordinaux* reconnaîtront que les indices que j'ai utilisés désignent justement ces nombres ordinaux. Que ceux qui ignorent ces choses ne s'en préoccupent pas. Il suffit de savoir que cette « gödelisation » peut se poursuivre encore plus loin — pour obtenir les systèmes $\mathbb{F}_{\omega^4}$, $\mathbb{F}_{\omega^5}$, ... – ce qui conduit ensuite à un système $\mathbb{F}_{\omega^\omega}$ encore plus vaste ; poursuivant l'itération, on atteint des nombres ordinaux de plus en plus grands tels que ω^{ω^ω}, etc. — aussi longtemps que l'on voit, à chaque étape, comment systématiser l'ensemble des gödelisations précédemment effectuées. C'est là qu'est en fait le nœud du problème : le « petit effort de réflexion » exige que l'on ait les bonnes intuitions pour systématiser les gödelisations antérieures. On peut accomplir cette systématisation, à condition que l'étape (l'ordinal) atteinte soit repérée par ce que l'on appelle un ordinal *récursif*, ce qui signifie en fait qu'il existe un algorithme permettant de générer la procédure. Toutefois, il n'existe pas d'algorithme préétabli qui permettrait d'effectuer d'un trait cette systématisation pour *tous* les ordinaux récursifs. Nous devons faire jouer notre intuition à chaque étape.

La procédure ci-dessus fut présentée pour la première fois par Alan Turing dans sa thèse de doctorat (et publiée dans Turing (1939))[8]. Il y montrait que

l'on peut, dans un certain sens, démontrer la vérité de *tout* énoncé Π_1 en itérant la gödelisation comme nous venons de le décrire (voir Feferman 1988). Toutefois, cette itération ne fournit pas une procédure permettant d'établir mécaniquement la vérité des énoncés Π_1, pour la simple raison que l'on ne peut systématiser mécaniquement cette gödelisation. Cela se *déduit* d'ailleurs directement du résultat de Turing. Nous avons vu en effet (§2.5) qu'*aucune* procédure algorithmique ne permet d'établir la vérité — ou la fausseté — des énoncés Π_1. Ainsi, l'itération de la gödelisation ne pourra jamais déboucher sur une procédure systématique échappant aux limitations dont nous avons parlé jusqu'ici. **Q19** n'est donc en rien une objection à $\mathscr{G}$.

Q20. La valeur réelle de la compréhension mathématique réside manifestement moins dans le fait qu'elle nous permet d'accomplir des choses non calculables, que dans le fait qu'elle nous permet de remplacer des calculs horriblement compliqués par des intuitions relativement simples. En d'autres termes, plutôt que de nous faire franchir les bornes de la calculabilité, l'esprit ne nous permet-il pas simplement de prendre des raccourcis par rapport à la théorie de la complexité ?

J'admets volontiers qu'*en pratique*, l'intuition d'un mathématicien lui sert davantage à contourner la complexité de ses calculs qu'à se débattre avec la non-calculabilité. Les mathématiciens sont somme toute des gens fondamentalement paresseux qui tentent souvent de trouver des moyens d'éviter les calculs (en dépit du fait que cela les oblige parfois à un travail intellectuel considérablement plus ardu que l'exécution desdits calculs !). Il arrive fréquemment que lorsqu'on demande à un ordinateur de sortir systématiquement les théorèmes de systèmes formels même modérément complexes, il s'embourbe rapidement dans une complexité opératoire pratiquement désespérée — tandis que pour peu qu'il ait une certaine compréhension de la signification des règles du système, un mathématicien humain n'éprouve guère de difficultés à obtenir des résultats intéressants dans ce système[9].

La raison qui m'a incité à me concentrer non sur la complexité, mais sur la non-calculabilité, est simplement que c'est seulement avec cette dernière que j'ai pu formuler les énoncés solides dont j'ai besoin. Il se peut que, dans la carrière d'un mathématicien, les problèmes de non-calculabilité jouent un rôle tout à fait mineur, voire nul. Là n'est pas la question. Ce que je tente de démontrer ici est que la compréhension (mathématique) est irréductible au calcul et que l'argumentation de Gödel(-Turing) est l'une des quelques approches possibles permettant de parvenir à cette démonstration. Il est tout à fait probable que la compréhension et l'intuition mathématiques servent souvent à accomplir des choses qui *pourraient* en principe être également réalisées par des calculs — mais où le calcul aveugle peu secondé par l'intuition s'avère parfois si laborieux qu'il en devient impraticable (*cf.* §3.26). C'est là toutefois un problème bien plus délicat que celui de la non-calculabilité.

Quoi qu'il en soit, quelle que puise être la valeur de l'argument **Q20**, il ne contredit en rien la conclusion $\mathscr{G}$.

Appendice A
Gödelisation à l'aide d'une machine de Turing : construction explicite

Soit A un algorithme dont nous savons qu'il vérifie correctement que certains calculs ne se terminent pas. Je vais construire explicitement, à partir de A, un calcul *sans fin* C qui met A en échec. Cette construction permettra de déterminer le degré de complexité de C et, en le comparant à celui de A, de retrouver les formules figurant aux sections 2.6 (*cf.* **Q8**) et 3.20.

Par souci de clarté, j'utiliserai les spécifications de la machine de Turing définie dans EOLP ; je renvoie donc le lecteur à cet ouvrage pour tous les détails concernant ces spécifications. Je limiterai ici ma description au minimum de ce qui est nécessaire pour aboutir aux résultats recherchés.

Une machine de Turing possède un nombre fini d'états internes mais agit sur un ruban infini. Ce ruban est une succession linéaire de « cases » contenant chacune éventuellement une marque, le nombre de marques sur l'ensemble du ruban étant fini. Nous désignerons chaque case marquée par le symbole `1` et chaque case non marquée par le symbole `0`. La machine comprend un dispositif de lecture qui examine une marque à la fois. La nature de la marque examinée et l'état interne de la machine déterminent (i) si la machine doit ou non modifier la marque qu'elle examine, (ii) quel doit être le nouvel état interne de la machine et (iii) si le dispositif de lecture doit effectuer, le long du ruban, un déplacement d'une case vers la droite (noté **D**) ou vers la gauche (noté **G**), ou encore d'une case vers la droite et provoquer l'arrêt de la machine (désigné par **STOP**). Lorsque la machine s'arrête, le résultat du calcul effectué s'affiche sous forme d'une succession de `0` et de `1` à gauche du dispositif de lecture. Au début, le ruban est entièrement vierge à l'exception des marques définissant les données particulières (exprimées par une suite finie de `0` et de `1`) sur lesquelles la machine exécutera ses opérations. Le dispositif de lecture se trouve initialement à gauche de toutes les marques.

Pour représenter des entiers naturels sur le ruban, qu'il s'agisse des données de départ ou du résultat final, on utilise la notation *binaire développée* dans laquelle chaque nombre s'écrit sous forme binaire standard, en remplaçant toutefois les chiffres binaires « 1 » et « 0 » par `10` et `0` respectivement. Ainsi, la traduction des entiers naturels en notation binaire développée est la suivante :

0 ↔ **0**
1 ↔ **10**
2 ↔ **100**
3 ↔ **1010**
4 ↔ **1000**
5 ↔ **10010**
6 ↔ **10100**
7 ↔ **101010**
8 ↔ **10000**
9 ↔ **100010**
10 ↔ **100100**
11 ↔ **1001010**
12 ↔ **101000**
13 ↔ **1010010**
14 ↔ **1010100**
15 ↔ **10101010**
16 ↔ **100000**
17 ↔ **1000010**
etc.

Remarquez que la notation binaire développée ne comprend jamais deux **1** successifs. Ainsi, on peut signaler le commencement et la fin de la spécification d'un entier naturel par une succession de deux ou plusieurs **1** et utiliser les suites **110**, **1110**, **11110**, etc., pour définir sur le ruban divers types d'instructions.

On peut également utiliser les marques du ruban pour définir des machines de Turing particulières, notamment lorsqu'on considère l'action d'une machine de Turing *universelle U*. Une machine de Turing *universelle U* agit sur un ruban dont la partie initiale donne la spécification d'une machine de Turing particulière *T*. Autrement dit, la machine de Turing universelle imite cette machine de Turing *T*. Les données sur lesquelles *T* elle-même est censée agir sont ensuite entrées dans *U* à droite de la portion de ruban spécifiant la machine *T*. Pour spécifier la machine *T*, on utilise les suites **110**, **1110**, **11110** qui désignent les diverses instructions destinées au dispositif de lecture de *T*, à savoir, respectivement, le déplacement d'une case vers la droite sur le ruban, d'une case vers la gauche, ou l'arrêt après un déplacement d'une case vers la droite :

D ↔ **110**
G ↔ **1110**
STOP ↔ **11110**.

Immédiatement avant chacune de ces instructions, on trouve soit le symbole **0** soit la paire **10** indiquant au dispositif de lecture de marquer respectivement un **0** ou un **1** sur le ruban à la place du symbole qu'il vient de lire. Immédiatement avant ce **0** ou ce **10** se trouve l'expression en binaire développé du numéro d'état interne dans lequel la machine de Turing doit ensuite

se placer en fonction de cette même instruction. (Remarquez que les états internes étant en nombre fini, on peut les repérer par les entiers naturels successifs 0, 1, 2, 3, 4, 5, 6, ..., *N*. Le codage de ces nombres sur le ruban se fait en notation binaire développée.)

L'instruction particulière correspondant à cette opération dépend de l'état interne dans lequel se trouve la machine juste avant qu'elle ne lise le ruban, et du symbole **0** ou **1** que le dispositif de lecture va lire et peut-être modifier. Par exemple, si la spécification de *T* contient l'instruction 230 → 171**D**, cela signifie: « Si *T* est dans l'état interne 23 et si le dispositif de lecture lit **0** sur le ruban, remplace alors ce nombre par **1**, mets-toi dans l'état interne 17 et déplace-toi sur le ruban d'une case vers la droite. » Dans ce cas, la partie « 171**D** » de l'instruction est codée sous la forme **100001010110**. En la décomposant en **1000010.10.110**, on voit que sa première partie est la forme binaire développée de 17, que la partie médiane est le code de la marque **1** sur le ruban, et que la troisième partie est le code de l'instruction « déplace-toi vers la droite ». Comment spécifier l'état interne antérieur (en l'occurrence l'état interne 23) et la marque qui va être examinée sur le ruban (ici **0**) ? On peut, si on le désire, utiliser ici aussi la notation binaire développée. Cela n'est toutefois pas réellement nécessaire, car l'ordre numérique des diverses instructions suffit (*i.e.* l'ordre **00** →, **01** →, **10** →, **11** →, **20** →, **21** →, **30** →, ...).

Voilà résumé le codage des machines de Turing tel qu'il est donné dans EOLP. Pour être complet, je dois toutefois mentionner quelques points supplémentaires. Tout d'abord, nous devons nous assurer qu'à chaque état interne correspond une instruction agissant sur **0** et sur **1** (sauf qu'il n'est pas toujours nécessaire d'avoir une instruction pour l'état interne d'ordre le plus élevé et agissant sur **1**). Lorsque le programme n'utilise jamais une instruction donnée, il faut insérer une « instruction fictive ». Par exemple, si l'état interne 23 ne rencontre jamais la marque **1** durant l'exécution du programme, on peut insérer l'instruction fictive 231 → **00D**.

Dans le codage binaire d'une machine de Turing sur un ruban, selon les prescriptions mentionnées plus haut, la paire **00** est représentée par la paire **00**. On peut cependant, sans risque d'ambiguïté, utiliser un seul **0** pour séparer les suites de (plus d'un) **1***. La machine de Turing démarre dans l'état interne 0 et le dispositif de lecture se déplace le long du ruban en conservant cet état interne jusqu'à ce qu'il rencontre le premier **1**. Cela suppose que l'opé-

---

* Cela signifie que dans le codage d'une machine de Turing, on peut remplacer chaque occurrence de la suite ...**110011**... par ...**11011**.... Dans le codage de la machine de Turing universelle figurant dans EOLP (*cf.* la note 7 du chapitre 2), j'ai par quinze fois omis d'opérer cette substitution. Cela m'a passablement irrité dans la mesure où, tenant compte des prescriptions contraignantes que j'avais données, j'avais dépensé une énergie considérable afin d'obtenir pour cette machine universelle un numéro aussi petit que possible. Ces simples substitutions donnent un numéro qui est plus de 30 000 fois inférieur à celui auquel je suis parvenu ! Je remercie Steven Gunhouse d'avoir attiré mon attention sur ces omissions et d'avoir en outre vérifié que la spécification, telle qu'elle est imprimée, *fournit bien* une machine de Turing universelle.

ration 00 → 00**D** figure toujours parmi les instructions d'une machine de Turing. Ainsi, si on spécifie une machine de Turing par une suite de **0** et de **1**, cette instruction n'a pas besoin de figurer explicitement ; on peut au lieu de cela commencer par **01** → **X**, où **X** désigne la première opération non triviale effectuée par la machine, *i.e.* lorsque le dispositif de lecture rencontre son premier **1** sur le ruban. Cela permet de se dispenser de la première suite **110** (désignant 00 → 00**D**) qui apparaît de toute façon toujours dans le codage d'une machine de Turing. En outre, nous supprimerons toujours la suite finale **110** puisqu'elle aussi est commune à toutes les machines de Turing.

La suite de **0** et de **1** qui en résulte est le *codage binaire* (ordinaire, *i.e.* non développé) du numéro n correspondant à la machine de Turing T. Nous appellerons cette machine T la « $n^{\text{ième}}$ machine de Turing » et nous écrirons $T = T_n$. Un tel nombre binaire n à la fin duquel on adjoint la suite **110** forme une suite de **0** et de **1** ne contenant jamais plus de quatre **1** successifs. Un nombre n pour lequel ce ne serait pas le cas donnerait une « machine de Turing ratée » qui cesserait de fonctionner dès qu'elle rencontrerait l'« instruction » contenant plus de quatre **1**. Une telle machine « T_n » est dite *incorrectement spécifiée*. Son action sur tout ruban est, *par définition*, considérée comme sans fin. De même, une machine de Turing rencontrant une instruction l'amenant dans un état interne défini par un nombre supérieur à celui de toute instruction figurant dans le programme se trouve elle aussi « bloquée » ; cette machine est une « ratée » et son action considérée également comme sans fin. (On peut relativement facilement supprimer ces défauts de conception en recourant à divers dispositifs, mais cela n'est pas réellement nécessaire ; *cf.* §2.6, **Q4**.)

Pour construire, à partir de l'algorithme A, le calcul sans fin sur lequel A va échouer, nous supposerons que A est donné sous forme de machine de Turing. Cette machine agit sur un ruban contenant le codage de deux entiers naturels p et q. On suppose que si le calcul $A(p,q)$ s'arrête, l'action du calcul T_p sur le nombre q *ne s'arrête pas*. Rappelons que si T_p n'est pas correctement spécifiée, nous considérons alors que son action sur q ne s'arrête pas, quelle que puisse être la valeur de q. Ainsi, si p prend une valeur « non permise », le résultat de $A(p,q)$, quel qu'il soit, sera cohérent avec nos hypothèses. Nous envisagerons donc uniquement des nombres p pour lesquels T_p *est* correctement spécifiée. Il en résulte que l'expression binaire du nombre p telle qu'elle apparaît sur le ruban ne contient aucune suite ... **11111** .... Cela permet d'utiliser cette suite **11111** pour indiquer sur le ruban le début et la fin du nombre p.

Nous devons toutefois faire de même pour q, qui *n'est pas*, lui, astreint à être un nombre de ce type. C'est là une difficulté technique qui affecte les prescriptions de la machine de Turing telles que je les ai données, mais nous pouvons la contourner en écrivant les nombres p et q en base *cinq*. (Dans cette base, « 10 » désigne le nombre *cinq*, « 100 » le nombre *vingt-cinq*, « 44 » le nombre *vingt-quatre*, etc.) Mais plutôt que d'employer les chiffres 0, 1, 2, 3, 4 de la base cinq, j'utiliserai respectivement les suites **0**, **10**, **110**, **1110** et **11110**. Ainsi :

0	est représenté par la suite	**0**
1	"	**10**
2	"	**110**
3	"	**1110**
4	"	**11110**
5	"	**100**
6	"	**1010**
7	"	**10110**
8	"	**101110**
9	"	**1011110**
10	"	**1100**
11	"	**11010**
12	"	**110110**
13	"	**1101110**
14	"	**11011110**
15	"	**11100**
16	"	**111010**
...		...
25	"	**1000**
26	"	**10010**
etc.		

La notation « C_p » désignera ici la machine de Turing correctement spécifiée T_r, où r est le nombre dont l'expression binaire ordinaire, avec la suite **110** accolée à la fin, est l'expression de p en base cinq. Le nombre q, sur lequel opère l'algorithme C_p, est lui aussi exprimé en base cinq. Le calcul $A(p,q)$ est représenté par une machine de Turing agissant sur un ruban où se trouve codée la paire p, q. Ce codage a la forme

...**00111110p111110q11111000**...,

où **p** et **q** sont respectivement les expressions de p et q en base cinq.

Nous devons maintenant trouver un p et un q pour lesquels nous savons non seulement que $C_p(q)$ ne se termine pas, mais aussi que $A(p,q)$ ne se termine pas. La procédure de la section 2.5 nous dit que nous devons trouver un nombre k pour lequel C_k agissant sur n est exactement $A(n,n)$ pour tout n, puis faire $p = q = k$. De manière explicite, nous devons trouver une prescription de machine de Turing K (= C_k) dont l'action sur un ruban marqué

...**00111110n11111000**...

(**n** étant l'expression de n en base cinq) est identique à celle de A sur

...**00111110n111110n11111000**...,

pour tout n. Autrement dit, K doit prendre le nombre n (écrit en base cinq) et le copier une fois, en séparant les deux occurrences de **n** par la suite **111110** (une suite similaire indiquant le début et la fin de l'ensemble des marques sur le ruban). K doit ensuite agir sur le ruban résultant exactement comme A.

Comment obtenir K à partir de A ? Premièrement, on cherche dans la spécification de A l'instruction initiale $\mathtt{01} \rightarrow \mathtt{X}$ et on note de quel « **X** » il s'agit. On va substituer ce « **X** » au « **X** » de la spécification donnée plus bas. Pour des raisons techniques, nous devons également supposer que l'état interne 0 de A ne réapparaît plus, une fois activée l'instruction $\mathtt{01} \rightarrow \mathtt{X}$. Cette restriction sur A n'entraîne aucune perte de généralité*. (On peut utiliser 0 dans les instructions fictives, mais pas ailleurs !)

Il faut ensuite déterminer le nombre total N d'états internes dans la spécification de A (y compris l'état 0, de sorte que le plus grand numéro d'état interne de A est égal à N - 1). Si la dernière instruction dans la spécification de A n'est pas de la forme $(N\text{-}1)\mathtt{1} \rightarrow \mathtt{Y}$, il faut alors ajouter à la fin une instruction fictive $(N\text{-}1)\mathtt{1} \rightarrow \mathtt{00D}$. Pour terminer, on supprime $\mathtt{01} \rightarrow \mathtt{X}$ de la spécification de A et on ajoute à cette spécification la liste d'instructions de machine de Turing ci-dessous, dans laquelle chaque numéro d'état interne doit être augmenté de N, dans laquelle également ϕ représente l'état interne résultant 0 et dans laquelle enfin le « **X** » de « $\mathtt{11} \rightarrow \mathtt{X}$ » ci-dessous est l'instruction que nous avons notée plus haut. (En particulier, les deux premières instructions de la liste deviennent $\mathtt{01} \rightarrow N\mathtt{1D}$, $N\mathtt{0} \rightarrow (N+4)\mathtt{0D}$.)

$\phi\mathtt{1} \rightarrow \mathtt{01D}$, $\mathtt{00} \rightarrow \mathtt{40D}$, $\mathtt{01} \rightarrow \mathtt{01D}$, $\mathtt{10} \rightarrow \mathtt{21D}$, $\mathtt{11} \rightarrow \mathtt{X}$, $\mathtt{20} \rightarrow \mathtt{31D}$,
$\mathtt{21} \rightarrow \phi\mathtt{0D}$, $\mathtt{30} \rightarrow \mathtt{551D}$, $\mathtt{31} \rightarrow \phi\mathtt{0D}$, $\mathtt{40} \rightarrow \mathtt{40D}$, $\mathtt{41} \rightarrow \mathtt{51D}$,
$\mathtt{50} \rightarrow \mathtt{40D}$,
$\mathtt{51} \rightarrow \mathtt{61D}$, $\mathtt{60} \rightarrow \mathtt{40D}$, $\mathtt{61} \rightarrow \mathtt{71D}$, $\mathtt{70} \rightarrow \mathtt{40D}$, $\mathtt{71} \rightarrow \mathtt{81D}$, $\mathtt{80} \rightarrow \mathtt{40D}$,
$\mathtt{81} \rightarrow \mathtt{91D}$, $\mathtt{90} \rightarrow \mathtt{100D}$, $\mathtt{91} \rightarrow \phi\mathtt{0D}$, $\mathtt{100} \rightarrow \mathtt{111D}$, $\mathtt{101} \rightarrow \phi\mathtt{0D}$,
$\mathtt{110} \rightarrow \mathtt{121D}$, $\mathtt{111} \rightarrow \mathtt{120D}$, $\mathtt{120} \rightarrow \mathtt{131D}$, $\mathtt{121} \rightarrow \mathtt{130D}$, $\mathtt{130} \rightarrow \mathtt{141D}$,
$\mathtt{131} \rightarrow \mathtt{140D}$, $\mathtt{140} \rightarrow \mathtt{151D}$, $\mathtt{141} \rightarrow \mathtt{10D}$, $\mathtt{150} \rightarrow \mathtt{00D}$, $\mathtt{151} \rightarrow \phi\mathtt{0D}$,
$\mathtt{160} \rightarrow \mathtt{170G}$, $\mathtt{161} \rightarrow \mathtt{161G}$, $\mathtt{170} \rightarrow \mathtt{170G}$, $\mathtt{171} \rightarrow \mathtt{181G}$, $\mathtt{180} \rightarrow \mathtt{170G}$,
$\mathtt{181} \rightarrow \mathtt{191G}$, $\mathtt{190} \rightarrow \mathtt{170G}$, $\mathtt{191} \rightarrow \mathtt{201G}$, $\mathtt{200} \rightarrow \mathtt{170G}$, $\mathtt{201} \rightarrow \mathtt{211G}$,
$\mathtt{210} \rightarrow \mathtt{170G}$, $\mathtt{211} \rightarrow \mathtt{221G}$, $\mathtt{220} \rightarrow \mathtt{220G}$, $\mathtt{221} \rightarrow \mathtt{231G}$, $\mathtt{230} \rightarrow \mathtt{220G}$,
$\mathtt{231} \rightarrow \mathtt{241G}$, $\mathtt{240} \rightarrow \mathtt{220G}$, $\mathtt{241} \rightarrow \mathtt{251G}$, $\mathtt{250} \rightarrow \mathtt{220G}$, $\mathtt{251} \rightarrow \mathtt{261G}$,
$\mathtt{260} \rightarrow \mathtt{220G}$, $\mathtt{261} \rightarrow \mathtt{271G}$, $\mathtt{270} \rightarrow \mathtt{321D}$, $\mathtt{271} \rightarrow \mathtt{281G}$, $\mathtt{280} \rightarrow \mathtt{330D}$,
$\mathtt{281} \rightarrow \mathtt{291G}$, $\mathtt{290} \rightarrow \mathtt{330D}$, $\mathtt{291} \rightarrow \mathtt{301G}$, $\mathtt{300} \rightarrow \mathtt{330D}$, $\mathtt{301} \rightarrow \mathtt{311G}$,
$\mathtt{310} \rightarrow \mathtt{330D}$, $\mathtt{311} \rightarrow \mathtt{110D}$, $\mathtt{320} \rightarrow \mathtt{340G}$, $\mathtt{321} \rightarrow \mathtt{321D}$, $\mathtt{330} \rightarrow \mathtt{350G}$,
$\mathtt{331} \rightarrow \mathtt{331D}$, $\mathtt{340} \rightarrow \mathtt{360D}$, $\mathtt{341} \rightarrow \mathtt{340D}$, $\mathtt{350} \rightarrow \mathtt{371D}$, $\mathtt{351} \rightarrow \mathtt{350D}$,
$\mathtt{360} \rightarrow \mathtt{360D}$, $\mathtt{361} \rightarrow \mathtt{381D}$, $\mathtt{370} \rightarrow \mathtt{370D}$, $\mathtt{371} \rightarrow \mathtt{391D}$, $\mathtt{380} \rightarrow \mathtt{360D}$,
$\mathtt{381} \rightarrow \mathtt{401D}$, $\mathtt{390} \rightarrow \mathtt{370D}$, $\mathtt{391} \rightarrow \mathtt{411D}$, $\mathtt{400} \rightarrow \mathtt{360D}$, $\mathtt{401} \rightarrow \mathtt{421D}$,
$\mathtt{410} \rightarrow \mathtt{370D}$, $\mathtt{411} \rightarrow \mathtt{431D}$, $\mathtt{420} \rightarrow \mathtt{360D}$, $\mathtt{421} \rightarrow \mathtt{441D}$, $\mathtt{430} \rightarrow \mathtt{370D}$,
$\mathtt{431} \rightarrow \mathtt{451D}$, $\mathtt{440} \rightarrow \mathtt{360D}$, $\mathtt{441} \rightarrow \mathtt{461D}$, $\mathtt{450} \rightarrow \mathtt{370D}$, $\mathtt{451} \rightarrow \mathtt{471D}$,

* En fait, l'une des propositions initiales de Turing était que la machine *s'arrête* chaque fois que l'état interne « 0 » réapparaît à partir d'un autre état interne. Dans ce cas, non seulement il est inutile d'imposer la restriction ci-dessus, mais on peut en outre se dispenser de l'instruction **STOP**. Cela permet une simplification dans la mesure où la suite **11110** n'étant dès lors plus nécessaire en tant qu'instruction, elle peut servir de marqueur à la place de **111110**. Cette idée allégerait significativement ma description de K qui pourrait alors être formulée en base quatre au lieu de cinq.

460 → **480D**, **461** → **461D**, **470** → **490D**, **471** → **471D**, **480** → **480D**, **481** → **490D**, **490** → **481D**, **491** → **501D**, **500** → **481D**, **501** → **511D**, **510** → **481D**, **511** → **521D**, **520** → **481D**, **521** → **531D**, **530** → **541D**, **531** → **531D**, **540** → **160G**, **541** → **ϕ0D**, **550** → **531D**.

Nous sommes maintenant en mesure de déterminer une limite précise à la taille de K en fonction de la taille de A. Mesurons cette « taille » par le « degré de complexité » tel qu'il a été défini à la section 2.6 (à la fin de la réponse à **Q8**). Pour une machine de Turing T_m (telle que A), ce degré de complexité est le nombre de chiffres composant la représentation binaire du nombre m. Pour une *action* particulière de machine de Turing $T_m(n)$ (telle que K), ce degré de complexité est le nombre de chiffres binaires contenus dans le plus grand des deux nombres m et n. Soient respectivement α et κ le nombre de chiffres binaires contenus dans a et k, où

$$A = T_a \text{ et } K = T_k (= C_k).$$

A ayant au moins $2N - 1$ instructions (en omettant la première) et la spécification de chacune de ces instructions prenant au moins trois chiffres binaires, le nombre total de chiffres binaires contenus dans son numéro de machine de Turing a vérifie certainement

$$\alpha \geq 6N - 6.$$

La liste d'instructions supplémentaires donnée plus haut pour K contient 105 places (à droite des flèches) auxquelles il faut ajouter le nombre N. Les nombres résultants étant tous inférieurs à $N + 55$, leurs représentations binaires développées contiennent chacune au plus $2 \log_2(N + 55)$ chiffres, ce qui donne pour les spécifications supplémentaires des états internes au plus $210 \log_2(N + 55)$ chiffres binaires. Il faut ajouter à cela les chiffres nécessaires pour les symboles supplémentaires **0**, **1**, **D** et **G**, ce qui donne 527 chiffres de plus (y compris une éventuelle instruction « fictive » et compte tenu du fait que l'on peut éliminer six des **0** en vertu de la règle selon laquelle on peut représenter **00** par **0**), de sorte que l'on est sûr que la spécification de K exige moins de $527 + 210 \log_2(N + 55)$ chiffres de plus que celle de A :

$$\kappa < \alpha + 527 + 210 \log_2(N + 55).$$

Utilisant la relation $\alpha \geq 6N - 6$ obtenue plus haut, on trouve (en notant que $210 \log_2 6 > 542$)

$$\kappa < \alpha - 15 + 210 \log_2(\alpha + 336).$$

Quel est à présent le degré de complexité η du calcul particulier $C_k(k)$ fourni par cette procédure ? Rappelons que le degré de complexité de $T_m(n)$ est défini par le nombre de chiffres binaires contenus dans le plus grand des deux nombres m et n. Maintenant, nous avons $C_k = T_k$, de sorte que le nombre de chiffres contenus dans le « m » est simplement égal à κ. Pour trouver le nombre de chiffres binaires contenus dans le « n » de ce calcul, nous devons examiner le ruban associé à $C_k(k)$. Ce ruban débute par la suite **111110**, immédiate-

ment suivie de l'expression binaire de k', et s'achève sur la suite **11011111**. Afin d'obtenir le « n » qui numérote le ruban dans le calcul $T_m(n)$, les conventions énoncées dans EOLP exigent que cette suite entière, à l'exception du dernier chiffre qui est supprimé, soit lue comme un nombre binaire. Il s'ensuit que le nombre de chiffres binaires contenus dans ce « n » particulier est exactement égal à $\kappa + 13$, et donc que $\kappa + 13$ est également le degré de complexité η de $C_k(k)$. On a alors $\eta = \kappa + 13 < \alpha - 2 + 210 \log_2(\alpha + 336)$, ce que l'on peut écrire plus simplement

$$\eta < \alpha + 210 \log_2(\alpha + 336).$$

Les détails de la présentation que je viens de donner correspondent aux codages propres aux machines de Turing. Ils seraient passablement différents pour d'autres types de codages, mais l'idée fondamentale est en elle-même très simple. En fait, si l'on avait adopté le formalisme du λ-calcul, toute cette procédure se serait en un sens pratiquement réduite à une trivialité. (Voir EOLP, fin du chapitre 2, pour une description du λ-calcul de Church ; voir aussi Church 1941.) Dans le λ-calcul, on peut considérer A comme un opérateur **A** dont l'action sur d'autres opérateurs **P** et **Q** se définit par **(AP)Q**. Ici **P** et **Q** représentent respectivement le calcul C_p et le nombre q. On exige alors que pour de tels **P** et **Q**, **A** vérifie :

si **(AP)Q** s'arrête, alors **PQ** ne s'arrête pas.

On peut facilement construire, en λ-calcul, une opération qui ne s'arrête pas, mais pour laquelle **A** échoue à vérifier ce fait. Pour cela, on pose

$$\mathbf{K} = \lambda\mathbf{x}.[(\mathbf{Ax})\mathbf{x}]$$

de sorte de **KY** = **(AY)Y** pour tout opérateur **Y**. On considère alors l'opération

KK.

Cette opération ne se termine manifestement pas, car **KK** = **(AK)K**, et son arrêt entraînerait, en vertu de l'hypothèse sur **A**, que **KK** ne se termine pas. En outre, **A** ne peut vérifier ce fait parce que **(AK)K** ne se termine pas. Si nous *sommes certains* que **A** s'arrête si le calcul testé ne s'arrête pas, nous sommes alors également *obligés d'admettre* que **KK** ne s'arrête pas.

Remarquez que cette procédure est extrêmement économique. Si l'on écrit **KK** sous la forme

$$\mathbf{KK} = \lambda\mathbf{y}.[\mathbf{yy}](\lambda\mathbf{x}.[(\mathbf{Ax})\mathbf{x}]),$$

on voit que le nombre de symboles contenus dans **KK** est simplement 16 fois supérieur au nombre de symboles contenus dans A (en ignorant les points, qui sont de toute façon redondants) !

Strictement parlant, cette estimation n'est cependant pas tout à fait légitime, car le symbole « **x** » peut aussi apparaître dans l'expression de **A**, et il faudrait tenir compte de cela. Il y a aussi une autre difficulté : le calcul sans fin engendré par cette procédure n'est pas une opération sur les entiers naturels (car le second **K** dans **KK** n'est pas un « nombre »). En fait, le λ-calcul ne

convient pas très bien pour traiter des opérations numériques explicites, et il n'est pas souvent facile de déterminer l'expression, sous forme d'une opération du λ-calcul, d'une procédure algorithmique s'appliquant aux entiers naturels. Ces raisons (entre autres) font que les machines de Turing sont plus adaptées à notre discussion et permettent d'obtenir plus clairement les résultats recherchés.

3
La non-calculabilité de la pensée mathématique

3.1. Ce que pensaient Gödel et Turing

Au chapitre 2, j'ai tenté de montrer la solidité de l'argumentation sous-tendant la conclusion (notée $\mathscr{G}$) selon laquelle la compréhension mathématique ne résulte pas de l'action d'un algorithme (ou d'algorithmes ; *cf.* **Q1**) que les mathématiciens utiliseraient consciemment et avec une confiance totale. Ce chapitre 2 n'a toutefois pas envisagé l'éventualité plus sérieuse — *non exclue* par $\mathscr{G}$ — selon laquelle les convictions mathématiques se fonderaient sur un algorithme inconnu utilisé inconsciemment, voire sur un algorithme connaissable dont on ne peut cependant savoir — ou dont on ne peut être fermement convaincu — s'il sous-tend ces convictions mathématiques. Je vais maintenant démontrer que si de telles possibilités sont logiquement permises, elles ne sont en rien plausibles.

Soulignons tout d'abord que lorsque les mathématiciens enchaînent soigneusement leurs raisonnements conscients pour établir des vérités mathématiques, ils ne *pensent* pas suivre de manière aveugle et inconsciente des règles échappant à leur entendement. Ils pensent au contraire qu'ils fondent leurs raisonnements sur des vérités incontestables — en définitive, des vérités fondamentalement « évidentes » — et uniquement sur de telles vérités. Et si ces raisonnements peuvent être parfois extrêmement longs, complexes, ou conceptuellement subtils, ils sont, en principe et fondamentalement, inattaquables, résolument convaincants et logiquement irréprochables. Les mathématiciens ne pensent généralement pas agir selon des procédures inconnues ou inaccessibles à l'entendement, qui guideraient mystérieusement leurs convictions.

Bien sûr, ils peuvent fort bien se tromper sur ce point. Peut-être existe-t-il une procédure algorithmique gouvernant sans qu'ils le sachent toutes leurs intuitions mathématiques. Cette éventualité apparaît probablement plus vraisemblable aux non-mathématiciens qu'à la plupart des mathématiciens. Dans ce chapitre, je tenterai de convaincre le lecteur que les mathématiciens ont raison de penser qu'ils ne réagissent pas simplement à un algorithme inconnu (et inconnaissable) — ni à un algorithme auquel ils ne croient pas fermement. Il se pourrait en fait que leurs pensées et leurs convictions obéissent inconsciemment à des principes inconnus ; mais je démontrerai que si tel est le cas, ces principes sont irréductibles à une description algorithmique.

Il est intéressant à cet égard de considérer les points de vue des deux éminents mathématiciens à l'origine du raisonnement même qui nous a conduits à la conclusion $\mathscr{G}$. Que pensait donc Gödel ? Et que pensait Turing ? Le lecteur sera peut-être surpris de découvrir qu'ils tirèrent des conclusions fondamentalement opposées à partir des mêmes données mathématiques. Qu'il soit bien clair cependant que les points de vue de ces deux grands penseurs sont en accord avec la conclusion $\mathscr{G}$. Gödel semble avoir pensé que l'esprit n'est en fait pas nécessairement une entité algorithmique et n'est même pas limité par la dimension finie du cerveau. Il critiquait d'ailleurs Turing qui était d'un avis contraire. Selon Hao Wang (1974, p. 326 ; *cf.* aussi Gödel, 1990, p. 297), s'il reconnaissait avec Turing que « le cerveau fonctionne fondamentalement comme un ordinateur numérique » et que « les lois physiques, dans leurs conséquences observables, ont une marge d'erreur finie », il rejetait cependant cette autre affirmation de Turing selon laquelle « il n'y a pas d'esprit distinct de la matière », la qualifiant de « préjugé de notre époque ». Ainsi, Gödel semble avoir estimé que si le cerveau *physique* a un comportement manifestement algorithmique, l'esprit, lui, ne se réduit pas au cerveau, de sorte que son activité n'obéit pas nécessairement à ces lois algorithmiques. Il ne considérait toutefois pas $\mathscr{G}$ comme une *preuve* que l'esprit fonctionne sur un mode non algorithmique. Il écrivit[1] en effet :

> En revanche, sur la base de ce qui a été démontré jusqu'ici, rien n'interdit l'existence (que l'on pourrait éventuellement découvrir par voie empirique) d'une machine à prouver les théorèmes qui *serait* en fait équivalente à l'intuition mathématique, mais pour laquelle on ne pourrait *démontrer* cette équivalence, ni même démontrer qu'elle donne uniquement des théorèmes *corrects* en théorie des nombres.

Cette conjecture est clairement compatible avec la conclusion $\mathscr{G}$ (et je suis persuadé que Gödel était intimement conscient de l'existence d'une conclusion analogue à $\mathscr{G}$). Il n'écartait pas la *possibilité logique* que l'esprit des mathématiciens humains puisse fonctionner selon un algorithme qu'ils ne connaîtraient pas, ou qu'ils connaîtraient sans être fermement convaincus de sa sûreté (« [on ne pourrait] *démontrer* qu'elle donne uniquement des théorèmes *corrects* ... »). C'est à ce type d'algorithme que je fais allusion lorsque je parle d'algorithme « dont on ne peut savoir s'il est sûr ». Bien sûr, c'eût été une tout autre démarche de sa part de penser qu'un tel algorithme sous-tendît

réellement le fonctionnement de l'esprit d'un mathématicien. Il semble que Gödel ne pensait pas cela et se trouva apparemment attiré par la position mystique que j'ai désignée par $\mathscr{D}$ et qui affirme que le discours scientifique est impuissant à expliquer le fonctionnement de l'esprit.

De son côté, Turing semble avoir rejeté le point de vue mystique. Comme Gödel, il croyait que le cerveau, comme tout autre objet physique, a un comportement algorithmique — rappelez-vous la « thèse de Turing » (§1.6). Ainsi, il dut trouver une autre interprétation de la conclusion $\mathscr{G}$. Turing accordait une grande importance au fait que les mathématiciens soient tout à fait capables de commettre des erreurs ; il affirma qu'un ordinateur ne peut être véritablement intelligent s'il ne commet pas, lui aussi, des erreurs[2] :

> En d'autres termes, si une machine est censée être infaillible, elle ne peut en même temps être intelligente. Il y a plusieurs théorèmes qui disent presque exactement cela. Mais ils ne disent pas jusqu'où peut aller l'intelligence d'une machine qui n'a aucune prétention à l'infaillibilité.

Les « théorèmes » auxquels pensait Turing sont sans aucun doute le théorème de Gödel et d'autres qui lui sont associés, tels sa propre version « algorithmique » de ce théorème. Ainsi, il semble avoir considéré comme essentielle la faillibilité de la pensée mathématique humaine et estimé que la soi-disant activité algorithmique imprécise de l'esprit s'avère bien plus efficace que n'importe quelle procédure algorithmique totalement sûre. Dès lors, pour échapper à la conclusion du raisonnement gödelien, il supposa que l'algorithme utilisé par les mathématiciens est techniquement non sûr et par conséquent certainement pas un algorithme « que l'on sait sûr ». Ce point de vue est donc compatible avec $\mathscr{G}$ et il semble que Turing aurait probablement approuvé le point de vue $\mathscr{A}$.

Dans la discussion qui va suivre, j'exposerai les raisons qui m'incitent à nier que la non-sûreté de l'algorithme utilisé par un mathématicien puisse *véritablement* expliquer ce qui se passe dans l'esprit dudit mathématicien. Il semble en tout cas intrinsèquement peu plausible que la supériorité de l'esprit sur un ordinateur infaillible réside dans son *im*précision — en particulier lorsqu'on considère, comme ici, l'aptitude du mathématicien à *déceler une vérité mathématique incontestable*, et non l'originalité ou la créativité mathématiques. Il est frappant de constater que face à des considérations telles que $\mathscr{G}$, Gödel et Turing ont adopté deux points de vue que nombre de personnes estiment passablement peu plausibles. On peut se demander s'ils auraient agi de la sorte s'ils avaient pu envisager — comme le fait le point de vue $\mathscr{C}$ que je défends ici — que des actions physiques puissent parfois être fondamentalement *non* calculables.

Dans les sections qui suivent (en particulier les sections 3.2 à 3.22), je donnerai en détail quelques arguments — dont certains sont assez complexes, déroutants ou techniques — visant à démontrer que les modèles « algorithmistes » $\mathscr{A}$ ou $\mathscr{B}$ ne fournissent pas une explication plausible de la compréhension mathématique. Au lecteur convaincu de l'inadéquation de ces deux modèles — ou que rebutent les détails — je conseillerai d'aller aussi loin

qu'il (ou elle) le pourra et, si cela devient ennuyeux, de passer directement à la section 3.23 où il (elle) trouvera un dialogue imaginaire résumant ces arguments. Mais ne revenez au raisonnement principal que si cela vous paraît indispensable.

3.2 Peut-on savoir si un algorithme non sûr simule la compréhension mathématique ?

Selon $\mathscr{G}$, rien n'interdit que la compréhension mathématique soit le résultat de l'action d'un algorithme non sûr ou inconnaissable, ou éventuellement de l'action d'un algorithme sûr et connaissable mais dont on ignore qu'il est sûr — voire encore que cet algorithme diffère d'un mathématicien à l'autre. Par « algorithme », j'entends simplement une procédure de calcul (*cf.* §1.5), c'est-à-dire tout ce qui peut en principe être simulé sur un ordinateur standard possédant une mémoire illimitée. (Rappelons qu'en vertu de la réponse à **Q8** (§2.6), le caractère « illimité » de cette mémoire — qui est bien sûr une idéalisation — n'affecte pas la validité de la conclusion $\mathscr{G}$.) Ce concept d'« algorithme » inclut des organisations descendantes, ascendantes, ou mixtes, et comprend par exemple tout ce que l'on peut obtenir à l'aide de réseaux de neurones formels (*cf.* §1.5) ou d'autres types de mécanismes ascendants, tels les « algorithmes génétiques » qui s'améliorent eux-mêmes en recourant à une procédure analogue à l'évolution darwinienne (*cf.* §3.11).

Aux sections 3.9 à 3.22 (que résume le dialogue imaginaire de la section 3.23), je montrerai explicitement que les arguments que je vais donner maintenant (ainsi que ceux exposés au chapitre 2) valent d'une manière générale pour les procédures ascendantes. Toutefois, par souci de clarté, je formulerai pour l'instant ces arguments dans le cadre d'une action algorithmique descendante, celle-ci pouvant correspondre à la procédure utilisée par un seul mathématicien ou par l'ensemble de la communauté mathématique. La discussion des questions **Q11** et **Q12** (§2.10) a montré que l'existence éventuelle d'algorithmes connus et sûrs *propres* à chaque mathématicien n'affecte pas de manière significative la conclusion $\mathscr{G}$. Nous examinerons plus loin (*cf.* §3.7) le cas d'algorithmes propres à chaque mathématicien, cette fois *non* sûrs ou *in*connaissables. Pour l'instant donc, nous allons essentiellement raisonner comme s'il n'existait qu'une seule procédure algorithmique sous-jacente à la compréhension mathématique, en nous restreignant même à la compréhension mathématique utilisée pour établir les énoncés Π_1 (rappelons que ces énoncés correspondent aux spécifications de machines de Turing dont l'action ne se termine pas — *cf.* le commentaire sous **Q10**). Dans ce qui suit, l'expression « compréhension mathématique » sera donc prise dans ce sens restreint (voir $\mathscr{G}^{**}$, p. 93).

La connaissabilité d'une hypothétique procédure algorithmique F — sûre ou non — sous-tendant la compréhension mathématique correspond à trois situations distinctes. F peut être :

I consciemment connaissable, le détail de son rôle étant également connaissable ;

II consciemment connaissable, son rôle étant cependant inconscient et inconnaissable ;

III inconsciente et inconnaissable.

Envisageons tout d'abord **I**, le cas complètement conscient. L'algorithme et son rôle étant tous deux connaissables, nous pouvons aussi bien considérer qu'ils sont *déjà* connus. Nous pouvons en effet supposer que nous faisons notre raisonnement à une époque où ces choses *sont* effectivement connues — car « connaissable » signifie que, du moins en principe, une telle époque peut survenir. Admettons donc que l'algorithme F et son rôle sous-jacent soient connus. Nous avons vu (§2.9) qu'un tel algorithme est de fait équivalent à un système formel $\mathbb{F}$. Cela signifie que la compréhension mathématique, ou du moins la compréhension de n'importe quel mathématicien, équivaut (pour ce mathématicien) à la démonstrabilité au sein d'un certain système formel $\mathbb{F}$. Pour avoir un espoir de vérifier notre conclusion $\mathscr{G}$ — conclusion imposée par les considérations développées au précédent chapitre —, nous devons supposer que le système $\mathbb{F}$ *n'est pas sûr*. Mais curieusement, cette non-sûreté n'est d'aucune utilité si $\mathbb{F}$, conformément à la situation **I**, est un système formel dont tout mathématicien *sait* — et donc a la *conviction* — qu'il sous-tend sa compréhension mathématique ! Car une telle conviction amène le mathématicien à croire (à tort) que $\mathbb{F}$ est sûr. (Un mathématicien ne peut raisonnablement nier ce qui constitue le fondement même de ses propres convictions !) Que $\mathbb{F}$ soit ou non réellement sûr, *croire* qu'il l'est signifie croire que $G(\mathbb{F})$ (ou $\Omega(\mathbb{F})$, *cf.* §2.8) est vrai ; mais $G(\mathbb{F})$ étant indémontrable dans $\mathbb{F}$ — en vertu du théorème de Gödel —, cela interdit que $\mathbb{F}$ sous-tende la *totalité* de la compréhension mathématique (associée aux énoncés Π_1). (Ce raisonnement vaut pareillement, qu'on l'applique aux mathématiciens considérés individuellement ou à l'ensemble de la communauté mathématique ; il s'applique en effet séparément à chacun des divers algorithmes qui seraient censés sous-tendre les processus de pensée des divers mathématiciens. En outre, il suffit de l'appliquer à la compréhension mathématique qui sert à établir les énoncés Π_1.) Ainsi, si l'on se donne un algorithme hypothétique F dont on sait qu'il n'est pas sûr et dont on suppose qu'il sous-tend la compréhension mathématique, on ne peut en fait *savoir* s'il a ce rôle. Cela exclut donc la situation **I**, que $\mathbb{F}$ soit sûr ou non. Si $\mathbb{F}$ est connaissable, on se retrouve face à la situation **II** : $\mathbb{F}$ sous-tend réellement la compréhension mathématique, mais on ne peut savoir s'il en est ainsi. Il se peut également, selon **III**, que $\mathbb{F}$ soit inconscient et inconnaissable.

La situation **I** (du moins dans le cadre d'algorithmes purement descendants) n'est donc pas une possibilité sérieusement envisageable ; le fait que $\mathbb{F}$ puisse

ne pas être sûr n'a, curieusement, aucune importance pour **I**. Le point crucial est que l'on ne peut *savoir* si ce système formel hypothétique, qu'il soit sûr ou non, sous-tend la compréhension mathématique ; ce n'est pas l'impossibilité de connaître l'algorithme qui est ici en jeu, mais l'impossibilité de savoir si cet algorithme est effectivement celui qui sous-tend la compréhension mathématique.

3.3 Un algorithme connaissable peut-il simuler, sans qu'on le sache, la compréhension mathématique ?

Venons-en maintenant au cas **II** et supposons que la compréhension mathématique soit équivalente à un algorithme ou un système formel dont on peut avoir une connaissance consciente, mais dont on ne peut savoir s'il joue ce rôle. Autrement dit, bien que le système formel hypothétique $\mathbb{F}$ soit connaissable, on ne peut être certain qu'il sous-tend notre compréhension mathématique. Cette situation est-elle plausible ?

Si ce $\mathbb{F}$ hypothétique n'est pas *déjà* connu, on peut, comme précédemment, supposer qu'il le sera un jour, du moins en principe. Imaginons que ce jour soit arrivé et que nous disposions des caractéristiques précises de $\mathbb{F}$. Bien qu'il puisse être très sophistiqué, ce système formel $\mathbb{F}$ est censé être suffisamment simple pour que nous puissions, du moins en principe, l'appréhender de manière parfaitement consciente. Rien cependant ne nous *assure* que $\mathbb{F}$ correspond à l'ensemble de notre intuition et de notre compréhension mathématiques (du moins par rapport aux énoncés Π_1). Nous allons voir que si elle est une possibilité logique, cette situation **II** est très peu plausible. En outre, je montrerai plus loin que même si elle était vérifiée, elle n'apporterait aucun réconfort aux adeptes de l'IA qui cherchent à construire un robot mathématicien ! Je reviendrai sur cet aspect du problème à la fin de cette section — je l'examinerai également plus à fond aux sections 3.15 et 3.29.

Pour souligner le fait qu'un tel $\mathbb{F}$ est effectivement une possibilité *logique*, rappelons-nous la « machine à prouver les théorèmes » dont Gödel nota qu'on ne pouvait (pour l'instant) logiquement exclure son existence (*cf.* la citation de la section 3.1). Nous allons le voir, cette « machine » serait en fait une procédure algorithmique F conforme aux cas **II** ou **III**. Soit, remarqua Gödel, c'est une machine « que l'on pourrait éventuellement découvrir par voie empirique », ce qui signifie que F est « consciemment connaissable », comme dans **II**, soit on ne pourrait la découvrir, ce qui est le cas **III**.

Se fondant sur son théorème, Gödel affirma que l'on ne pouvait « démontrer » que la procédure F (ou le système formel $\mathbb{F}$ qui lui est associé ; *cf.* §2.9) est « équivalente à l'intuition mathématique » — comme le dit sa citation. Dans **II** (et, par voie de conséquence, également dans **III**), j'ai formulé

cette limitation fondamentale sur F d'une manière légèrement différente : « son rôle, en tant qu'algorithme réel sous-tendant la compréhension mathématique, est inconscient et inconnaissable ».

Cette limitation (dont la nécessité découle du rejet de **I** ; *cf.* §3.2) entraîne clairement que l'on ne peut démontrer que F est équivalente à l'intuition mathématique, car une telle démonstration établirait que F tient le rôle que nous sommes en principe incapables de lui découvrir. Inversement, si ce rôle de F — être le fondement de la compréhension mathématique — *pouvait* être consciemment connu — dans le sens où l'on comprendrait pleinement en quoi consiste ce rôle —, on devrait alors accepter également la sûreté de F. Car *ne pas reconnaître* la sûreté indiscutable de F équivaut à refuser certaines de ses conséquences, alors que ces conséquences sont justement les propositions mathématiques (du moins les énoncés Π_1) qui *sont* acceptées. Ainsi, la connaissance du rôle de F équivaudrait à la possession d'une *preuve* de F, bien qu'une telle « preuve » ne soit pas une preuve formelle déduite au sein d'un système formel donné à l'avance.

Notons également que les énoncés Π_1 valides peuvent être considérés comme des exemples de « théorèmes corrects en théorie des nombres » auxquels Gödel fit allusion. En fait, si la « théorie des nombres » était considérée comme englobant la μ-opération « trouver le plus petit entier naturel possédant telle ou telle propriété arithmétique » — auquel cas elle inclurait l'action des machines de Turing (voir la fin de la section 2.8) —, *tous* les énoncés Π_1 feraient alors partie de la théorie des nombres. Ainsi, il semble qu'un raisonnement de type gödelien ne fournisse aucun moyen clair d'exclure le cas **II** pour des raisons d'ordre purement logique — du moins si l'on accepte l'autorité de Gödel !

En revanche, on peut se demander si **II** est réellement une éventualité *plausible*. Quelles sont les conséquences de l'existence d'un système formel $\mathbb{F}$ connaissable mais dont on ignore s'il est équivalent à une (indéniable) compréhension mathématique humaine ? Nous l'avons indiqué, nous pouvons toujours supposer qu'il existera une époque à partir de laquelle nous aurons effectivement connaissance des caractéristiques de ce $\mathbb{F}$. Rappelons (§2.7) qu'un système formel se définit par un ensemble d'*axiomes* et un ensemble d'*opérations* (de *règles de procédures*). Les *théorèmes* de $\mathbb{F}$ (les « propositions ») se déduisent des axiomes à l'aide de ces opérations, tous les théorèmes se formulant à l'aide des mêmes symboles que ceux servant à exprimer les axiomes. Nous supposerons que les théorèmes de $\mathbb{F}$ sont précisément les propositions (écrites à l'aide de ces symboles) dont les mathématiciens peuvent *en principe* établir qu'elles sont indéniablement vraies.

Supposons pour l'instant que les axiomes de $\mathbb{F}$ soient en nombre *fini*. Ces axiomes ne sont jamais que des formes particulières de théorèmes. Mais puisqu'un mathématicien détermine la validité d'un théorème en recourant à son intuition et à sa compréhension, le sens de chaque *axiome individuel* est en principe accessible à la compréhension mathématique. Ainsi, il arrivera ou pourrait arriver (*en principe*) une époque où chaque axiome individuel apparaîtra comme indéniablement vrai. Cela valant pour chaque axiome de $\mathbb{F}$, *tous*

ces axiomes apparaîtront finalement, au bout d'un certain temps, indéniablement vrais (ou pourraient en principe apparaître comme tels).

Qu'en est-il des règles de procédure ? Peut-on envisager une époque à laquelle celles-ci apparaîtront comme incontestablement sûres ? Pour nombre de systèmes formels, ces opérations sont simplement des choses que l'on ne peut manifestement qu'accepter, comme : « Si P et P $\Rightarrow$ Q sont deux théorèmes, alors Q est également un théorème » (*cf.* EOLP, p. 427, ou Kleene 1952 pour le symbole d'inférence « $\Rightarrow$ »). L'évidence de telles règles ne présente aucune difficulté. Toutefois, ces règles peuvent contenir des procédures d'inférence bien plus subtiles pouvant exiger, de la part du mathématicien, des considérations délicates avant qu'il puisse être en mesure de décider s'il doit ou non les accepter comme « incontestablement sûres ». En fait, nous le verrons dans un instant, $\mathbb{F}$ contient nécessairement des règles de procédure dont les mathématiciens humains *ne peuvent* savoir si elles sont sûres ou non — toujours dans le cas où les axiomes de $\mathbb{F}$ sont en nombre fini.

Pour voir qu'il en est ainsi, plaçons-nous à l'époque où le fondement des axiomes est indéniablement établi et examinons le système $\mathbb{F}$ tout entier. Supposons également que nous soyons fermement convaincus de la légitimité des règles de procédure de $\mathbb{F}$. Bien que nous ne soyons pas censés pouvoir savoir que $\mathbb{F}$ recouvre *tout* ce qui est en principe mathématiquement accessible à la compréhension et à l'intuition humaines, nous devrions du moins être alors convaincus que $\mathbb{F}$ est incontestablement sûr, car nous reconnaissons la légitimité de ses axiomes et de ses règles de procédure. Nous devrions donc être également convaincus que $\mathbb{F}$ est *consistant*. Nous en arriverions alors naturellement à penser que $G(\mathbb{F})$ est aussi nécessairement vraie, en vertu de cette consistance — et même *indéniablement* vraie ! Mais $\mathbb{F}$ étant censé — bien que nous l'ignorions — englober la totalité de ce qui selon nous est indéniablement vrai, $G(\mathbb{F})$ est en vérité un théorème de $\mathbb{F}$. En vertu du théorème de Gödel, il ne peut en être ainsi que si $\mathbb{F}$ est en fait *in*consistant. Si $\mathbb{F}$ est inconsistant, alors « 1 = 2 » est un théorème de $\mathbb{F}$. Ainsi, l'assertion « 1 = 2 » serait nécessairement, en principe, une proposition que nous jugerions indéniablement vraie — ce qui est manifestement une contradiction !

Cela étant, nous devons au moins envisager l'*éventualité* selon laquelle les mathématiciens humains utiliseraient (inconsciemment) un système formel $\mathbb{F}$ qui *ne* serait en fait *pas* sûr. J'examinerai ce problème à la section 3.4 ; admettons pour l'instant que les procédures sous-tendant la compréhension mathématique soient parfaitement sûres. Dans ce cas, nous venons de voir que l'on aboutit à une contradiction. Il y a donc au moins une des règles de procédure de $\mathbb{F}$ que les mathématiciens humains ne peuvent considérer comme incontestablement sûre (bien qu'elle le soit effectivement).

Toutes ces considérations reposaient sur l'hypothèse que $\mathbb{F}$ ne possède qu'un nombre fini d'axiomes. Examinons maintenant le cas où ces axiomes sont en nombre *infini*. Pour que $\mathbb{F}$ mérite le qualificatif de système formel — de sorte que l'on puisse toujours vérifier, à l'aide d'une procédure opératoire prédéfinie, qu'une preuve d'une certaine proposition est bien conforme aux règles de $\mathbb{F}$ —, il faut que son système d'axiomes infini soit exprimable en ter-

mes finis. En fait, la définition d'un système formel laisse toujours une certaine latitude, ses opérations pouvant être classées soit parmi les « axiomes », soit parmi les « règles de procédure ». Par exemple, bien que le système axiomatique standard de la théorie des ensembles — le système de Zermelo-Fraenkel (désigné ici par $\mathbb{ZF}$) — possède un nombre infini d'axiomes exprimés à l'aide de structures appelées « schémas axiomatiques », on peut le reformuler de sorte que le nombre de ses axiomes devienne fini[3]. En un certain sens, cette reformulation est toujours possible pour les systèmes axiomatiques qui sont des « systèmes formels » au sens opératoire exigé ici*.

Il semble donc que l'on puisse appliquer l'argumentation précédente — qui a permis d'exclure le cas **II** – à *tout* système $\mathbb{F}$ (sûr), que ses axiomes soient ou non en nombre fini. C'est effectivement possible, mais la réduction d'un système axiomatique infini à un système fini peut introduire de nouvelles règles dont le fondement n'est pas nécessairement évident. Ainsi, lorsque, comme précédemment, on se place à une époque où les axiomes et les règles de procédure de $\mathbb{F}$ nous sont connus dans leur totalité — les théorèmes de ce $\mathbb{F}$ hypothétique étant censés être précisément ceux en principe accessibles à la compréhension et à l'intuition humaines —, nous ne pouvons être certains que les règles de procédure de ce $\mathbb{F}$, contrairement à ses axiomes, puissent être perçues comme incontestablement sûres, même si elles le sont effectivement. Car, contrairement aux axiomes, les règles de procédure ne peuvent être assimilées aux théorèmes d'un système formel. Ce sont seulement les *théorèmes* de $\mathbb{F}$ dont on peut reconnaître la validité incontestable.

Sur un plan strictement logique, je ne pense pas que l'on puisse tirer plus de ces arguments. Si l'on croit en **II**, on doit admettre un certain système formel $\mathbb{F}$ (sur lequel les mathématiciens se fondent pour valider les énoncés Π_1) dont l'existence est clairement perçue par les mathématiciens humains, dont la liste finie d'axiomes est (incontestablement) acceptable, mais dont l'ensemble (fini) des règles de procédure $\mathscr{R}$ contient au moins une opération considérée comme fondamentalement douteuse. Tous les théorèmes de $\mathbb{F}$ s'avèrent alors individuellement valides — de manière quelque peu miraculeuse, car nombre d'entre eux s'obtiennent à l'aide des règles douteuses $\mathscr{R}$. Les mathématiciens humains reconnaissent (en principe) la validité *individuelle* de ces théorèmes, sans cependant disposer pour cela d'un critère *systématique*. Limitons-nous aux théorèmes de $\mathbb{F}$ qui sont des énoncés Π_1. L'ensemble $\mathscr{R}$ permet d'établir algorithmiquement la liste de tous les énoncés Π_1 pouvant être perçus comme vrais par les mathématiciens humains. Individuellement, chacun de ces énoncés Π_1 est donc perçu intuitivement comme vrai, mais au prix d'un raisonnement très différent des règles de $\mathscr{R}$ par lesquelles il a été obtenu. Afin de réduire chaque énoncé Π_1 à une pure vérité, on doit faire appel à de nouvelles intuitions, de plus en plus sophistiquées. Ainsi, presque par magie, tous les énoncés Π_1 se révèlent vrais, certains ne pouvant être perçus comme tels

* Cette reformulation peut s'obtenir de manière assez triviale. Il suffit d'interpréter les spécifications de la machine de Turing qui exécute l'algorithme *F* en termes de règles de procédure du système considéré.

qu'en recourant, à maintes reprises et à un niveau de plus en plus profond, à un type de raisonnement fondamentalement nouveau. En outre, *tout* énoncé Π_1 perçu comme vrai — par n'importe quel moyen que ce soit — appartient à la liste générée par $\mathcal{R}$. À la fin, il y a un énoncé Π_1 particulier, $G(\mathbb{F})$, qui est *vrai* et que l'on peut construire explicitement à partir de la connaissance du système $\mathbb{F}$, mais dont la vérité *n'apparaît pas* incontestable. Au mieux, les mathématiciens verront que la vérité de $G(\mathbb{F})$ dépend précisément de la sûreté de l'ensemble douteux $\mathcal{R}$ qui semble miraculeusement générer exactement tous les énoncés Π_1 *pouvant* être perçus comme incontestablement vrais.

Certaines personnes estimeront peut-être que cet argument n'est pas *totalement* déraisonnable. Il existe en effet de nombreux exemples de résultats mathématiques s'obtenant à l'aide de principes que l'on pourrait qualifier d'« heuristiques » qui, s'ils ne fournissent pas une *preuve* du résultat recherché, permettent cependant de prévoir que ce dernier a de fortes chances d'être vrai. Une preuve suit donc parfois des cheminements extrêmement variés. Toutefois, il me semble que ces principes heuristiques ont assez peu à voir avec notre $\mathcal{R}$ hypothétique. Ils permettent en fait de mieux comprendre *pourquoi* certains résultats mathématiques sont vrais*. Une fois que l'on dispose de techniques mathématiques plus développées, on comprend pourquoi ils fonctionnent. La plupart du temps, on parvient seulement à connaître les *situations* dans lesquelles on peut être sûr que ces principes fonctionnent, et celles dans lesquelles on doit les utiliser avec prudence — si l'on n'y prend garde, ils peuvent conduire à des conclusions erronées. Mais si l'on fait suffisamment attention, ces principes deviennent des auxiliaires puissants et sûrs lors de l'établissement de vérités mathématiques. Plutôt que de fournir des processus algorithmiques miraculeusement fiables — mais dont l'efficacité est inaccessible à l'intuition humaine — pour établir des énoncés Π_1, les principes heuristiques permettent de conforter notre intuition et d'accroître notre compréhension. C'est là une caractéristique très différente de celle de l'algorithme F (ou du système formel $\mathbb{F}$) intervenant dans notre cas **II**. De plus, personne n'a jamais proposé un principe heuristique qui précisément générerait *tous* les énoncés Π_1 accessibles à la compréhension des mathématiciens humains.

Bien sûr, rien de tout cela ne nous dit qu'un tel F — l'hypothétique « machine à prouver les théorèmes » de Gödel — soit une impossibilité ; mais du point de vue de la compréhension mathématique, son existence semble très improbable. Quoi qu'il en soit, on n'a aujourd'hui encore aucune idée de ce que serait un tel F, ni le moindre indice sur son éventuelle existence. Il

* Un principe heuristique de ce type peut prendre la forme d'une *conjecture*, telle l'importante conjecture de Taiyama (dont une version généralisée porte le nom de « philosophie de Langland »), à partir de laquelle on a pu déduire le plus célèbre des énoncés Π_1, à savoir le « dernier théorème de Fermat » (*cf.* note p. 187). Toutefois, la démonstration du théorème de Fermat par Andrew Wiles n'était pas indépendante de la conjecture de Taiyama — comme elle aurait dû l'être si cette conjecture avait été un « $\mathcal{R}$ » ; elle consistait en une argumentation permettant de *prouver* (dans le cas pertinent) la conjecture même de Taiyama !

pourrait n'être, au mieux, qu'une *conjecture* — et, en tout état de cause, une conjecture indémontrable. (Sa démonstration la contredirait !) Il me semble qu'il serait extrêmement imprudent de la part d'un adepte de l'IA (qu'il partage le point de vue $\mathcal{A}$ ou $\mathcal{B}$) de placer tous ses espoirs dans la découverte d'une procédure algorithmique* telle que F, dont l'existence même est pour le moins douteuse, d'autant que si elle existait, sa construction explicite serait hors de portée de l'intelligence de n'importe quel mathématicien ou logicien actuel.

Peut-on cependant concevoir qu'un tel F puisse exister et être construit en recourant à des procédures algorithmiques ascendantes suffisamment sophistiquées ? Dans la discussion du cas **III** développée aux sections 3.5 à 3.23, je présenterai un solide argument logique montrant qu'aucune procédure ascendante connaissable ne permettra jamais de trouver un tel F, même s'il existe. Ainsi, même la « machine à prouver les théorèmes » de Gödel n'est pas une possibilité logique sérieusement envisageable, sauf si l'ensemble de la compréhension mathématique est sous-tendue par des « mécanismes inconnaissables » dont la nature ne conforterait cependant en rien les partisans de l'IA !

Avant d'aborder la discussion du cas **III** et des procédures ascendantes en général, nous devons achever l'examen du cas **II**. Il est en effet encore possible que l'algorithme sous-jacent F — ou le système formel $\mathbb{F}$ — *ne* soit *pas sûr* (une alternative qui ne s'applique pas au cas **I**). Se peut-il que la compréhension mathématique soit équivalente à l'action d'un algorithme connaissable fondamentalement erroné ? C'est cette possibilité que nous allons maintenant envisager.

3.4 Les mathématiciens utilisent-ils inconsciemment un algorithme qui n'est pas sûr ?

Peut-être la compréhension mathématique est-elle sous-tendue par un système formel $\mathbb{F}$ qui *n'est pas* sûr ? Qui sait si un jour les mécanismes qui nous amènent à considérer qu'un énoncé mathématique est indéniablement vrai ne nous induiront pas fondamentalement en erreur ? Peut-être même nous

* Bien sûr, on peut toujours affirmer que la construction d'un robot mathématicien est très loin des objectifs immédiats de l'intelligence artificielle et donc soutenir que la découverte d'un tel F est prématurée ou inutile. Cette attitude témoignerait cependant d'une incompréhension totale de la présente discussion. Les points de vue qui affirment que l'on peut expliquer l'intelligence humaine en termes de processus algorithmiques exigent implicitement la présence potentielle d'un tel F — connaissable ou inconnaissable — dans la mesure où c'est simplement en recourant à l'intelligence que nous avons abouti à nos conclusions. L'aptitude mathématique ne joue à cet égard aucun rôle particulier ; voir notamment les sections 1.18 et 1.19.

sommes-nous déjà fourvoyés ? Cette situation diffère légèrement de celle que nous avons considérée lors de la discussion du cas **I**, au cours de laquelle nous avons été conduits à exclure que nous puissions *savoir* si un système formel $\mathbb{F}$ sous-tend la compréhension mathématique. Le cas qui nous occupe maintenant correspond à la situation où le *rôle* de $\mathbb{F}$ est inconnaissable, de sorte que nous devons envisager la possibilité que $\mathbb{F}$ ne soit pas sûr. Est-il toutefois réellement plausible que nos convictions mathématiques inébranlables puissent reposer sur un système sans sûreté — si peu sûr que, par exemple, « 1 = 2 » ferait en principe partie de ces convictions ? Si nous ne pouvons nous fier au raisonnement mathématique, nous ne pouvons plus dès lors nous fier *au moindre* raisonnement que nous tenons sur le fonctionnement du monde physique, car le raisonnement mathématique constitue un élément essentiel de la compréhension scientifique.

Certes on *ne peut exclure* que la logique mathématique que nous jugeons légitime aujourd'hui (ou que nous pourrions juger légitime dans l'avenir) contienne intrinsèquement une contradiction. On peut invoquer à l'appui de cette éventualité le fameux paradoxe (sur « l'ensemble des ensembles qui n'appartiennent pas à eux-mêmes ») que Bertrand Russell mentionna en 1902 à Gottlob Frege, alors que celui-ci s'apprêtait à publier l'œuvre de sa vie sur les fondements des mathématiques (voir aussi la réponse à **Q9**, §2.7, et EOLP, p. 106-107). Frege ajouta en appendice (*cf.* Frege 1964) :

> Rien ne peut probablement plus désagréablement affecter un auteur scientifique que de voir ébranlé l'un des fondements de son édifice une fois son travail achevé. Une lettre de Mr Bertrand Russell m'a mis dans cette situation ...

Bien sûr, on peut simplement dire que Frege a commis une erreur. On sait que les mathématiciens en commettent de temps en temps — et parfois de graves. De plus, l'erreur de Frege était *rectifiable*, ainsi qu'il l'admet lui-même dans la citation ci-dessus et, nous l'avons vu à la section 2.10 (dans le commentaire sous **Q13**), les erreurs rectifiables ne nous concernent pas. Comme à la section 2.10, nous ne nous intéressons ici qu'aux questions de principe et non à la faillibilité des mathématiciens individuels — aux erreurs que l'on peut déceler et décortiquer. Toutefois, la situation est ici légèrement différente de celle examinée sous **Q13**, car nous sommes maintenant en présence d'un système formel $\mathbb{F}$ dont nous *ignorons* s'il sous-tend la compréhension mathématique. Comme avant, nous ne nous intéressons pas aux erreurs — aux « lapsus » — que tel ou tel mathématicien peut commettre en raisonnant dans le cadre d'un système général cohérent. Mais nous avons maintenant une situation dans laquelle c'est le système lui-même qui peut renfermer une contradiction fondamentale. C'est exactement ce qui s'est produit avec Frege. S'il n'avait pas eu connaissance du paradoxe de Russell (ou d'un paradoxe similaire), il n'aurait probablement *pas* été persuadé que son système contenait une erreur fondamentale. Ce n'est pas que Russell ait remarqué une erreur technique de raisonnement que Frege aurait reconnue comme telle selon ses propres critères de raisonnement ; ce sont ces critères eux-mêmes qui se sont avérés

contenir intrinsèquement une contradiction. Ce fut cette *contradiction* qui persuada Frege qu'il y avait une erreur — et ce que Frege considérait auparavant comme un raisonnement inattaquable lui apparut alors fondamentalement contestable. L'erreur de raisonnement ne fut perçue qu'une fois la contradiction révélée. Sans cette révélation, les méthodes de raisonnement de Frege auraient certainement reçu la confiance des mathématiciens, et ceux-ci les auraient probablement adoptées durant de nombreuses années.

En vérité, je dois dire que dans ce cas précis, il est peu probable que de nombreux mathématiciens auraient longtemps accepté de raisonner (sur les ensembles infinis) en s'accordant la liberté qu'autorisait le système de Frege. La raison en est que les paradoxes de type russellien étaient extrêmement faciles à déceler. On pourrait cependant imaginer l'existence d'un paradoxe bien plus subtil, même contenu implicitement dans les procédures mathématiques utilisées aujourd'hui (que l'on considère comme totalement irréprochables), un paradoxe qui ne serait pas décelé avant plusieurs siècles. Ce serait une fois ce paradoxe mis au jour que l'on éprouverait la nécessité de modifier les règles de raisonnement. On interpréterait alors cette situation en disant que l'intuition mathématique ne repose pas sur des considérations intemporelles, mais qu'elle est fortement influencée par ce qui semble avoir marché *jusqu'alors* et grâce à quoi, en fait, « on arrive à s'en sortir ». Dans cette perspective, il y aurait alors un algorithme ou un système formel sous-jacent à la compréhension mathématique actuelle, un algorithme non pas définitivement fixé, mais continuellement soumis à des modifications à mesure que l'on disposerait de nouvelles informations. Je reviendrai sur le problème des algorithmes variables aux sections 3.9 à 3.11 (voir aussi §1.5). Nous verrons alors que ces « algorithmes variables » se réduisent en fait à une forme habituelle d'algorithme.

Bien sûr, il serait naïf de ma part de ne pas admettre que dans leur pratique quotidienne, les mathématiciens accordent souvent une certaine confiance à une procédure « qui semble avoir marché jusqu'ici ». Au cours de mes propres activités mathématiques, par exemple, cette forme de « laxisme » constitue indiscutablement un élément de mes réflexions. Mais si cette procédure joue ordinairement un rôle important lors du cheminement incertain vers la compréhension d'une idée, cela ne signifie pas qu'on la considère comme absolument légitime. Je doute que Frege lui-même eût affirmé de manière dogmatique la validité de son système, même s'il n'avait jamais eu connaissance du paradoxe de Russell. Une forme de raisonnement aussi générale doit en tout état de cause être proposée avec une certaine prudence et exige d'être passablement « ruminée » avant que l'on puisse lui décerner un brevet d'incontestable légitimité. Dans le cadre d'un système aussi général que celui de Frege, il me semble que l'usage aurait été de formuler les énoncés sous la forme « le système de Frege étant supposé sûr, telle proposition est vraie » et non directement « telle proposition est vraie », sans autre précaution (voir les commentaires sous **Q11** et **Q12**).

Les mathématiciens d'aujourd'hui font peut-être preuve d'une plus grande prudence envers les énoncés qu'ils considèrent comme « indéniablement vrais » — contrairement à la période de témérité excessive (due en grande

partie à l'œuvre de Frege) de la fin du XIX^e^ siècle. Depuis la mise au jour de paradoxes tels que celui de Russell, la nécessité d'une telle prudence est particulièrement évidente. Les mathématiciens avaient surtout commencé de s'enhardir après avoir perçu la puissance de la théorie des nombres infinis et des ensembles infinis publiée par Cantor à la fin du XIX^e^ siècle. (Signalons que Cantor était tout à fait conscient du problème posé par les paradoxes de type russellien ; il les avait d'ailleurs rencontrés bien avant Russell[4] et tenta de formuler une théorie qui en aurait rendu compte.) Pour ce qui nous concerne ici, la plus extrême prudence est bien entendu de rigueur. Par chance, notre discussion ne fait intervenir que des énoncés dont la vérité est indéniable, et rien qui mette en jeu les énoncés incertains liés aux ensembles infinis. Le point essentiel est que, où que se situe la frontière entre vérités indéniables et vérités suspectes, les énoncés déduits de l'argumentation gödelienne sont en fait tous des vérités indéniables (*cf.* le commentaire sous **Q13**). Le raisonnement même de Gödel (et de Turing) ne fait intervenir aucune des questions problématiques liées à l'existence de certains ensembles infinis. Les problèmes, liés à l'extrême liberté d'argumentation, qui préoccupaient Cantor, Frege et Russell ne nous concernent pas dans la mesure où ils ont trait à des raisonnements « douteux » et non « incontestables ». Cela étant, il m'apparaît vraiment peu plausible que les mathématiciens utilisent *réellement* un système formel $\mathbb{F}$ *non sûr* comme point de départ de leur intuition et de leur compréhension mathématiques. J'espère que le lecteur reconnaîtra avec moi que cette éventualité, pour réelle qu'elle soit, n'est certainement pas *plausible.*

Pour en terminer avec l'éventualité d'un hypothétique système formel $\mathbb{F}$ qui ne serait pas sûr, nous allons brièvement revenir sur les autres aspects de la faillibilité humaine que nous avons déjà discutés sous **Q12** et **Q13**. Je tiens d'abord à rappeler que nous *ne sommes pas* concernés ici par les inspirations, les conjectures et les critères heuristiques qui peuvent guider les mathématiciens vers leurs découvertes ; seuls nous intéressent les raisonnements et les intuitions sur lesquels se fondent leurs convictions à l'égard de la vérité mathématique. Ces convictions peuvent résulter d'une simple adhésion aux arguments d'autrui, sans mettre nécessairement en jeu le cheminement vers une découverte mathématique. Lorsqu'on avance à tâtons vers une découverte originale, il importe certes de donner libre cours aux spéculations, sans brider son esprit par des considérations de précision et de certitude absolue (j'ai d'ailleurs le sentiment que c'est précisément ce à quoi Turing fait allusion dans la citation de la section 3.1). Mais quand vient le moment de trancher entre l'acceptation ou le rejet des arguments censés conforter la vérité d'un nouvel énoncé mathématique, on doit alors recourir à des raisonnements et à des intuitions — souvent épaulés par de longs calculs — dépourvus d'erreurs.

Cela ne signifie pas que les mathématiciens se trompent rarement lorsqu'ils pensent avoir correctement conduit leurs raisonnements. Bien au contraire. Mais leur tendance à commettre de telles erreurs n'*accroît* pas fondamentalement leurs facultés de compréhension (bien que j'admette volontiers qu'une erreur puisse faire surgir fortuitement un éclair de compréhension). Le plus important est cependant que ces erreurs sont *rectifiables* ; une fois détectées —

que ce soit par un autre mathématicien ou par leur auteur lui-même —, elles sont *identifiables* en tant qu'erreurs. Ce n'est pas comme s'il y avait un système formel $\mathbb{F}$ fondamentalement erroné qui contrôlerait la compréhension des mathématiciens, car ce système serait incapable de reconnaître ses propres erreurs. (L'éventualité d'un système formel auto-améliorant — se modifiant lui-même chaque fois qu'il découvre une incohérence — sera examinée dans la discussion qui conduira à la section 3.14. On verra alors que cette éventualité ne permet pas vraiment d'échapper à la non-calculabilité de la pensée mathématique ; *cf.* aussi §3.26.)

La formulation incorrecte d'un énoncé mathématique donne lieu à un type d'erreur légèrement différent, car il se peut que l'auteur de l'énoncé veuille en réalité *signifier* une idée légèrement différente de celle qu'il a effectivement formulée. Ici encore, on est en présence d'une erreur rectifiable, et non d'une erreur *intrinsèque* résultant d'un système formel $\mathbb{F}$ manquant de sûreté. (Rappelez-vous la réponse de Feynman mentionnée sous **Q13** : « N'écoutez pas ce que je dis, écoutez ce que je *veux dire* ! ») Répétons-le, nous nous préoccupons uniquement ici de ce qui peut être *en principe* vérifié par un mathématicien (humain) et les erreurs de ce type — *i.e.* les erreurs rectifiables — ne nous concernent pas. Le point le plus important pour toute notre discussion est que l'idée centrale de l'argumentation de Gödel-Turing est nécessairement compréhensible par un mathématicien, et c'est cela qui nous oblige à rejeter le cas **I** et à considérer le cas **II** comme extrêmement peu plausible. Nous l'avons vu lors de la réponse à **Q13**, l'*idée* sous-jacente à l'argumentation de Gödel-Turing est en principe compréhensible par un mathématicien, même si un certain énoncé « $G(\mathbb{F})$ » sur lequel ce mathématicien peut se fonder peut être erroné — pour des raisons *rectifiables*.

L'éventualité que des algorithmes sans sûreté puissent sous-tendre la compréhension mathématique soulève d'autres problèmes. Ceux-ci concernent les procédures ascendantes telles que les algorithmes auto-améliorants, les algorithmes d'apprentissage (notamment les réseaux de neurones formels), les algorithmes incluant des ingrédients aléatoires et ceux dont l'action dépend de leur environnement. J'ai déjà abordé certains de ces problèmes (*cf.* le commentaire sous **Q2**) ; je vais les traiter plus à fond dans le cadre de l'examen du cas **III** qui va débuter à la section suivante.

3.5 Un algorithme peut-il être inconnaissable ?

Selon **III**, la compréhension mathématique serait le produit d'un algorithme inconnaissable. Mais que faut-il exactement entendre par « inconnaissable » ? Dans les précédentes sections de ce chapitre, nous nous sommes intéressés aux problèmes *de principe*. Ainsi, dire qu'un mathématicien humain peut reconnaître la vérité d'un énoncé Π_1 quelconque signifie que cet énoncé Π_1 est

accessible *en principe* à sa compréhension, et non que tout mathématicien connaît nécessairement une démonstration de cet énoncé. L'inconnaissabilité d'un *algorithme* exige une définition légèrement différente de l'adjectif « inconnaissable ». Par algorithme inconnaissable, j'entends un algorithme dont on ne peut, *en pratique*, définir la spécification.

Lorsque nous nous sommes intéressés à ce que l'on peut déduire au sein d'un système formel connaissable ou à ce que l'on peut accomplir en utilisant un algorithme connu, nous avons dû examiner ce qui pouvait ou ne pouvait pas être obtenu en principe. La possibilité d'obtenir une proposition à partir de ce système formel ou de cet algorithme était *nécessairement* considérée d'un point de vue « de principe ». Comparons cette situation à celle de la vérité des énoncés Π_1. Un énoncé Π_1 est considéré comme *vrai* s'il correspond à une machine de Turing dont l'action ne se termine pas en principe, indépendamment de ce que l'on peut accomplir en pratique par le calcul direct (*cf.* la discussion de **Q8**). De même, l'affirmation que l'on peut ou non obtenir une certaine proposition au sein d'un système formel doit être considérée d'un point de vue « de principe », cette affirmation prenant elle-même la forme d'une affirmation disant qu'un certain énoncé Π_1 est respectivement vrai ou faux (*cf.* la fin de la discussion sous **Q10**). Ainsi, lorsque nous nous intéressons à la possibilité d'obtenir des propositions au sein d'un système de règles formelles, la notion de « connaissabilité » doit être toujours considérée d'un point de vue « de principe ».

En revanche, la connaissabilité de ces règles formelles est un problème qui se situe sur le plan « pratique ». *Tout* système formel, machine de Turing ou énoncé Π_1 peut être spécifié en principe, de sorte que le problème de l'« inconnaissabilité », s'il a un sens, concerne ce qui est ou non accessible en pratique. Tout algorithme est en principe connaissable — dans le sens où l'action de la machine de Turing qui lui est associée est « connue » dès que l'entier naturel qui code cette action est lui-même connu (*e.g.* à l'aide de la procédure de classification des machines de Turing donnée dans EOLP). Rien n'autorise à penser qu'un entier naturel puisse être inconnaissable en principe. On peut dresser — *en principe* — la liste 0, 1, 2, 3, 4, 5, 6, ... de tous les entiers naturels (et donc de toutes les actions algorithmiques), et pour peu qu'on la prolonge suffisamment, cette liste contient tout nombre naturel donné à l'avance, aussi grand qu'il soit ! En pratique cependant, il existe des nombres qui sont trop grands pour qu'on puisse espérer les rencontrer lors d'une telle énumération. Par exemple, le numéro de la machine de Turing universelle donné à la page 58 de EOLP est bien trop grand pour figurer dans une énumération concrète des entiers naturels. Même si l'on disposait d'un ordinateur pouvant écrire ces nombres les uns après les autres dans le plus petit intervalle de temps théoriquement définissable (*i.e.* le temps de Planck, soit environ $0{,}5 \times 10^{-43}$ seconde — *cf.* §6.11), et si cet ordinateur s'était attelé à la tâche depuis le commencement même de l'Univers, il n'aurait pu à ce jour afficher aucun nombre dont le développement binaire contiendrait plus de 203 chiffres. Le nombre mentionné dans EOLP contient vingt fois plus de chiffres, mais cela

n'empêche pas qu'il soit, en pratique, « connaissable » (puisqu'il est explicitement présenté dans EOLP).

L'inconnaissabilité en pratique d'un entier naturel ou d'une action de machine de Turing suppose que la spécification même de ce nombre, ou de cette action, soit si complexe qu'elle dépasse les possibilités humaines. Si cela paraît excessif, on peut affirmer que la finitude de l'être humain impose probablement une *certaine* limite aux nombres pouvant être spécifiés par l'homme. (Voir à ce propos la discussion sous **Q8**.) Ainsi, dans la situation **III**, ce serait l'immense complexité des détails infimes de la spécification de l'algorithme F censé sous-tendre la compréhension mathématique qui soustraitrait cet algorithme à la connaissabilité humaine — « connaissabilité » étant pris dans le sens de *spécifiabilité* et non de connaissance de l'algorithme que nous sommes censés en fait utiliser. C'est cette exigence de non-spécifiabilité qui distingue **III** de **II**. Ainsi, la situation **III** signifie qu'il est humainement impossible ne serait-ce que de spécifier le nombre en question, en dehors du fait de savoir si ce nombre possède ou non les caractéristiques qui en feraient un déterminant de l'action algorithmique sous-tendant la compréhension mathématique humaine.

Qu'il soit bien clair ici que la taille même de ce nombre n'est pas en soi limitative. Il est très facile de spécifier des nombres dont la taille *excède* celle des nombres pouvant spécifier les actions algorithmiques simulant le comportement de tout organisme de l'Univers observable (ainsi, le nombre $2^{2^{65\,536}}$ mentionné dans la réponse à **Q8** est aisément spécifiable, bien qu'il dépasse considérablement le nombre d'états d'univers possibles de toute la matière contenue dans notre Univers observable[5]). Ce n'est pas seulement la taille d'un nombre, mais surtout l'impossibilité de sa spécification *précise*, qui le fait échapper aux capacités humaines.

Supposons qu'en vertu de **III** la spécification d'un tel F échappe effectivement aux capacités humaines. Quelles indications cela nous donne-t-il sur les perspectives de succès d'une intelligence artificielle (qu'il s'agisse de l'IA « forte » ou « faible » — *i.e.* des points de vue $\mathcal{A}$ ou $\mathcal{B}$) ? Les adeptes des systèmes IA contrôlés par ordinateur (du moins les partisans de $\mathcal{A}$, mais peut-être aussi ceux de $\mathcal{B}$) affirmeraient que les robots qui résulteraient de cette technologie pourraient atteindre voire dépasser les capacités mathématiques de l'homme. Dès lors, si l'on accepte **III**, un tel algorithme F non spécifiable par l'homme ferait partie du système de contrôle de ces robots. Cela semble impliquer qu'une telle technologie est inaccessible. En effet, s'il faut pour cela un algorithme F impossible à spécifier par l'être humain, ce dernier n'a aucun espoir de la réaliser.

Mais ce n'est pas ce que nous annoncent les partisans les plus ambitieux de l'IA. Ils affirment que cet algorithme F n'apparaîtra pas immédiatement, mais sera construit par étapes, à mesure que les robots eux-mêmes amélioreront leurs performances à travers des expériences d'apprentissage fondées sur des organisations algorithmiques ascendantes. En outre, les robots les plus perfectionnés ne seront pas des créations directes des êtres humains, mais plus probablement celles d'autres robots[6]. Une forme d'évolution darwinienne

pourrait également intervenir, améliorerant les aptitudes des robots d'une génération à l'autre. Certaines personnes affirment d'ailleurs que c'est grâce à des processus généraux de ce type que nous avons *nous-mêmes* pu acquérir, au sein de notre propre « ordinateur neuronal », un algorithme F qui nous est inconnaissable et qui contrôle notre compréhension mathématique.

Dans les prochaines sections, je montrerai que les raisonnements de ce type ne permettent pas vraiment d'éviter le problème : si les procédures initiales de l'IA sont algorithmiques et connaissables, tout algorithme résultant F sera également connaissable. Ainsi, le cas **III** se ramène à **II** ou à **I**, autrement dit aux cas qui ont déjà été exclus — aux sections 3.2 à 3.4 — soit au titre d'impossibilité effective (cas **I**), soit à celui d'éventualité pour le moins peu plausible (cas **II**). En fait, si l'on suppose que ces procédures algorithmiques sous-jacentes sont connaissables, on est réellement ramené au cas **I**. Ainsi, le cas **III** (et, par voie de conséquence, le cas **II**) se trouvera effectivement indéfendable.

Tout lecteur fermement convaincu que **III** fournit une piste probable vers un modèle algorithmique de l'esprit fera bien d'accorder à ces arguments l'attention qu'ils méritent et de les poursuivre jusqu'au bout. Car ils montrent que si l'on considère **III** comme constituant le fondement de notre compréhension mathématique, la seule interprétation plausible de l'émergence de notre propre F est une intervention divine — recoupant essentiellement la thèse $\mathscr{A}/\mathscr{D}$ mentionnée à la fin de la section 1.3. Or cette interprétation n'offre manifestement aucune consolation à ceux que préoccupent les objectifs à long terme, ô combien ambitieux, d'une IA reposant sur l'informatique !

3.6 Sélection naturelle ou acte divin ?

Peut-être devons-nous cependant considérer sérieusement l'éventualité d'une intervention divine dans les mécanismes sous-tendant notre intelligence — et l'impossibilité d'une explication de cette intelligence à l'aide d'une science qui s'est pourtant montrée si féconde dans la description du monde inanimé. Nous devons certes rester ouverts à toutes les possibilités ; qu'il soit bien clair toutefois que dans les discussions qui vont suivre, je m'en tiendrai à un point de vue scientifique. C'est donc en termes purement scientifiques que je vais maintenant analyser l'éventualité selon laquelle notre compréhension mathématique résulterait de l'action d'un algorithme impénétrable — et examiner le problème de l'émergence d'un tel algorithme. Certains lecteurs inclinent probablement à penser qu'un tel algorithme pourrait simplement avoir été implanté dans notre cerveau par un acte divin. Je ne peux réfuter de façon décisive une telle vision ; mais si on choisit à un certain stade de renoncer aux méthodes de la science, pourquoi le faire à ce point précis ? Si en effet on renonce à l'explication scientifique, pourquoi ne pas dissocier entièrement l'âme de toute action algorithmique, plutôt que de dissimuler son soi-disant

libre arbitre dans la complexité et l'insondabilité d'un algorithme censé contrôler son activité ? Il semble plus raisonnable d'adopter le point de vue — qui fut apparemment celui de Gödel — selon lequel l'activité de l'esprit est irréductible à celle du cerveau physique, et donc de souscrire au point de vue $\mathscr{D}$. Toutefois, je pense que, de nos jours, même ceux qui soutiennent que notre esprit est en un certain sens un don de Dieu tendent néanmoins à considérer que notre comportement admet une interprétation à l'intérieur du cadre de la science. Leurs arguments sont certainement défendables, mais je ne vais pas chercher ici à m'opposer à $\mathscr{D}$. J'espère que les lecteurs qui adhèrent à une forme ou une autre de $\mathscr{D}$ resteront en ma compagnie et essaieront de voir jusqu'où peut nous mener la méthode scientifique.

Quelles sont donc les conséquences scientifiques d'une hypothèse affirmant que nous parvenons à nos conclusions mathématiques à l'aide d'une action algorithmique nécessairement insondable ? En gros, elle implique que les procédures algorithmiques exceptionnellement complexes nécessaires pour simuler une authentique compréhension mathématique sont le résultat d'un bon nombre de centaines de milliers d'années (au moins) de sélection naturelle, ainsi que de plusieurs milliers d'années d'apprentissage et de collecte de données sur l'environnement physique. Les aspects héréditaires de ces procédures se seraient progressivement construits à partir d'ingrédients algorithmiques plus simples (primitifs) et résulteraient de pressions sélectives, comme c'est le cas de tous les autres éléments mécaniques, fantastiquement efficaces, qui constituent notre corps et notre cerveau. Les algorithmes mathématiques innées (*i.e.* tous les aspects héréditaires que pourrait contenir notre pensée mathématique — supposée algorithmique) se trouveraient d'une manière ou d'une autre codés dans l'ADN, sous forme de séquences particulières. Ils auraient été engendrés par une procédure semblable à celle qui donne naissance aux perfectionnements survenant progressivement ou par intermittence en réponse aux pressions sélectives. En outre, il y aurait divers types d'influences extérieures, telles l'éducation mathématique directe et l'expérience au contact de notre environnement physique, ainsi que d'autres facteurs intervenant sous forme de données supplémentaires purement aléatoires. Nous allons voir si cette explication est réellement plausible.

3.7 Combien d'algorithmes ?

Se pourrait-il que les divers modes de compréhension mathématique propres à des individus différents résultent de l'existence de nombreux algorithmes très différents, voire non équivalents ? Une chose est claire dès le départ : la perception des mathématiques est souvent très différente d'un individu à l'autre, même parmi les mathématiciens professionnels. Pour certains, ce sont les images visuelles qui comptent avant tout, tandis que pour d'autres, ce

peuvent être la structure logique précise, l'argument conceptuel subtil, ou les détails du raisonnement analytique, ou encore la pure manipulation algébrique. Signalons à cet égard que les pensées analytique et géométrique, par exemple, sont censées être situées dans des zones cérébrales opposées[7] — respectivement à droite et à gauche. Pourtant, ces deux modes de pensée permettent souvent de percevoir une même vérité mathématique. Du point de vue algorithmique, il pourrait sembler au premier abord qu'il doive y avoir des différences profondes entre les divers algorithmes mathématiques détenus par chaque mathématicien individuel. Mais en dépit des images très différentes que divers mathématiciens (ou d'autres personnes) peuvent se former pour comprendre ou pour communiquer des idées mathématiques, il est extrêmement frappant de constater que lorsque finalement ils décident de ce qui est indéniablement vrai, les mathématiciens sont d'accord entre eux — sauf dans les cas où surgit un désaccord à la suite de la détection d'une erreur (rectifiable) dans le raisonnement de l'un d'eux, ou éventuellement lorsque ce désaccord résulte de différences d'attitudes concernant un petit nombre de problèmes fondamentaux (*cf.* **Q11**, en particulier $\mathcal{G}^{***}$). Par souci de clarté, j'ignorerai ce dernier problème dans la discussion qui suit. Sa prise en compte n'affecterait pas substantiellement mes conclusions. (Dans ma démonstration, l'existence d'un très petit nombre de points de vue différents équivaut pratiquement à l'existence d'un seul point de vue.)

La découverte de la vérité mathématique emprunte parfois des chemins très divers. Il ne fait guère de doute que l'activité physique — quelle qu'elle soit — mise en jeu lors de la perception de la vérité d'un énoncé mathématique diffère très substantiellement d'un individu à l'autre, même si elle concerne une même vérité mathématique. Ainsi, si les mathématiciens s'appuient sur des algorithmes pour apprécier la vérité mathématique, ces algorithmes diffèrent probablement, dans leurs détails, d'un individu à l'autre. Pourtant, en un certain sens, ils doivent tous être *équivalents*.

Cette idée n'est pas aussi déraisonnable qu'elle paraît, du moins du point de vue de ce qui est mathématiquement *possible*. Des machines de Turing très différentes peuvent donner des résultats identiques. (Considérez par exemple la machine de Turing construite comme suit : lorsqu'elle agit sur l'entier naturel n, elle donne 0 chaque fois que n peut se décomposer en une somme de quatre carrés, et 1 chaque fois que cette décomposition est impossible. Les résultats affichés par cette machine sont identiques à ceux d'une autre qui afficherait 0 *quel que soit* le nombre n sur lequel elle opère — puisque *tout* entier naturel est la somme de quatre carrés ; *cf.* §2.3.) Deux algorithmes n'ont pas besoin d'être similaires au niveau de leurs opérations internes pour être identiques au niveau de leurs éventuels effets externes. Toutefois, en un certain sens, cela *renforce* le mystère de la gestation de nos hypothétiques algorithmes insondables censés percevoir la vérité mathématique, car nous avons maintenant besoin d'une multitude de tels algorithmes, tous parfaitement distincts les uns des autres dans le détail de leurs constructions, mais cependant tous essentiellement équivalents au niveau de leurs résultats.

3.8 Sélection naturelle de mathématiciens idéalistes et éthérés

Qu'en est-il du rôle de la sélection naturelle ? Se peut-il qu'il existe un algorithme F (voire plusieurs), gouvernant notre compréhension mathématique, qui serait inconnaissable (selon **III**) ou du moins dont le rôle serait inconnaissable (selon **II**) ? Avant de répondre à cette question, permettez-moi de rappeler un point déjà mentionné au début de la section 3.1. Lorsqu'ils déduisent ce qu'ils considèrent comme des conclusions mathématiques inattaquables, les mathématiciens *n'ont pas la conviction* d'obéir à un ensemble de règles inconnaissables — de règles si complexes qu'elles seraient en principe mathématiquement impénétrables. Ils pensent au contraire que ces conclusions sont le résultat de raisonnements, certes parfois longs et tortueux, mais reposant en définitive sur des vérités claires et inattaquables pouvant en principe être perçues comme telles par tout être humain.

En fait, que ce soit au niveau logique ou à celui du sens commun, ce qu'ils croient faire *est* effectivement ce qu'ils font. Cela ne doit pas être mis en doute ; c'est là un point sur lequel on ne saurait trop insister. Car si on soutient, en accord avec **III** ou **II**, qu'ils suivent un ensemble de règles de calcul inconnaissables ou insondables, alors c'est qu'ils font *aussi* cela — concurremment à ce qu'ils pensent qu'ils font, mais à un niveau de description différent. D'une manière ou d'une autre, le fait de suivre algorithmiquement ces règles devrait avoir les mêmes *effets* que ceux auxquels conduisent l'intuition et la compréhension mathématiques — du moins en pratique. On ne peut adhérer à l'un des deux points de vue $\mathscr{A}$ ou $\mathscr{B}$ sans croire automatiquement que cette possibilité est authentiquement plausible.

Rappelons le rôle de ces algorithmes. Ils dotent leurs détenteurs — du moins en principe — de l'aptitude à suivre correctement des raisonnements mathématiques sur des entités abstraites très éloignées de l'expérience directe, aptitude qui ne confère aucun avantage pratique discernable à ceux qui la possèdent. Quiconque a eu l'occasion de feuilleter une revue de mathématiques a pu constater à quel point les préoccupations des mathématiciens sont éloignées de tout ce qui concerne la vie pratique. Les détails des raisonnements généralement présentés dans ces revues ne sont pas immédiatement compréhensibles, sauf par une infime minorité de gens ; pourtant, ces raisonnements se développent petit à petit, étape par étape, et chacune de ces étapes pourrait *en principe* être comprise par tout être pensant, même si elle fait partie d'un raisonnement abstrait sur des ensembles infinis à la définition complexe. On doit donc supposer qu'un algorithme — ou peut-être l'un parmi beaucoup d'algorithmes possibles, tous mathématiquement équivalents — donnant aux gens la possibilité de suivre un tel raisonnement, est généré par une séquence particulière d'ADN. Si l'on croit qu'il en est ainsi, on doit alors sérieusement se demander comment diable cet algorithme — ou ces algorithmes — pourrait avoir été généré par la sélection naturelle. Il semble clair en effet que, du point de vue de la sélection naturelle, il n'y a aujourd'hui aucun avantage à

être mathématicien. (Je soupçonne même que cela pourrait être un handicap. Victimes de leurs étranges passions, les puristes des mathématiques ont une certaine propension à terminer leur carrière dans des fonctions académiques maigrement rétribuées — parfois même en n'occupant aucune fonction !) Mais, ce qui est encore plus important, nos lointains ancêtres n'ont probablement tiré aucun avantage d'une aptitude au raisonnement sur des ensembles infinis tout à fait abstraits, sans parler des ensembles infinis d'ensembles infinis, etc. Ils se préoccupaient des problèmes concrets de la vie quotidienne — construire des abris, se vêtir, concevoir des pièges à mammouths, ou, plus tard, domestiquer les animaux et pratiquer l'agriculture. (Fig. 3.1.)

Figure 3.1. Pour nos lointains ancêtres, la possession d'une aptitude spécifique à la résolution de problèmes mathématiques sophistiqués n'a certainement guère représenté d'avantage du point de vue de la sélection ; en revanche, la possession d'une aptitude générale à la *compréhension* constitua probablement un atout.

Il serait plus rationnel de supposer que les avantages dont bénéficièrent nos ancêtres venaient de qualités utiles pour accomplir toutes ces choses et qui, *par hasard*, s'avérèrent, bien plus tard, être exactement ce qu'il fallait pour mener des raisonnements mathématiques. C'est en fait plus ou moins mon point de vue. Ainsi, grâce aux contraintes de la sélection naturelle, ce serait l'aptitude générale à *comprendre* que l'Homme aurait d'une manière ou d'une autre acquise et développée de façon très poussée. Cette faculté de compréhension aurait été généraliste et se serait exercée sur de nombreux plans au bénéfice de l'Homme. La construction d'abris ou de pièges à mammouths, par exemple, serait un exemple de contexte où les facultés de compréhension de l'Homme auraient été extrêmement précieuses. À mon avis toutefois, l'aptitude à comprendre ne fut pas uniquement l'attribut de l'*Homo sapiens*. Elle a pu égale-

ment être présente, mais à un degré moindre, chez nombre d'autres animaux avec lesquels l'Homme était en compétition de sorte que ce dernier, grâce à un développement *accru* de son aptitude à comprendre, aurait acquis un avantage tout à fait considérable sur les animaux.

Ce point de vue soulève une difficulté dès que l'on attribue une nature algorithmique à cette faculté de compréhension. Car nous avons vu qu'en vertu des arguments donnés plus haut, toute faculté de compréhension suffisamment développée pour que son détenteur puisse apprécier la validité des raisonnements mathématiques, et en particulier de l'argumentation gödelienne telle que je l'ai exposée, doit, si elle est algorithmique, être une action si complexe et obscure qu'elle (ou son rôle) est inconnaissable par le détenteur même de cette faculté. Notre algorithme hypothétique issu de la sélection naturelle aurait alors nécessairement, à l'époque même de nos lointains ancêtres, été suffisamment puissant pour englober, dans son champ d'application potentiel, les règles de tout système formel que les mathématiciens d'aujourd'hui considèrent comme indéniablement consistant (ou indéniablement sûr pour ce qui concerne les énoncés Π_1 ; *cf.* §2.10, réponse sous **Q10**). Ces règles incluraient presque certainement non seulement celles du système formel $\mathbb{ZF}$ de Zermelo-Fraenkel, voire de son extension au système $\mathbb{ZFC}$ (le système $\mathbb{ZF}$ auquel on aurait ajouté l'axiome du choix) — deux systèmes dont nombre de mathématiciens estiment aujourd'hui qu'ils fournissent toutes les méthodes de raisonnement nécessaires aux mathématiques ordinaires —, mais aussi celles de tout système formel que l'on peut obtenir en appliquant à $\mathbb{ZF}$, autant de fois qu'on veut, la procédure de gödelisation, ainsi que de tout autre système formel que l'on pourrait découvrir en utilisant les intuitions accessibles aux mathématiciens — par exemple à travers la prise de conscience que la gödelisation donnera toujours des systèmes formels incontestablement sûrs — ou d'autres types de raisonnements inattaquables encore plus puissants. Cet algorithme devrait avoir intégré, au nombre de ses opérations, le pouvoir de distinguer précisément les raisonnements corrects de ceux qui ne le sont pas, et ce dans tous les domaines de l'activité mathématique non encore découverts à l'époque de nos lointains ancêtres et qui de nos jours emplissent les pages des revues spécialisées. Cet algorithme hypothétique, inconnaissable ou incompréhensible, devrait avoir renfermé, codé en son sein, le pouvoir d'accomplir tout cela. Or il aurait été entièrement généré par une sélection naturelle sévissant dans un monde où nos lointains ancêtres luttaient pour survivre. Mais dans un tel environnement, la possession d'une aptitude particulière à faire d'obscures mathématiques ne conférait probablement aucun avantage sélectif, et je tends à penser qu'il n'y a aucune raison pour qu'un tel algorithme ait alors émergé.

La situation est très différente si l'on suppose que la « compréhension » est une activité non algorithmique : elle n'a alors plus besoin d'être si complexe qu'elle en devient inconnaissable ou incompréhensible, et pourrait ainsi être bien plus proche de l'idée que les mathématiciens ont de leur activité. La compréhension est de fait perçue comme une qualité simple et de sens commun. Elle est certes difficile à définir de manière précise, mais elle nous est si

familière que nous avons peine à croire qu'on ne puisse, même en principe, la simuler correctement à l'aide d'une procédure de calcul. Pourtant, c'est bien ce que j'affirme ici. Voici pourquoi. Si l'on considère la compréhension comme une activité algorithmique, il faut qu'elle consiste en un algorithme « à tout faire », un algorithme qui contienne, préprogrammées, les réponses à toutes les questions mathématiques que l'on pourrait un jour lui poser. Si ces réponses ne sont pas directement préprogrammées, il doit alors disposer de moyens de calcul permettant de les trouver. Or, nous l'avons vu, s'ils doivent englober tout résultat accessible à la compréhension humaine, ces « préprogrammations » ou ces « moyens de calcul » sont nécessairement eux-mêmes inaccessibles à cette compréhension humaine. Comment les processus aveugles de la sélection naturelle, chargés de favoriser la survie de nos lointains ancêtres, auraient-ils pu « prévoir » que telle ou telle procédure algorithmique, sûre quoique inconnaissable, pourrait résoudre d'obscurs problèmes mathématiques qui n'avaient pas le moindre rapport avec la survie ?

3.9 Des algorithmes capables d'apprendre

De crainte que le lecteur ne soit tenté de convenir trop rapidement de l'absurdité d'une telle possibilité, je vais préciser la vision que les tenants du point de vue algorithmiste ont tendance à présenter. Nous l'avons déjà indiqué à la section 3.5, ils n'envisagent pas tant un algorithme qui aurait été, en un sens, « préprogrammé » pour fournir des réponses aux problèmes mathématiques, qu'un système algorithmique qui aurait la capacité d'*apprendre*. Autrement dit, ils envisagent quelque chose qui, conjointement aux procédures descendantes éventuellement nécessaires (*cf.* §1.5), contiendrait une part importante d'ingrédients ascendants*.

Certains peuvent avoir le sentiment qu'une description en termes de procédures descendantes est totalement inadéquate pour un système engendré uniquement par les processus aveugles de la sélection naturelle. Par procédures descendantes, j'entends ici les aspects de notre algorithme hypothétique qui sont génétiquement *fixés* dans l'organisme et ne sont pas modifiés par les expériences et l'apprentissage ultérieurs de chaque individu. Bien sûr, ces procédures descendantes n'auraient pas été conçues par quoi que ce soit qui aurait « su » d'avance ce qu'elles devaient finalement accomplir (à l'instar des

* On dispose aujourd'hui d'une théorie mathématique de l'apprentissage qui est relativement bien définie ; voir Anthony et Biggs (1992). Toutefois, cette théorie ne concerne pas tant les problèmes de la calculabilité que ceux de la *complexité* — *i.e.* les aspects liés à la vitesse ou à l'espace mémoire nécessaires à la résolution des problèmes ; *cf.* EOLP, p. 151-157. Rien n'indique que ces systèmes apprentis puissent simuler les processus permettant aux mathématiciens humains de parvenir à la notion de « vérité inattaquable ».

séquences d'ADN se transformant en une action cérébrale appropriée). Elles pourraient cependant fournir des règles précises dans le cadre desquelles fonctionnerait un cerveau s'adonnant à une activité mathématique. Ces procédures descendantes engendreraient les actions algorithmiques constituant un cadre fixe à l'intérieur duquel opéreraient des « procédures d'apprentissage » (ascendantes) plus souples.

Quelle est la nature de ces procédures d'apprentissage ? Imaginons que notre système d'apprentissage soit placé dans un environnement extérieur et que les réactions de cet environnement au comportement du système modifient en permanence son fonctionnement ultérieur. Deux éléments principaux sont alors en jeu. Un facteur *externe*, qui correspond au comportement de l'environnement et à sa réaction aux actions du système, et un facteur *interne*, qui correspond aux transformations opérées par le système sur son propre comportement en réaction aux changements survenant dans l'environnement. Nous allons d'abord examiner la nature algorithmique du facteur externe. Se peut-il que la réaction de l'environnement soit non algorithmique même si la construction interne de notre système apprenti est entièrement algorithmique ?

Dans certaines circonstances — c'est par exemple souvent le cas lors de l'« apprentissage » des réseaux de neurones formels —, la réaction de l'environnement est fournie par le comportement d'un expérimentateur, d'un formateur ou d'un enseignant — nous dirons simplement « enseignant » — ayant l'intention délibérée d'améliorer les performances du système. Lorsque ce dernier fonctionne de la manière souhaitée par l'enseignant, il reçoit un signal qui l'en avertit, de sorte que les mécanismes internes qui modifient son propre comportement lui permettent, par la suite, de fonctionner selon les désirs de l'enseignant. Considérons par exemple l'apprentissage d'un réseau de neurones formels destiné à reconnaître des visages humains. On contrôle en permanence les performances du système et on lui injecte à chaque étape la précision de ses estimations, de sorte qu'il peut améliorer ses résultats en modifiant convenablement sa structure interne. En pratique, l'enseignant ne contrôle pas nécessairement lui-même à chaque étape ces estimations, car la procédure d'apprentissage peut être en grande partie automatisée. Mais dans ce type de situation, ce sont les objectifs et les jugements de l'enseignant humain qui forment le critère ultime pour décider de la qualité de la performance. Dans d'autres types de situation, la réaction de l'environnement extérieur n'est pas forcément aussi « délibérée ». Par exemple, dans le cas du développement d'un système *vivant* — mais considéré cependant comme fonctionnant selon un certain type de réseau de neurones formels (ou d'autres procédures algorithmiques, *e.g.* les algorithmes génétiques, *cf.* §3.7) tel celui envisagé dans les modèles numériques —, il n'y a ni évaluation ni objectif externes. Au lieu de cela, le système vivant modifie éventuellement son comportement dans un sens interprétable en termes de *sélection naturelle* et agit en fonction de critères qui se sont constitués au fil de nombreuses années d'évolution et accroissent les chances de sa propre survie et de celle de sa progéniture.

3.10 L'environnement peut-il fournir un facteur externe non algorithmique ?

Nous supposons ici que le système lui-même (qu'il soit ou non vivant) est un *robot* contrôlé par ordinateur, de sorte que ses procédures automodificatrices sont entièrement algorithmiques. (J'emploie ici le mot « robot » pour souligner simplement que l'on doit considérer notre système comme une entité entièrement algorithmique en interaction avec son environnement. Cela n'implique pas que ce système soit un dispositif mécanique qui ait été délibérément construit par des êtres humains. Il pourrait être lui-même un être humain en développement — en accord avec $\mathscr{A}$ ou $\mathscr{B}$ —, ou encore un objet artificiellement construit.) Nous supposons donc ici que le facteur *interne* est entièrement algorithmique. Regardons maintenant si le facteur *externe* fourni par l'environnement est algorithmique ; autrement dit, voyons si un algorithme peut simuler correctement cet environnement tant dans le cas *artificiel* — lorsque l'environnement est contrôlé par un enseignant humain — que dans le cas *naturel* — lorsque ce contrôle est exercé par les seules forces de la sélection naturelle. Dans chacun de ces deux cas, les règles internes permettant au robot de modifier son comportement doivent être adaptées au type de signaux que l'environnement transmet au système.

Le problème de la possibilité d'une simulation de l'environnement dans le cas artificiel — *i.e.* peut-on simuler numériquement un enseignant humain — est en fait le problème que nous ne cessons de considérer depuis le début de ce livre. Les points de vue $\mathscr{A}$ ou $\mathscr{B}$ — dont nous explorons ici les conséquences — affirment qu'une telle simulation est en principe possible. Examinons la plausibilité générale de cette affirmation. Nous avons alors en présence un « robot » numérique et un environnement lui aussi numérique. Le système *global*, réunion du robot et de l'environnement, est donc une entité que l'on pourrait en principe simuler numériquement, de sorte que l'environnement n'offrirait au robot aucune possibilité de se comporter de manière non algorithmique.

Certains affirment parfois que c'est dans le fait que les êtres humains forment une *communauté* soumise à une interaction permanente entre ses membres que réside notre avantage sur les ordinateurs. Selon ce point de vue, les êtres humains pourraient être individuellement considérés comme des systèmes numériques dont la réunion donnerait quelque chose de plus. Cet argument s'appliquerait en particulier à la communauté des mathématiciens — de sorte que cette communauté, contrairement aux mathématiciens individuels qui la composent, agirait de manière non algorithmique. Une telle idée m'apparaît difficilement défendable, car on pourrait aussi bien considérer une communauté d'ordinateurs en constante interaction mutuelle. Une telle « communauté » formerait elle aussi un système numérique, et l'action de cette communauté tout entière pourrait, éventuellement, être simulée sur un seul ordinateur. Bien sûr, si le nombre de ses individus est important, la communauté peut représenter un système numérique incomparablement plus

vaste que ceux formés par ses individus, mais cela ne constitue pas une différence *de principe*. S'il est vrai que notre planète abrite plus de 5×10^9 êtres humains (sans parler de l'immense fonds de connaissances accumulé dans les bibliothèques), ce chiffre n'est pas une objection au point de vue algorithmiste : le développement de l'informatique pourrait en quelques décennies permettre de construire des ordinateurs capables de traiter l'accroissement numérique résultant du passage de l'individu à la communauté. Il semble clair que dans le cas artificiel — *i.e.* lorsque l'environnement extérieur se compose d'enseignants humains —, ce passage de l'individu à la communauté ne fait en principe rien gagner, et donc ne peut expliquer comment une entité non algorithmique pourrait émerger de l'association d'éléments entièrement algorithmiques.

Qu'en est-il du cas naturel ? L'environnement *physique*, entièrement dissocié de l'action des enseignants humains qu'il contient, peut-il renfermer des ingrédients ne pouvant, même en principe, être simulés numériquement ? Il me semble que si l'on croit qu'il en est ainsi, on reconnaît par là même l'insuffisance de la principale objection opposée à $\mathcal{C}$. Car la seule raison claire de contester $\mathcal{C}$ est que l'on doute que le comportement des objets du monde physique puisse être non algorithmique. Reconnaître l'existence d'*une* action physique non algorithmique entraîne que l'on est prêt à admettre l'éventualité de l'existence d'actions non algorithmiques dans le cerveau — ce qui rend caduque la principale objection opposée à $\mathcal{C}$. D'une manière générale toutefois, il semble très improbable que l'environnement non humain contienne quelque chose qui se déroberait au calcul à un niveau plus profond que ne le ferait un être humain. (Voir également la section 1.9 et **Q2** à la section 2.6.) Je pense que peu de gens soutiendraient sérieusement que l'environnement d'un robot puisse contenir quoi que ce soit qui se déroberait, *en principe*, à tout calcul.

Il me faut signaler un point important à propos du caractère « en principe » algorithmique de l'environnement. Il ne fait aucun doute que l'environnement *réel* de tout organisme vivant en développement (ou de tout système robotique sophistiqué) dépend de facteurs incroyablement complexes qui rendent probablement irréalisable toute simulation raisonnablement précise de cet environnement. Le comportement dynamique de systèmes physiques même relativement simples peut être excessivement complexe et peut dépendre de manière tellement cruciale du détail précis de leur état initial qu'il échappe à toute prédiction numérique — la prédiction météorologique à long terme en est un exemple souvent cité. De tels systèmes sont dits *chaotiques* ; *cf.* §1.7. (Les systèmes chaotiques possèdent un comportement complexe et effectivement imprédictible. Ils ne sont toutefois pas mathématiquement incompréhensibles ; leur étude forme une branche très active de la recherche mathématique actuelle[8].) Je l'ai dit à la section 1.7, les systèmes chaotiques *font partie* de ce que j'appelle les systèmes « algorithmiques ». En ce qui nous concerne ici, le point essentiel à propos des systèmes chaotiques est qu'il n'est pas nécessaire de pouvoir simuler un environnement chaotique *réel* : la simulation d'un environnement *typique* suffit. Pour reprendre l'exemple du temps,

nous n'avons pas besoin de connaître *la* situation météo ; *toute* situation météo plausible suffit !

3.11 Comment un robot peut-il apprendre ?

Admettons donc que nous n'ayons pas à nous préoccuper du problème de la simulation numérique de l'environnement. Nous pourrons en principe nous débrouiller avec l'environnement *à condition* de ne rencontrer aucun obstacle lors de la simulation des règles *internes* du robot lui-même. Demandons-nous alors comment ce robot fait pour apprendre. De quelles procédures d'apprentissage dispose-t-il ? Parfois ce sont des règles claires et prédéfinies de nature algorithmique. C'est par exemple le cas avec les réseaux de neurones formels généralement utilisés (*cf.* §1.5), où, en fonction des critères (naturels ou artificiels) déterminés par l'environnement extérieur, ces règles renforcent ou affaiblissent les connexions entre les « neurones » artificiels composant le réseau, de manière à améliorer sa performance globale. Un autre type de système apprenti est fourni par ce que l'on appelle les « algorithmes génétiques », dans lesquels intervient une forme de sélection naturelle des différentes procédures algorithmiques mises en jeu dans la machine : l'algorithme le plus efficace pour contrôler le système apparaît au terme d'une forme de compétition débouchant la « survie du plus apte ».

Qu'il soit bien clair que, comme c'est habituellement le cas avec une telle organisation ascendante, ces règles sont différentes des algorithmes standard descendants qui agissent selon des procédures connues et donnent des solutions précises à des problèmes mathématiques. Les règles ascendantes se contentent de guider le système, d'une manière générale, pour lui permettre d'améliorer ses performances. Toutefois, ces règles restent entièrement algorithmiques — au sens où l'on peut les faire exécuter par un ordinateur standard (une machine de Turing).

Outre ce type de règles bien définies, le robot peut également, pour améliorer ses performances, intégrer des éléments *aléatoires.* Ces ingrédients aléatoires pourraient être introduits de manière physique, peut-être en recourant à des processus quantiques tels les périodes de désintégration de noyaux atomiques radioactifs. En pratique toutefois, on utilise plutôt une procédure numérique donnant un résultat qui est *de fait* aléatoire — et qualifié de *pseudo-aléatoire* —, même s'il est totalement déterminé par l'exécution d'un calcul déterministe (*cf.* §1.9). Une autre procédure assez proche consisterait à utiliser l'*instant* précis où la quantité « aléatoire » intervient, puis à intégrer cet aspect temporel dans un calcul complexe associé, en fait, à un système chaotique, de sorte que d'infimes variations du temps donneraient des résultats réellement aléatoires. Si, strictement parlant, ces ingrédients aléatoires donnent des processus indescriptibles en termes d'« action de machine de Turing », cela ne

nous est guère utile. Une entrée aléatoire dans le fonctionnement d'un robot est, en pratique, équivalente à une entrée pseudo-aléatoire ; et une entrée pseudo-aléatoire ne donne rien qui ne soit simulable par une machine de Turing.

Le lecteur peut à ce stade se demander si, bien qu'une entrée aléatoire ne soit pas *en pratique* différente d'une entrée pseudo-aléatoire, il n'existe pas entre ces deux types d'entrée une différence *de principe*. Lors de nos discussions précédentes — *cf.* en particulier les sections 3.2 à 3.4 —, nous nous sommes effectivement intéressés à ce que les mathématiciens humains peuvent accomplir, non pas en pratique, mais en principe. De fait, techniquement parlant, il existe certains types de situations mathématiques dans lesquelles seule une entrée *réellement* aléatoire, contrairement à une entrée pseudo-aléatoire, permet de trouver la solution d'un problème. De telles situations surviennent lorsque le problème contient un élément de « compétition », comme par exemple en théorie des jeux ou en cryptographie. Dans certains types de « jeux à deux », la stratégie optimale pour chacun des deux joueurs implique un ingrédient purement aléatoire[9]. Si, lors d'un nombre de parties suffisamment grand, l'un des deux joueurs commet en permanence un écart par rapport au caractère aléatoire assurant la stratégie optimale, il laisse à son adversaire — du moins en principe — la possibilité de prendre l'avantage. Cet avantage peut éventuellement se concrétiser si cet adversaire parvient à deviner la nature de l'ingrédient pseudo-aléatoire (ou autre) que le premier joueur emploie au lieu de laisser jouer le pur hasard. Une situation analogue se produit en cryptographie. Ici, la préservation du code dépend de l'emploi d'une suite de chiffres obtenue par un procédé réellement aléatoire. Si à la place on utilise un procédé pseudo-aléatoire, il se peut qu'une personne tentant de percer le code parvienne à découvrir le détail de ce procédé — et donc le message codé.

À première vue, il peut sembler que puisque le vrai hasard joue un rôle déterminant dans ces situations de compétition, il pourrait être favorisé dans la sélection naturelle. De fait, je suis sûr qu'il *est* à de nombreux égards un facteur important dans le développement des organismes. Nous le verrons toutefois dans la suite de ce chapitre, un élément aléatoire ne permet pas d'échapper à la conclusion $\mathscr{C}$. Les arguments qui vont suivre s'appliquent même aux ingrédients *authentiquement* aléatoires et montreront que ces composants ne permettent pas de contourner les contraintes liant les systèmes algorithmiques. En fait, les processus *pseudo*-aléatoires laissent sur ce plan un petit peu plus d'espoir que les processus aléatoires (*cf.* §3.22).

Pour le moment, supposons que notre robot soit en fait une machine de Turing (bien que doté d'une capacité mémoire limitée). Pour être plus exacts, puisque ce robot est en interaction constante avec son environnement et que nous examinons le cas où cet environnement peut lui aussi être simulé numériquement, ce seraient le robot *et* son environnement que nous devrions assimiler à une machine de Turing. Il sera cependant utile de considérer séparément le robot comme une machine de Turing à part entière et de regarder son environnement comme la source des informations s'inscrivant sous forme d'entrées sur le ruban de la machine. En fait, cette analogie n'est pas tout à fait

appropriée, car techniquement parlant, une machine de Turing est un dispositif *fixe* qui n'est pas censé modifier sa structure avec l'« expérience ». On pourrait imaginer que la machine de Turing modifie sa structure tout en continuant de tourner et de recevoir l'information de l'environnement, mais cela ne marche pas. D'une part la *sortie* d'une machine de Turing n'est pas censée être examinée avant que la machine ait atteint l'instruction interne **STOP** (voir la section 2.1 et l'appendice A ; voir aussi EOLP, chapitre 2), d'autre part, une fois cette instruction atteinte, la machine n'est pas censée examiner le ruban contenant les informations d'entrée, sauf si elle repart depuis le début. Or pour pouvoir tourner de nouveau, elle doit revenir à son état initial : elle ne peut donc « apprendre ».

Il est toutefois facile de remédier à cette difficulté en utilisant la procédure technique suivante. On prend une machine de Turing qui, lorsqu'elle atteint l'instruction **STOP**, sort *deux* choses (codées sous forme d'un seul nombre). La première est le codage de son comportement externe, tandis que la seconde, réservée à son usage *interne*, est le codage de l'expérience qu'elle vient d'acquérir au contact de l'environnement extérieur. Lors de l'exécution suivante, la machine lit *d'abord* l'information « interne », *puis*, sur la suite du ruban, l'information « externe » fournie par son environnement, contenant la réaction détaillée que cet environnement a eue lors de l'exécution précédente. Ainsi, la partie *interne* du ruban (qui tend à devenir de plus en plus longue avec le temps) contient tout l'apprentissage de la machine, et celle-ci se l'*auto-injecte* lors de chaque nouvelle exécution.

3.12 Un robot peut-il avoir des « convictions mathématiques inébranlables » ?

Cette procédure permet donc de décrire à l'aide d'une machine de Turing tout « robot » numérique capable d'améliorer ses performances. Or notre robot est censé pouvoir former des jugements sur la vérité mathématique, avec tous les moyens accessibles à un mathématicien humain. Comment va-t-il s'y prendre ? Il est inutile de tenter de coder, d'une manière entièrement « descendante », toutes les règles mathématiques (telles celles contenues dans $\mathbb{ZF}$ et les autres systèmes formels que l'on peut obtenir à partir de lui — *cf.* §3.8) qui lui fourniraient directement les intuitions mathématiques dont disposent les mathématiciens, car, nous l'avons vu, il n'existe aucun moyen raisonnable (sauf une « intervention divine » — *cf.* §3.5, §3.6) de construire un tel algorithme descendant, fantastiquement complexe, et dont nous ne pourrions savoir s'il est efficace. Nous devons supposer que quels que soient ces éléments « descendants », ils ne servent pas à faire uniquement des mathématiques sophistiquées, mais sont au contraire des règles générales dont on pourrait imaginer qu'elles sont sous-jacentes à la « compréhension ».

Rappelons que l'environnement fournit deux types d'entrées — *artificielles* et *naturelles* ; *cf.* §3.9 — pouvant influencer de manière significative le comportement de notre robot. En ce qui concerne les aspects artificiels de l'environnement, on peut imaginer un enseignant (ou des enseignants) qui décrirait au robot diverses vérités mathématiques et tenterait de l'aider à trouver un moyen interne de distinguer les vérités des faussetés. L'enseignant signalerait au robot les erreurs qu'il commettrait, lui parlerait de divers concepts mathématiques, lui exposerait différentes méthodes de démonstration. Il disposerait à cette fin d'un éventail de procédures au nombre desquelles pourraient figurer des pédagogies fondées sur l'« exemple », des méthodes « suggestives », « directives », voire même tout simplement la « fessée » ! Les aspects naturels de l'environnement, eux, fourniraient au robot des « idées » sur le comportement des corps physiques, ou lui permettraient de prendre concrètement conscience des concepts mathématiques, par exemple sous forme de diverses objectivations des entiers naturels — deux oranges, sept bananes, quatre pommes, zéro chaussure, une chaussette, etc. — et de bonnes approximations de concepts géométriques — tels la ligne droite ou le cercle — ou de certains concepts liés aux ensembles infinis (comme l'ensemble des points contenus dans un cercle).

Notre robot n'étant pas préprogrammé selon une stratégie totalement descendante et étant censé forger son propre concept de vérité mathématique à l'aide de ses procédures d'apprentissage, il commettrait de temps en temps des *erreurs* — il *apprendrait* par essais et erreurs. Au début du moins, ces erreurs pourraient être corrigées par son enseignant. Ce robot pourrait aussi parfois remarquer, à partir de l'observation de son environnement, que certaines de ses idées antérieures à propos des vérités mathématiques sont en fait erronées, ou le sont probablement. Il pourrait aussi parvenir à cette conclusion par des considérations de cohérence, etc. Il commettrait de moins en moins d'erreurs à mesure que s'accroîtrait son expérience. Avec le temps, il s'appuierait de moins en moins sur l'enseignant et l'environnement physique — et peut-être pourrait-il finalement s'en dispenser totalement — et se fonderait de plus en plus sur ses capacités de calcul internes. Ainsi, notre robot serait censé pouvoir dépasser les vérités mathématiques particulières apprises au contact de son enseignant et de son environnement physique. On pourrait même imaginer qu'il puisse finalement apporter des contributions originales à la recherche.

Pour voir si tout cela est vraiment plausible, nous devons revenir sur ce que nous avons dit plus haut. Pour que notre robot possède réellement les aptitudes, la compréhension et l'intuition d'un mathématicien humain, il devra disposer d'un concept de « vérité mathématique inattaquable ». Ses premiers essais, qui auront été corrigés par son enseignant ou rendus peu plausibles par l'observation de son environnement physique, *n'entreront pas* dans cette catégorie. Ils appartiendront à la catégorie des « conjectures » formulées à titre exploratoire et susceptibles d'être erronées. Si notre robot doit se comporter comme un vrai mathématicien, il pourra certes de temps en temps commettre des erreurs, mais celles-ci seront rectifiables — et rectifiables, en principe, selon ses *propres* critères internes de « vérité inattaquable ».

Nous l'avons vu, un mathématicien humain ne peut parvenir au concept de « vérité inattaquable » en se fondant sur un ensemble de règles mécaniques (humainement) inconnaissables et en lesquelles il aurait une confiance absolue. Si nous supposons que notre robot peut égaler (voire dépasser) le niveau d'aptitude mathématique qu'un être humain est *en principe* capable d'atteindre, *son* concept de vérité mathématique inattaquable ne peut, lui non plus, s'obtenir à partir d'un ensemble de règles mécaniques pouvant, en principe, être perçues comme sûres — par un mathématicien humain ou, en l'occurrence, par notre robot mathématicien !

La question importante qui se pose alors est de savoir quels concepts, intuitions ou vérités inattaquables nous devons considérer comme pertinents : ceux des mathématiciens humains ou ceux des robots ? Peut-on considérer qu'un robot *a* réellement des intuitions ou des convictions ? S'il est partisan du point de vue $\mathcal{B}$, le lecteur éprouvera peut-être quelques difficultés à répondre à cette question, car les concepts mêmes d'« intuition » et de « conviction », étant des attributs *mentaux*, sont en tant que tels étrangers à un robot entièrement contrôlé par ordinateur. Toutefois, dans la discussion qui précède, il n'est pas réellement nécessaire que le robot possède d'authentiques qualités mentales si l'on suppose qu'il peut se comporter *extérieurement* comme un mathématicien humain — ce que pensent tant les partisans inconditionnels de $\mathcal{B}$ que ceux de $\mathcal{A}$. Ainsi, il n'est pas nécessaire que le robot comprenne, entrevoie ou croie *réellement* quoi que ce soit si, à travers les énoncés qu'il formule publiquement, il se comporte exactement comme s'il possédait ces attributs mentaux. Je reviendrai plus en détail sur ce point à la section 3.17.

En principe, le point de vue $\mathcal{B}$ ne diffère pas du point de vue $\mathcal{A}$ au niveau des limitations affectant le comportement d'un robot. Toutefois, les partisans de $\mathcal{B}$ sont probablement *moins optimistes* quant à ses compétences ou à la probabilité de concevoir un système numérique qui soit capable de simuler le cerveau d'un être humain en train de prendre conscience de la légitimité d'un argument mathématique. Une telle activité humaine implique une certaine compréhension du *sens* des concepts mathématiques mis en jeu. Selon le point de vue $\mathcal{A}$, la notion même de « sens » ne contient rien qui soit indescriptible en termes de calcul, tandis que selon $\mathcal{B}$, cette notion ressortit au domaine sémantique de l'esprit et diffère de tout ce que l'on peut décrire en termes purement algorithmiques. Selon $\mathcal{B}$, il est illusoire d'escompter que notre robot puisse avoir conscience d'un « sens » *réel*, quel qu'il soit. Ainsi, les partisans de $\mathcal{B}$ inclinent moins que les partisans de $\mathcal{A}$ à penser que tout robot construit sur les principes que nous venons de considérer puisse réellement simuler les manifestations externes de la compréhension humaine. Cela me semble indiquer (de manière somme toute naturelle) que les partisans de $\mathcal{B}$ devraient être plus faciles à convertir au point de vue $\mathcal{C}$ que ceux de $\mathcal{A}$; si l'on considère les divers résultats que nous devons établir pour conforter le point de vue $\mathcal{C}$, les différences entre $\mathcal{A}$ et $\mathcal{B}$ ne sont toutefois pas significatives.

En conclusion, nous pouvons dire que si, au début, notre robot émet à titre exploratoire des énoncés mathématiques — contrôlés par des procédures algorithmiques reposant largement sur une stratégie ascendante — dont la vérité

n'a aucune valeur définitive, nous devons cependant supposer qu'au bout d'un certain temps, il émet des énoncés — attestés par une sorte d'*imprimatur* que je désignerai ici par le symbole « ☆ » — qui lui sont dictés par des « convictions » mathématiques *inattaquables* et fondées sur ses *propres* critères. À la section 3.19, j'examinerai l'éventualité d'erreurs — mais rectifiables par le robot — intervenant lors de l'attribution du label ☆ à ces énoncés. Pour l'instant, nous supposerons que si le robot émet un énoncé ☆, cet énoncé est dépourvu d'erreur.

3.13 Les mécanismes sous-jacents aux raisonnements mathématiques des robots

Considérons maintenant les divers mécanismes intervenant dans les procédures qui gouverneraient le comportement du robot et lui permettraient de formuler ses assertions ☆. Certains de ces mécanismes seraient *internes* au robot. Il y aurait des contraintes internes de type descendant, réglant le fonctionnement du robot. Il y aurait également des procédures ascendantes prédéterminées lui permettant d'améliorer ses performances (et de parvenir progressivement au niveau des assertions ☆). Ces procédures seraient en principe connaissables par l'être humain (même si les conséquences ultimes de ces divers facteurs internes pourraient échapper aux capacités de calcul d'un mathématicien humain). De fait, si l'on affirme que l'être humain pourra un jour construire un robot capable de faire de vraies mathématiques, il faut que les mécanismes internes sur lesquels reposera la construction de ce robot *soient* connaissables par l'homme ; sinon, toute tentative de construction du robot sera irrémédiablement vouée à l'échec !

Bien entendu, cette construction pourrait connaître de nombreuses étapes : elle pourrait être effectuée entièrement par des robots de « niveau inférieur » (incapables de faire de vraies mathématiques) pouvant eux-mêmes avoir été construits par des robots de niveau encore plus bas. Toute cette hiérarchie aurait cependant été mise en branle par des êtres humains et selon des règles (probablement un mélange de procédures descendantes et ascendantes) qui seraient elles aussi nécessairement connaissables.

Parmi les ingrédients essentiels au développement du robot figureraient également les divers facteurs *externes* provenant de son environnement. Celui-ci pourrait en effet contribuer de manière déterminante à ce développement grâce à l'information fournie tant par les enseignants humains (ou robotiques) que par les corps matériels qu'il contient. L'information « naturelle » fournie par l'environnement non humain ne serait pas censée être « inconnaissable ». Elle pourrait certes être très complexe dans ses détails et avoir souvent un caractère interactif, mais nous disposons déjà, grâce à la « réalité virtuelle », de simulations efficaces d'importants aspects de notre environnement (*cf.* §1.20).

Rien ne semble interdire que l'on puisse étendre ces simulations de manière à fournir à notre robot, comme s'ils étaient des facteurs naturels externes, tous les éléments dont il aurait besoin pour son développement — en gardant à l'esprit (*cf.* §1.7, §1.9) qu'il suffit de simuler non un environnement réel, mais un environnement *typique*.

L'intervention des enseignants (humains ou robotiques) — les facteurs externes « artificiels » — pourrait se manifester à diverses étapes. On peut supposer qu'elle pourrait elle aussi être mécanisée, donc rien de nouveau au niveau de la connaissabilité de principe des mécanismes sous-jacents. Cette hypothèse de mécanisation est-elle justifiée ? Je suis persuadé qu'il n'apparaîtra pas excessif — du moins aux partisans de $\mathscr{A}$ et de $\mathscr{B}$ — de supposer que l'on puisse remplacer toute intervention humaine dans le développement du robot par une procédure entièrement algorithmique. Rien en effet n'exige que cet « enseignement » ait un caractère mystérieux — par exemple qu'il consiste en une sorte d'« influx » indéfinissable que l'enseignant humain transmettrait au robot pour faire son éducation. Il se trouve tout simplement que certaines informations de base nécessaires au robot sont de celles qu'un être humain peut communiquer sans problème. Ce processus a de fortes chances de s'avérer plus efficace s'il s'effectue sur le mode interactif, l'enseignant interagissant avec le robot à l'instar d'un précepteur dont le comportement est dicté par les réactions de son élève. Mais cela n'est pas contradictoire avec le fait que l'action de l'enseignant puisse prendre une forme algorithmique. Toute la discussion de ce chapitre s'articule en définitive autour d'une *démonstration par l'absurde* dont l'hypothèse de départ est justement que le comportement d'un être humain ne contient rien qui soit fondamentalement non algorithmique. Pour les partisans des points de vue $\mathscr{C}$ et $\mathscr{D}$, qui sont davantage portés à croire en la possibilité d'un « influx » non algorithmique transmis au robot par l'humanité même de son enseignant, cette discussion est en tout état de cause superflue !

Il ne semble pas raisonnable de supposer que tous ces mécanismes (*i.e.* les procédures de calcul internes et les données extraites de l'interaction avec l'environnement) soient inconnaissables, même si certaines personnes peuvent estimer que les conséquences détaillées des mécanismes externes peuvent ne pas être calculables par un être humain — voire par tout ordinateur actuel ou à venir. À la fin de la section 3.15, je reviendrai brièvement sur ce problème de la connaissabilité de ces mécanismes numériques. Mais pour l'instant, nous supposerons que l'ensemble de tous ces mécanismes — que je désignerai par la lettre **M** — est en fait connaissable. Se peut-il maintenant que certaines des assertions ☆ auxquelles peuvent conduire ces mécanismes *ne puissent* être connues d'un être humain ? Est-ce là une réelle éventualité ? Elle n'est pas réelle si nous continuons de donner à l'adjectif « connaissable » le sens de « connaissable *en principe* » adopté lors de la discussion des cas **I** et **II** et défini explicitement au début de la section 3.5. Le fait que quelque chose (*e.g.* la formulation d'une assertion ☆) puisse être hors de portée des capacités de *calcul* propres à un être humain ne nous concerne pas, que cet être humain s'aide d'une feuille de papier et d'un crayon, d'une calculette, voire même d'un

ordinateur standard programmé selon une stratégie descendante. L'intégration d'ingrédients ascendants aux procédures de calcul n'ajoute rien de nouveau à ce qu'il peut obtenir *en principe* — à condition que les *mécanismes* de base mis en jeu dans ces procédures ascendantes soient compréhensibles par l'être humain. En revanche, lorsque nous parlons de la connaissabilité des mécanismes **M** eux-mêmes, nous devons l'envisager « en pratique », dans le sens défini à la section 3.5. Ainsi, nous supposerons pour l'instant que les mécanismes **M** sont en fait des mécanismes connaissables *en pratique.*

Si l'on connaît les mécanismes **M**, nous pouvons supposer qu'ils constituent les éléments de base permettant de construire un *système formel* $\mathbb{Q}(\mathbf{M})$ dont les *théorèmes* seraient : (i) les assertions ☆ découlant de la mise en place de ces mécanismes ; (ii) les propositions pouvant être déduites de ces assertions ☆ à l'aide des lois de la logique élémentaire. Par « logique élémentaire », j'entends, par exemple, les règles du *calcul des prédicats* — *cf.* §2.9 — ou tout autre système de règles logiques (algorithmiques) dont la légitimité s'imposerait avec la force de l'évidence. La construction d'un tel système formel $\mathbb{Q}(\mathbf{M})$ se justifie par le fait que ces assertions ☆ se déduisent de **M**, individuellement, à l'aide d'une procédure algorithmique $Q(\mathbf{M})$ — bien que celle-ci soit assez lente en pratique. Remarquez que $Q(\mathbf{M})$, tel qu'il est défini, engendre les assertions ☆ de (i), mais pas nécessairement toutes celles de (ii) (car on peut supposer que notre robot s'ennuierait prodigieusement s'il avait à déduire toutes les conséquences logiques des théorèmes ☆ qu'il produit !). Ainsi, $Q(\mathbf{M})$ n'est pas strictement équivalente à $\mathbb{Q}(\mathbf{M})$, mais la différence n'est pas importante. Bien sûr, on pourrait, si on le désirait, étendre la procédure algorithmique $Q(\mathbf{M})$ de manière à en obtenir une autre qui serait, elle, équivalente à $\mathbb{Q}(\mathbf{M})$.

En ce qui concerne l'interprétation du système formel $\mathbb{Q}(\mathbf{M})$, il est clair qu'à mesure que se développera le robot, la présence du label ☆ sur une assertion *signifiera* — et continuera de signifier — que la validité de cette assertion est incontestablement établie. Si aucune information n'est transmise par l'enseignant humain (sous une forme ou une autre), nous ne pourrons être certains que le robot n'élaborera pas, par lui-même, un langage différent dans lequel le label ☆ aura une tout autre signification, si tant est qu'il en aura une. Pour garantir la cohérence du langage du robot avec notre propre définition de $\mathbb{Q}(\mathbf{M})$, nous devrons nous assurer que lors de l'apprentissage du robot (par exemple, au contact de l'enseignant humain), le sens attaché au label ☆ ne variera pas. D'une manière analogue, nous devrons nous assurer que les notations utilisées par le robot pour formuler, par exemple, ses énoncés Π_1, seront identiques à (ou explicitement traduisibles en) celles que nous-mêmes nous utilisons. Si les mécanismes de **M** nous sont connaissables, il en résulte que les axiomes et les règles de procédure du système formel $\mathbb{Q}(\mathbf{M})$ sont également connaissables. En outre, tout théorème que l'on peut obtenir au sein de $\mathbb{Q}(\mathbf{M})$ doit, *en principe*, être connaissable (dans le sens où sa formulation est connaissable, même si sa vérité ne l'est pas nécessairement), les procédures algorithmiques permettant d'obtenir nombre de ces théorèmes pouvant échapper aux capacités de calcul de l'être humain.

3.14 La contradiction fondamentale

La discussion précédente a montré que l'« algorithme inconscient et inconnaissable F » dont **III** suppose qu'il sous-tend la perception des vérités mathématiques peut se réduire à un algorithme consciemment connaissable — à condition que l'on puisse, conformément aux objectifs de l'IA, mettre en action un système de procédures aboutissant à la construction d'un robot capable de faire des mathématiques d'un niveau égal à (voire dépassant) celui des mathématiques d'un mathématicien humain. L'algorithme inconnaissable F est ainsi remplacé par un système formel connaissable $\mathbb{Q}(\mathbf{M})$.

Avant d'examiner cela en détail, je voudrais attirer votre attention sur un problème important qu'à ma connaissance personne n'a encore correctement abordé, à savoir l'éventualité qu'au lieu de correspondre à un ensemble de mécanismes fixes, le développement du robot puisse dépendre de l'introduction, de temps à autre, d'*éléments aléatoires*. Ce problème sera examiné le moment venu, mais pour l'instant, je considérerai simplement que tous ces éléments aléatoires sont susceptibles d'être générés par un calcul *pseudo*-aléatoire (chaotique). Nous l'avons vu (*cf.* §1.9, §3.11), ces ingrédients pseudo-aléatoires devraient, en pratique, convenir. Je donnerai à la section 3.18 une discussion plus complète du problème des entrées véritablement aléatoires ; pour l'instant, lorsque je parlerai des « mécanismes **M** », je supposerai qu'ils sont en fait purement algorithmiques et exempts de toute incertitude.

L'idée centrale de notre contradiction est, schématiquement parlant, que $\mathbb{Q}(\mathbf{M})$ doit en fait prendre la place du « F » intervenu dans les discussions précédentes, en particulier dans celle de la section 3.2 à propos du cas **I**. Il en résulte que le cas **III** se réduit effectivement à **I** et se trouve donc exclu. Nous supposons — conformément aux points de vue $\mathcal{A}$ ou $\mathcal{B}$ — que notre robot pourrait, *en principe*, grâce à des procédures d'apprentissage analogues à celles que nous avons définies, aboutir finalement à tout résultat mathématique qu'un être humain pourrait lui-même obtenir. Nous supposons qu'il *pourrait* également obtenir des résultats *inaccessibles* en principe aux capacités de calcul de l'être humain. Dans chacun de ces deux cas, le robot devrait être en mesure de percevoir la légitimité du raisonennement gödelien (ou du moins *simuler* qu'il perçoit cette légitimité — selon le point de vue $\mathcal{B}$). Ainsi, pour tout système formel (suffisamment vaste) $\mathbb{H}$, il lui apparaîtrait avec évidence que si $\mathbb{H}$ est sûr, la proposition gödelienne* $G(\mathbb{H})$ est alors non seulement vraie, mais aussi qu'elle n'est pas un théorème de $\mathbb{H}$. En particulier, le robot percevrait que la vérité de $G(\mathbb{Q}(\mathbf{M}))$ découle avec évidence de la sûreté de $\mathbb{Q}(\mathbf{M})$, mais aussi que la sûreté de $\mathbb{Q}(\mathbf{M})$ entraîne avec la même évidence que $G(\mathbb{Q}(\mathbf{M}))$ n'est pas un théorème de $\mathbb{Q}(\mathbf{M})$.

En reprenant exactement le raisonnement fait dans le cas **I** (pour les êtres humains — *cf.* §3.2), il s'ensuit immédiatement que le robot ne peut avoir la

* Dans les tirages précédents de cet ouvrage, $\Omega(\mathbb{F})$ était utilisé à la place de $G(\mathbb{F})$ dans le reste du chapitre 3. Il est cependant plus correct d'utiliser $G(\mathbb{F})$. (*cf.* §2.8 et p. 89.)

ferme conviction que le système formel $\mathbb{Q}(\mathbf{M})$ *est* équivalent à sa propre notion de vérité mathématique incontestable, et ce en dépit du fait que *nous* (c'est-à-dire les vrais spécialistes de l'IA) sachions que les mécanismes **M** *sous-tendent* l'ensemble des convictions mathématiques du robot, et donc que cet ensemble *est* équivalent à $\mathbb{Q}(\mathbf{M})$. Car si le robot croyait fermement que ses convictions sont réductibles à $\mathbb{Q}(\mathbf{M})$, il devrait alors également croire que $\mathbb{Q}(\mathbf{M})$ est sûr. Par conséquent, il devrait croire que $G(\mathbb{Q}(\mathbf{M}))$ est vraie, et donc que $G(\mathbb{Q}(\mathbf{M}))$ est irréductible à l'ensemble de ses convictions — ce qui est une contradiction ! Ainsi, le robot ne peut savoir qu'il a été construit à l'aide des mécanismes **M**. Puisque *nous*, nous savons — ou du moins, pouvons savoir — qu'il a été construit à l'aide de ces mécanismes, cela semble indiquer que nous avons accès à des vérités mathématiques, *e.g.* $G(\mathbb{Q}(\mathbf{M}))$, qui sont inaccessibles aux capacités de calcul du robot, en dépit du fait que ces capacités soient censées être égales (voire être supérieures) à celles de l'être humain.

3.15 Comment lever cette contradiction ?

On peut envisager les choses sous deux angles différents : du point de vue des créateurs du robot, ou du point de vue du robot lui-même. Du point de vue des créateurs, rien n'interdit qu'un mathématicien humain doute des prétentions du robot lorsqu'il affirme détenir la vérité vraie, sauf si cet être humain admet chacun des *arguments* utilisés par le robot. Ce mathématicien pourrait contester certains théorèmes de $\mathbb{Q}(\mathbf{M})$ — rappelons que les capacités de raisonnement du robot peuvent éventuellement *excéder* celles de l'être humain. Ainsi, on pourrait affirmer que le simple fait de savoir que le robot a été construit à l'aide des mécanismes **M** peut, du point de vue humain, ne pas être un gage de la validité d'une démonstration mathématique. Examinons donc cette contradiction du point de vue du *robot*. Quelles issues s'offrent à lui ?

Il semble avoir essentiellement quatre grandes possibilités — en supposant qu'il reconnaisse qu'il *est* lui-même un robot numérique.

(a) Bien qu'acceptant que sa propre construction *puisse* reposer sur les mécanismes **M**, le robot n'est pas *fermement* convaincu de ce fait.

(b) Bien qu'étant fermement convaincu, à l'instant où il les formule, de l'incontestable légitimité de chacune de ses assertions ☆, le robot doute de la validité de la *totalité* de ces assertions ☆ — et donc reste peu convaincu que ses convictions à l'égard des énoncés Π_1 reposent *entièrement* sur $\mathbb{Q}(\mathbf{M})$.

(c) Peut-être les vrais mécanismes **M** dépendent-ils essentiellement d'éléments *aléatoires* et ne peuvent-ils être correctement décrits à l'aide d'entrées numériques pseudo-aléatoires connues.

(d) Peut-être les vrais mécanismes **M** sont-ils en fait *in*connaissables.

Au cours des neuf prochaines sections, je vais tenter de démontrer qu'aucune des trois possibilités (a), (b) et (c) ne permet vraiment au robot d'échapper à la contradiction. Si l'on soutient que la compréhension mathématique se réduit à des opérations algorithmiques, il est contraint — tout comme nous — d'accepter à contrecœur la possibilité (d). Je suis sûr que ceux qui s'intéressent à l'intelligence artificielle considèrent comme moi que (d) est une éventualité très peu satisfaisante. Elle conforte peut-être le point de vue synthétisé par la thèse $\mathcal{A}/\mathcal{D}$ mentionnée à la fin de la section 1.3 et selon laquelle l'implantation d'un algorithme inconnaissable dans notre cerveau-ordinateur exige l'*intervention de Dieu* (le « meilleur programmeur de la profession »). Mais elle n'est guère réconfortante pour ceux qui espèrent *construire* un robot réellement artificiellement intelligent ! Elle est également plutôt démoralisante pour ceux d'entre nous qui espèrent comprendre, dans ses principes, l'émergence de l'intelligence chez l'homme à l'aide de lois scientifiques compréhensibles, telles celles de la physique, de la chimie, de la biologie et de la sélection naturelle — indépendamment de tout désir de recréer une telle intelligence à l'intérieur d'un robot. Selon moi, une conclusion aussi pessimiste n'est cependant pas inévitable, pour la simple raison que la « compréhensibilité scientifique » est très différente de la « calculabilité ». Il est plus probable que la vraie conclusion est non pas que les lois sous-jacentes à l'intelligence sont incompréhensibles, mais qu'elles sont *non algorithmiques*. J'en dirai plus à ce sujet dans la deuxième partie de ce livre.

3.16 Le robot doit-il croire en M ?

Imaginons que nous présentions au robot un ensemble de mécanismes **M** — éventuellement, mais non nécessairement, ceux qui ont présidé à sa construction. Je vais tenter de convaincre le lecteur que le robot ne peut que rejeter l'éventualité que ces mécanismes puissent sous-tendre sa compréhension mathématique — *indépendamment du fait* qu'il en soit ou non effectivement ainsi ! Si nous supposons pour l'instant que le robot rejette les possibilités (b), (c) et (d), nous sommes amenés à conclure, curieusement, que (a) ne peut, par elle-même, nous permettre d'échapper au paradoxe.

Le raisonnement est le suivant. Soit $\mathcal{M}$ l'hypothèse :

> « Les mécanismes **M** sous-tendent la compréhension mathématique du robot. »

Considérons maintenant les assertions de la forme :

> « Tel énoncé Π_1 est une conséquence de $\mathcal{M}$. »

Si le robot est fermement convaincu de la validité d'une telle assertion, je la qualifierai d'assertion $\bigstar_{\mathcal{M}}$. Autrement dit, les assertions $\bigstar_{\mathcal{M}}$ ne se réfèrent pas

nécessairement aux énoncés Π_1 que le robot considère comme incontestablement vrais en eux-mêmes, mais à ceux dont il considère qu'ils se déduisent indéniablement de l'hypothèse $\mathscr{M}$. Le robot n'a pas initialement besoin d'imaginer le moins du monde qu'il a été *effectivement* construit à l'aide de **M**. Il peut même penser que c'est là une possibilité improbable, tout en pouvant parfaitement envisager — dans la vraie tradition scientifique — les conséquences que l'on peut déduire en supposant simplement à titre d'*hypothèse* qu'il a été construit ainsi.

Existe-t-il des énoncés Π_1 que le robot considère comme des conséquences inévitables de $\mathscr{M}$, mais qui cependant ne sont pas des assertions ☆ ? La réponse est oui. Car, nous l'avons vu à la fin de la section 3.14, la vérité de l'énoncé Π_1 $G(\mathbb{Q}(\mathbf{M}))$ et le fait $G(\mathbb{Q}(\mathbf{M}))$ n'est pas un théorème de $\mathbb{Q}(\mathbf{M})$ découlent de la sûreté de $\mathbb{Q}(\mathbf{M})$. En outre, le robot est fermement convaincu de la légitimité de ces deux déductions. Si l'on suppose qu'il admet que ses convictions inébranlables seraient contenues dans $\mathbb{Q}(\mathbf{M})$ s'il avait été construit à l'aide de **M** — autrement dit qu'il rejette la possibilité (b)* —, il est alors nécessairement convaincu que la sûreté de $\mathbb{Q}(\mathbf{M})$ est une conséquence de $\mathscr{M}$. Ainsi, il est fermement persuadé non seulement que l'énoncé Π_1 $G(\mathbb{Q}(\mathbf{M}))$ découle de l'hypothèse $\mathscr{M}$, mais aussi (lorsqu'il accepte $\mathscr{M}$) que la validité de cet énoncé Π_1 n'est pas immédiatement perceptible si l'on ne recourt pas à $\mathscr{M}$ (car $\mathscr{M}$ n'appartient pas à $\mathbb{Q}(\mathbf{M})$). Ainsi, $G(\mathbb{Q}(\mathbf{M}))$ est une assertion $☆_{\mathscr{M}}$, mais non une assertion ☆.

Supposons maintenant que le système formel $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ soit construit exactement de la même manière que $\mathbb{Q}(\mathbf{M})$, sauf que ce sont maintenant les assertions $☆_{\mathscr{M}}$ qui jouent le rôle que les assertions ☆ jouaient dans la construction de $\mathbb{Q}(\mathbf{M})$. Autrement dit, les théorèmes de $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ sont soit (i) les assertions $☆_{\mathscr{M}}$ elles-mêmes, soit (ii) les propositions obtenues à partir de ces assertions $☆_{\mathscr{M}}$ à l'aide de la logique élémentaire (*cf.* §3.13). De même que lorsqu'il accepte l'hypothèse $\mathscr{M}$, le robot admet que $\mathbb{Q}(\mathbf{M})$ englobe ses convictions inébranlables à l'égard de la vérité des énoncés Π_1, de même il admet que le système $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ englobe ses convictions inébranlables à l'égard de la vérité des énoncés Π_1 *dépendants* de l'hypothèse $\mathscr{M}$.

Supposons ensuite qu'il examine l'énoncé Π_1 de Gödel $G(\mathbb{Q}_{\mathscr{M}}(\mathbf{M}))$. Il serait certainement résolument convaincu que cet énoncé Π_1 découle de la sûreté de $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$. Il croirait tout aussi fermement que cette sûreté est une conséquence de $\mathscr{M}$, car il admet que $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ *contient* ses convictions les plus intimes à l'égard de son aptitude à déduire des énoncés Π_1 à partir de l'hypothèse $\mathscr{M}$. (Il raisonnerait ainsi : « Si j'accepte $\mathscr{M}$, j'accepte alors tous les énoncés Π_1 qui engendrent le système $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$. Ainsi, si j'accepte $\mathscr{M}$, je dois accepter que $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ est sûr. Par conséquent, si j'accepte $\mathscr{M}$, je dois accepter que $G(\mathbb{Q}_{\mathscr{M}}(\mathbf{M}))$ est vrai. »)

---

* Bien entendu, la possibilité (d) n'intervient pas ici, car le robot *connaît* **M** ; en outre, cet ensemble **M** étant pour l'instant exempt d'éléments purement aléatoires, (c) n'intervient pas non plus.

Mais s'il croit (intimement) que l'énoncé Π_1 gödelien $G(\mathbb{Q}_{\mathscr{M}}(\mathbf{M}))$ *est* une conséquence de $\mathscr{M}$, il doit également croire que $G(\mathbb{Q}_{\mathscr{M}}(\mathbf{M}))$ est un théorème de $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$. Et il ne peut croire cela que s'il croit que $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ *n'est pas sûr* — ce qui contredit manifestement le fait qu'il accepte $\mathscr{M}$!

On a implicitement supposé, en certains endroits du raisonnement précédent, que l'ensemble des convictions du robot *est* en fait sûr — bien que ce qui soit effectivement requis est que le robot *croie* que ses convictions sont sûres. Dans les deux cas, le robot est censé jouir d'une compréhension mathématique d'un niveau au moins égal à celui d'un être humain ; et, ainsi qu'il a été dit à la section 3.4, la compréhension mathématique de l'être humain est en principe sûre.

L'hypothèse $\mathscr{M}$ et la définition d'une assertion $\stackrel{\star}{\mathscr{M}}$ semblent contenir un certain flou. Il faut toutefois souligner que parce qu'elle est un énoncé Π_1, une assertion $\stackrel{\star}{\mathscr{M}}$ est mathématiquement parfaitement définie. On pourrait imaginer que la plupart des assertions $\stackrel{\star}{\mathscr{M}}$ que le robot peut formuler sont en fait des assertions $\stackrel{\star}{}$ ordinaires, car il est improbable qu'il juge utile d'invoquer l'hypothèse $\mathscr{M}$ pour chacune d'elles. L'énoncé $G(\mathbb{Q}(\mathbf{M}))$ ferait cependant exception, car pour le robot, $\mathbb{Q}(\mathbf{M})$ joue le rôle de la « machine à prouver les théorèmes » envisagée par Gödel (*cf.* §3.1 et §3.3). En présence de $\mathscr{M}$, le robot a accès à sa propre « machine à prouver les théorèmes », et bien qu'il ne puisse être fermement convaincu (il ne peut en fait l'être) de la sûreté de cette machine, il pourrait envisager qu'elle *soit* sûre et tenter de déduire les conséquences de cette hypothèse.

À ce stade, le robot n'est pas plus proche du paradoxe relevé par Gödel pour les êtres humains (voir sa citation à la section 3.1). Toutefois, puisqu'il a connaissance non seulement du système formel $\mathbb{Q}(\mathbf{M})$, mais aussi des hypothétiques *mécanismes* $\mathbf{M}$, il peut reproduire le raisonnement et passer de $\mathbb{Q}(\mathbf{M})$ à $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$, système dont la sûreté lui apparaît aussi n'être qu'une conséquence de l'hypothèse $\mathscr{M}$. Et c'est *cela* qui le conduit à la contradiction (recherchée). (Voir aussi la section 3.24 pour une discussion plus approfondie du système $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ et de son lien apparent avec le « raisonnement paradoxal ».)

Il s'ensuit qu'aucun être mathématiquement conscient — aucun être capable d'une authentique compréhension mathématique — ne peut fonctionner selon un ensemble de mécanismes, quel qu'il soit, accessible à son entendement, indépendamment du fait qu'il *sache* ou non que ces mécanismes sont censés guider son propre cheminement vers la découverte de vérités mathématiques incontestables. (Rappelons également que ces « vérités mathématiques incontestables » correspondent simplement aux résultats qu'il peut établir mathématiquement — *i.e.* à l'aide de « démonstrations mathématiques », et non nécessairement de preuves « formelles ».)

Ainsi, le raisonnement précédent nous amène à conclure qu'il n'existe aucun ensemble de mécanismes algorithmiques connaissables par un robot, exempts d'ingrédients purement aléatoires, et dont le robot pourrait accepter, ne serait-ce qu'à titre de *possibilité*, qu'ils sous-tendent l'ensemble de ses convictions mathématiques — *à condition* que le robot reconnaisse que les procédures spécifiques que j'ai proposées pour construire le système formel $\mathbb{Q}(\mathbf{M})$

à partir des mécanismes **M** produisent *réellement* tous les énoncés Π_1 qu'il considère comme indéniablement vrais — et, corrélativement, que le système formel $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ contient tous les énoncés Π_1 qui, selon lui, découlent indubitablement de l'hypothèse $\mathscr{M}$.

Reste cependant la possibilité — (c) — que le robot ne puisse parvenir à un système de convictions mathématiques potentiellement cohérent sans que l'on intègre des ingrédients véritablement aléatoires dans les mécanismes **M**. Elle sera examinée dans les sections qui suivent (§3.17-§3.22) et, par souci de commodité, dans le cadre de la discussion générale de la possibilité (b). Mais celle-ci ne peut être sérieusement abordée sans reconsidérer en premier lieu le problème crucial des « convictions » du robot — que la fin de la section 3.12 n'a fait qu'effleurer.

3.17 Des erreurs du robot et du sens de ses assertions ☆

Le robot est-il prêt à reconnaître comme une évidence que *si* sa construction repose sur un ensemble de mécanismes **M** — respectivement, s'il admet l'hypothèse $\mathscr{M}$ —, le système formel $\mathbb{Q}(\mathbf{M})$ — respectivement $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ — exprime correctement ses convictions à l'égard des énoncés Π_1 ? Pour qu'il en soit ainsi, il faut avant tout qu'il soit prêt à croire que $\mathbb{Q}(\mathbf{M})$ est *sûr* — autrement dit, il doit croire que tous les énoncés Π_1 qui sont des assertions ☆ sont réellement *vrais*. Dans les arguments que j'ai présentés, il faut également que *tout* énoncé Π_1 que le robot considère comme incontestablement vrai soit en fait un théorème de $\mathbb{Q}(\mathbf{M})$ (de sorte que $\mathbb{Q}(\mathbf{M})$ permettrait au robot de définir une « machine à prouver les théorèmes » analogue à celle que Gödel avait envisagée pour les mathématiciens humains ; *cf.* §3.1, §3.3). De fait, il n'est *pas* essentiel que $\mathbb{Q}(\mathbf{M})$ — respectivement $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ — ait ce rôle de référence obligée pour les aptitudes potentielles du robot à juger les énoncés Π_1 ; il suffit qu'il soit suffisamment vaste pour permettre une formulation du raisonnement gödelien applicable au système $\mathbb{Q}(\mathbf{M})$ — respectivement $\mathbb{Q}_{\mathscr{M}}(\mathbf{M})$ — lui-même. Nous verrons plus loin que c'est là une condition relativement précise — et qu'il suffit de l'appliquer à un système d'énoncés Π_1 *fini*.

Ainsi, nous devons — tout comme le robot — accepter l'éventualité que les assertions ☆ soient parfois erronées, bien que le robot puisse les rectifier en se fondant sur ses propres critères. L'idée est que le comportement du robot est tout à fait semblable à celui d'un mathématicien humain. Il arrive qu'un mathématicien humain croie fermement à la vérité (ou à la fausseté) d'un énoncé Π_1 donné, puis se rende compte que le raisonnement qui a abouti à cet énoncé contient une erreur. C'est seulement alors qu'il perçoit la fausseté de ce raisonnement, et ce en vertu des mêmes critères qui lui avaient fait croire, à tort, que le raisonnement était correct. Ainsi, un énoncé Π_1 qui était

d'abord apparu comme incontestablement vrai peut se révéler incontestablement faux (et vice versa).

On peut escompter que le robot ait un comportement analogue et que l'on ne puisse dès lors se fier à ses assertions ☆, même s'il leur a décerné le label d'imprimatur « ☆ ». Par la suite, il pourrait certes corriger son erreur, mais celle-ci n'en aurait pas moins été commise. En quoi cela affecte-t-il notre conclusion concernant la sûreté du système formel $\mathbb{Q}(\mathbf{M})$? Manifestement, $\mathbb{Q}(\mathbf{M})$ n'est maintenant plus entièrement sûr, et le robot ne peut même pas le « percevoir » comme entièrement sûr, de sorte que l'on peut douter de la validité de la proposition gödelienne $G(\mathbb{Q}(\mathbf{M}))$. C'est essentiellement cette conclusion que sous-entend la possibilité (b).

Revenons sur la signification qu'il faut attribuer au fait que notre robot parvient à des vérités mathématiques « incontestables ». Nous allons comparer cette situation à celle que nous avons examinée dans le cas d'un mathématicien humain. Rappelons que lors de cet examen, nous avons vu qu'il fallait envisager non pas ce qu'un mathématicien peut affirmer *en pratique*, mais ce qu'il peut, *en principe*, considérer comme une vérité mathématique incontestable. (Rappelons-nous aussi la formule de Feynman : « N'écoutez pas ce que je dis, écoutez ce que je *veux dire* ! ») Nous devons donc concentrer notre attention sur ce que notre robot *veut dire* plutôt que sur ce qu'il dit. Mais, pour un partisan de $\mathscr{B}$ encore plus que pour un partisan de $\mathscr{A}$, l'idée même que le robot *veuille dire* quelque chose n'admet pas d'interprétation claire. Si l'on pouvait se fier, non aux assertions ☆ du robot, mais à ce qu'il « veut dire » réellement ou à ce qu'il « voudrait dire » en principe, le problème de l'éventuelle inexactitude de ses assertions ☆ se trouverait réglé. Malheureusement, il semble que nous n'ayons aucun moyen de percevoir extérieurement le « sens » ou le « sens implicite » de ses assertions. Si nous nous en tenons au système formel $\mathbb{Q}(\mathbf{M})$, nous devons apparemment accepter les assertions ☆ elles-mêmes, sans être totalement certains de leur validité.

Peut-être les deux points de vue $\mathscr{A}$ et $\mathscr{B}$ entraînent-ils sur ce plan des conséquences véritablement distinctes, car bien que les tenants de $\mathscr{A}$ et ceux de $\mathscr{B}$ s'accordent sur ce que peut accomplir extérieurement un système physique, ils sont apparemment en désaccord sur le type de système numérique qui pourrait fournir une simulation correcte de l'activité cérébrale d'une personne en train de percevoir la validité d'un raisonnement mathématique (voir la fin de la section 3.12). Toutefois, cette différence de vue n'intervient pas particulièrement dans la discussion qui nous occupe.

3.18 Comment intégrer les ingrédients aléatoires ; ensemble de l'activité de tous les robots

En l'absence de moyens efficaces pour traiter ces problèmes de sémantique, nous devons nous reposer sur les assertions ☆ formulées par le robot à l'aide des mécanismes contrôlant son comportement. Nous devons accepter que certaines de ces assertions ☆ puissent contenir des erreurs, bien que ces erreurs soient rectifiables et, en tout état de cause, extrêmement rares. Il semble raisonnable de supposer que chaque fois que le robot commet une erreur, celle-ci est, du moins en partie, imputable à des facteurs de hasard présents dans son environnement ou ses mécanismes internes. On peut imaginer qu'un second robot, fonctionnant selon les mêmes mécanismes que le premier mais pour lequel ces facteurs seraient différents, ne ferait pas les erreurs commises par le premier — bien qu'il puisse en faire d'autres. Ces facteurs de hasard pourraient être des ingrédients purement aléatoires participant soit des données fournies au robot par l'environnement, soit des spécifications de ses mécanismes internes. Ce pourraient également être des ingrédients pseudo-aléatoires, d'ordre externe ou interne, résultant d'un processus de calcul déterministe mais chaotique.

Pour les besoins de la discussion, je supposerai que l'action de chacun de ces ingrédients pseudo-aléatoires ne se distingue pas, du moins au niveau de ses effets, de celle d'un ingrédient purement aléatoire. Cette hypothèse traduit le point de vue habituel. Reste cependant la possibilité que le comportement des systèmes chaotiques renferme un élément qui — allant *au-delà* de la simulation du hasard — serait une approximation d'une forme utile de comportement non algorithmique. À ma connaissance, une telle éventualité n'a jamais été sérieusement envisagée, en dépit du fait que certaines personnes soient persuadées que le comportement chaotique est un aspect fondamental de l'activité cérébrale. Selon moi, cette idée restera cependant peu convaincante tant que l'on n'aura pas démontré la présence, dans les systèmes chaotiques, d'un comportement fondamentalement *non* aléatoire (ou plutôt, non pseudo-aléatoire) — qui, au sens fort, serait une bonne approximation d'un comportement authentiquement non algorithmique. Je ne connais à ce jour aucune ébauche de démonstration qui aille dans ce sens. Comme nous le verrons plus loin (§3.22), il est cependant très improbable qu'un comportement chaotique permette d'échapper aux difficultés que les arguments de type gödelien soulèvent pour les modèles algorithmiques de l'esprit.

Supposons pour l'instant que tout élément pseudo-aléatoire (ou chaotique) présent dans les mécanismes internes du robot ou dans son environnement puisse, sans la moindre perte d'efficacité, être remplacé par un élément purement aléatoire. Pour analyser l'action de ces éléments, nous devons considérer l'*ensemble* de tous les robots possibles. Puisque nous supposons que notre robot est contrôlé numériquement et qu'en outre, l'action de son environnement est pareillement exprimable sous forme de données numériques (rappelez-vous les deux zones, « interne » et « externe », définies sur le ruban de la

machine de Turing ; *cf.* aussi §1.8), ces robots possibles seront en nombre *fini*. Ils peuvent certes s'avérer extrêmement nombreux, mais leur description n'en reste pas moins une affaire de calcul. Ainsi, l'ensemble de tous les robots possibles, chacun fonctionnant selon des mécanismes que nous avons nous-mêmes définis, constitue lui aussi un système algorithmique — bien que ne pouvant, en pratique, assurément être simulé par aucun ordinateur envisageable aujourd'hui. Cependant, bien qu'elle serait impossible à réaliser, cette simulation de l'action simultanée de tous les robots possibles fonctionnant selon les mécanismes **M** resterait dans le domaine du « connaissable » ; autrement dit, on saurait (théoriquement) construire un ordinateur — une machine de Turing — qui effectuerait cette simulation, bien qu'il soit impensable de l'effectuer *concrètement*. C'est là un point capital de notre discussion. Un mécanisme connaissable ou un calcul connaissable se caractérisent par le fait qu'ils peuvent être *spécifiés* par un être humain ; peu importe qu'ils soient ou non le résultat d'un calcul que pourrait effectuer un être humain, voire un ordinateur pouvant être construit concrètement. Ce point est très semblable à celui que nous avons rencontré plus haut lors de l'examen de **Q8** et est cohérent avec la terminologie introduite au début de la section 3.5.

3.19 La suppression des assertions ☆ erronées

Revenons maintenant au problème des assertions ☆ erronées (rectifiables) émises de temps en temps par notre robot. Supposons que celui-ci vienne effectivement d'émettre un tel énoncé erroné. Si l'on suppose qu'un autre robot, ou le même robot un instant plus tard — ou une autre *matérialisation* de ce même robot — ne ferait probablement pas la même erreur, nous pouvons, *en principe*, reconnaître la nature erronée d'une telle assertion ☆ en examinant la simulation de tous les comportements possibles de notre ensemble de robots. On peut effectuer cette simulation de sorte qu'au lieu de survenir au fil du temps, les diverses matérialisations du robot surviennent toutes simultanément. (C'est là simplement un moyen pratique de représenter les choses. Il n'exige pas que notre simulation opère vraiment de façon « parallèle ». Nous l'avons vu d'ailleurs, rien en principe ne distingue une action parallèle d'une action séquentielle — hormis l'efficacité opératoire ; *cf.* §1.5.) L'idée est qu'en examinant le résultat de cette simulation, on devrait pouvoir, en principe, séparer le nombre comparativement petit d'assertions ☆ erronées de la multitude des assertions ☆ correctes en tirant avantage du fait que les assertions ☆ erronées sont « rectifiables » et sont donc perçues comme erronées par l'immense majorité des matérialisations du robot participant à la simulation — du moins à mesure que l'acquisition des « expériences » parallèles des diverses matérialisations du robot se déroule dans le temps (simulé). Je ne dis pas que cette procédure soit concrètement réalisable, mais simple-

ment qu'elle est algorithmique, dans le sens où les *règles* **M** qui sous-tendent tout ce calcul sont en principe « connaissables ».

Pour que notre simulation soit plus proche de ce qui se passerait avec la communauté des mathématiciens humains, et pour être doublement certains que toutes les erreurs présentes dans les assertions ☆ ont bien été supprimées, décomposons l'environnement de notre robot en, d'une part, la *communauté* des autres robots et, d'autre part, l'environnement résiduel dépourvu de robots (et d'êtres humains) — en admettant qu'outre cet environnement résiduel, il puisse y avoir quelques enseignants, du moins dans les premières phases du développement des robots, de sorte, notamment, que les robots aient une conscience claire de la signification de leur geste lorsqu'ils décernent le label d'imprimatur ☆. Tous les comportements possibles de *tous* ces autres robots, plus tous les environnements résiduels (pertinents) possibles, plus toutes les données (pertinentes) possibles fournies par les enseignants, tout cela variant en fonction des différents choix des paramètres aléatoires mis en jeu, contribuent aux différentes matérialisations composant la simulation de notre ensemble. Ici encore, on peut considérer que les règles — que je désignerai encore par **M** — sont parfaitement connaissables, malgré la rebutante complexité du détail des calculs qu'il faudrait mettre en œuvre si l'on effectuait réellement la simulation.

Supposons que nous consignions (en principe) tous les énoncés Π_1 — ou leurs négations — émis avec le label ☆ par les diverses matérialisations des robots (simulés numériquement). Comment distinguer ceux d'entre eux qui sont *exempts d'erreur* ? Nous pouvons procéder en *ignorant* toute assertion ☆ concernant un énoncé Π_1 sauf si, pendant un intervalle de temps T orienté vers le passé ou le futur, le nombre r des diverses occurrences de cette assertion ☆, dans l'ensemble de toutes les simulations simultanées, vérifie $r > L + Ns$, où L et N sont des nombres suffisamment grands et où s est le nombre d'assertions ☆ qui, durant le même intervalle de temps, donnent un jugement inverse sur l'énoncé Π_1 ou simplement affirment que le raisonnement sous-jacent à l'assertion ☆ originelle est erroné. On peut éventuellement exiger que T (qui ne correspond pas nécessairement à une simulation du temps « réel » mais peut être un temps mesuré en unités adaptées aux procédures de calcul), L et N croissent avec la « complexité » de l'énoncé Π_1 émis avec le label ☆.

Cette notion de « complexité » des énoncés Π_1 admet une définition précise en termes de spécifications de machines de Turing (voir la section 2.6, fin de la réponse à **Q8**) : si l'on reprend les formules explicites données au chapitre 2 de EOLP et brièvement rappelées dans l'appendice A du présent livre, le *degré de complexité* d'un énoncé Π_1 affirmant le non-arrêt de l'action $T_m(n)$ d'une machine de Turing T_m sur un nombre n est égal au nombre ρ de chiffres binaires contenus dans *le plus grand* des deux nombres m et n.

La raison pour laquelle on inclut le nombre L dans l'inégalité ci-dessus, plutôt que de s'en tenir à une sorte de majorité écrasante donnée par le seul grand nombre N, est que l'on doit tenir compte de l'éventualité suivante. Supposez qu'à intervalles très espacés apparaisse dans notre ensemble de robots un robot

« fou » qui sorte une « assertion ☆ » totalement insensée, qui ne soit jamais venue à l'« esprit » des autres robots — une assertion si absurde qu'aucun des autres robots ne songe même à en exprimer la négation ! Sans l'inclusion du facteur L, cette assertion ☆ serait, selon notre critère, considérée comme « exempte d'erreur ». Mais avec un L suffisamment grand, cette éventualité ne se présente pas — à condition, répétons-le, que ces cas de « folie » soient passablement rares. (Il se peut bien entendu que j'aie oublié d'autres éventualités de ce type qui nécessiteraient d'autres précautions. Mais pour l'instant du moins, il semble raisonnable de procéder en suivant les critères que je viens de donner.)

Si l'on garde à l'esprit que les assertions ☆ sont de toute façon, *a priori*, des affirmations « parfaitement valides » — fondées sur une logique implacable excluant tout élément sur lequel le robot aurait le moindre doute —, on peut raisonnablement considérer que la procédure ci-dessus, dans laquelle les fonctions $T(\rho)$, $L(\rho)$ et $N(\rho)$ n'ont pas besoin de sortir de l'ordinaire, permet d'éliminer toute erreur qui aurait tout de même pu se glisser dans le raisonnement du robot. Cela étant admis, on a à nouveau affaire à un système *algorithmique* — un système *connaissable* (en ce sens que ses *règles* sous-jacentes sont connaissables) à partir du moment où les mécanismes originels **M** gouvernant le comportement du robot sont connaissables. Ce système algorithmique engendre un nouveau système formel (connaissable) $\mathbb{Q}'(\mathbf{M})$, dont les théorèmes sont ces assertions ☆ *exemptes d'erreurs* (ou des assertions pouvant être obtenues à partir d'elles à l'aide des simples opérations logiques du calcul des prédicats).

En fait, l'important pour notre objectif n'est pas tellement que ces assertions soient *réellement* exemptes d'erreurs, mais que les robots eux-mêmes soient *convaincus* qu'elles le sont (n'oubliez pas toutefois que pour les partisans de $\mathcal{B}$, cette « conviction » a pour seule réalité le fait que l'on peut la *simuler* ; *cf.* §3.12, §3.17)

Plus précisément, il faut que les robots soient prêts à croire, en faisant l'*hypothèse* que ce sont les mécanismes **M** qui sous-tendent leur comportement — l'hypothèse $\mathcal{M}$ de la section 3.16 —, que ces assertions ☆ sont exemptes d'erreurs. Dans la présente section, nous nous sommes jusqu'ici intéressés à la suppression des éventuelles erreurs contenues dans les assertions ☆ du robot. Mais en vérité, ce qui nous intéresse *réellement* du point de vue de la contradiction fondamentale présentée à la section 3.16, c'est la suppressions des erreurs contenues dans les assertions $☆_{\mathcal{M}}$, à savoir dans les énoncés Π_1 dont le robot est convaincu qu'ils découlent logiquement de l'hypothèse $\mathcal{M}$. L'acceptation du système $\mathbb{Q}'(\mathbf{M})$ par les robots étant conditionnée par l'acceptation de l'hypothèse $\mathcal{M}$, nous pouvons aussi bien considérer qu'ils envisagent un système plus vaste $\mathbb{Q}'_{\mathcal{M}}(\mathbf{M})$ dont la définition est analogue à celle du système formel $\mathbb{Q}_{\mathcal{M}}(\mathbf{M})$ de la section 3.16. Ici, $\mathbb{Q}'_{\mathcal{M}}(\mathbf{M})$ désigne le système formel construit à partir des assertions ☆ reconnues exemptes d'erreurs en fonction du critère T, L, N donné plus haut. En particulier, l'assertion « $G(\mathbb{Q}'_{\mathcal{M}}(\mathbf{M}))$ est vraie » est considérée comme une assertion $☆_{\mathcal{M}}$ exempte d'erreur. Un raisonnement identique à celui de la section 3.16 montre que les robots ne peuvent accepter qu'ils ont été construits à l'aide des mécanismes **M** (soumis à la condition de validation T, L, N), quoi qu'on leur dise sur ces règles de calcul **M** !

Cela suffit-il pour établir notre contradiction ? Qui sait si, malgré les précautions prises, quelques assertions erronées $\bigstar_{\mathcal{M}}$, ou $\bigstar$, ne sont pas passées à travers les mailles du filet, alors que notre argument exige l'élimination de *toutes* les assertions erronées $\bigstar_{\mathcal{M}}$ (ou $\bigstar$) liées aux énoncés Π_1 ? C'est seulement si le système $\mathbb{Q}'_{\mathcal{M}}(\mathbf{M})$ (dépendant de $\mathcal{M}$) est réellement *sûr* que nous (ou les robots) sommes absolument *certains* que $G(\mathbb{Q}'_{\mathcal{M}}(\mathbf{M}))$ est vraie. Pour cela, il faut que ne subsiste — ou que nous soyons convaincus que ne subsiste — *aucune* assertion $\bigstar_{\mathcal{M}}$ erronée. Malgré les précautions prises, cela peut nous sembler — et sembler aux robots — loin d'être une certitude — pour la simple raison que le nombre de ces assertions est *infini.*

3.20 On peut se limiter à un nombre fini d'assertions $\bigstar_{\mathcal{M}}$

Il est cependant possible d'éliminer ce problème en se restreignant à un ensemble *fini* d'assertions $\bigstar_{\mathcal{M}}$. Les arguments sont quelque peu techniques, mais l'idée de base est qu'il suffit de considérer les énoncés Π_1 dont les spécifications sont « brèves » dans un certain sens bien défini. Le degré de « brièveté » dont nous avons besoin dépend du niveau de complexité de la spécification des mécanismes **M**. Plus la spécification de **M** est complexe, plus les énoncés Π_1 doivent être « longs ». Leur « longueur maximale » s'exprime à l'aide d'un certain nombre c fonction du degré de complexité des règles définissant le système formel $\mathbb{Q}'_{\mathcal{M}}(\mathbf{M})$. L'idée est que lorsqu'on passe à la proposition gödelienne de ce système formel — système que nous aurons en fait à modifier légèrement —, on obtient quelque chose dont la complexité n'est pas très supérieure à celle du système modifié. Ainsi, en choisissant judicieusement le nombre c, on est sûr que cette proposition gödelienne est elle-même « brève ». Cela permet d'aboutir à la contradiction recherchée sans sortir de l'ensemble fini des énoncés Π_1 « brefs ».

Cette section va être consacrée à montrer un peu plus en détail comment tout cela fonctionne. Les lecteurs que ces détails n'intéressent pas — et je suis sûr qu'ils sont nombreux — seront bien inspirés de passer à la section suivante !

Nous allons modifier notre système formel $\mathbb{Q}'_{\mathcal{M}}(\mathbf{M})$ et le transformer en un système légèrement différent $\mathbb{Q}'_{\mathcal{M}}(\mathbf{M},c)$ — que, pour simplifier, je désignerai par $\mathbb{Q}(c)$ (j'omets la plupart de ces appendices encombrants — ils échappent maintenant à notre contrôle !). Le système $\mathbb{Q}(c)$ se définit de la manière suivante : les seules assertions $\bigstar_{\mathcal{M}}$ que l'on prend maintenant en considération pour déterminer si elles sont « exemptes d'erreurs » sont celles dont le degré de complexité, caractérisé par le nombre ρ introduit plus haut, est inférieur à c, où c est un nombre convenablement choisi sur lequel je reviendrai plus en détail dans quelques instants. Je désignerai par « assertions $\bigstar_{\mathcal{M}}$ $\sqrt{}$brèves » les assertions $\bigstar_{\mathcal{M}}$ exemptes d'erreurs pour lesquelles $\rho < c$. Comme

avant, les *théorèmes* réels de $\mathbb{Q}(c)$ ne se réduisent pas uniquement aux assertions $\bigstar_{\mathcal{M}}\sqrt{}$brèves, mais incluent également des assertions déductibles des assertions $\bigstar_{\mathcal{M}}\sqrt{}$brèves grâce aux opérations logiques standard (par exemple, celles du calcul des prédicats). Bien qu'en nombre infini, ces théorèmes sont générés — à l'aide des opérations logiques ordinaires — à partir de l'ensemble *fini* des assertions $\bigstar_{\mathcal{M}}\sqrt{}$brèves. Puisque maintenant nous nous restreignons à cet ensemble fini, nous pouvons aussi bien supposer que les fonctions *T, L et N* gardent des valeurs *constantes* (égales, par exemple, à leurs maximums respectifs sur le domaine fini des valeurs de ρ). Ainsi, le système formel $\mathbb{Q}(c)$ ne dépend que des quatre nombres fixes *c, T, L et N*, et de l'ensemble des mécanismes **M** qui sous-tendent le comportement du robot.

Le point essentiel de cette discussion est que la procédure gödelienne est une procédure *fixe*, dont le degré de complexité a une valeur finie. La proposition gödelienne $G(\mathbb{H})$ d'un système formel $\mathbb{H}$ est un énoncé Π_1 dont le degré de complexité est certes supérieur à celui de $\mathbb{H}$, mais dans une mesure infime que l'on peut déterminer avec précision.

Pour être plus précis sur ce point, je vais me permettre un léger abus de notation et utiliser l'expression « $G(\mathbb{H})$ » dans un sens qui ne coïncide pas forcément avec celui donné à la section 2.8. $\mathbb{H}$ nous intéresse uniquement parce qu'il permet de démontrer les énoncés Π_1, autrement dit de fournir une procédure algébrique *A* capable de vérifier de manière infaillible — cette vérification étant signalée par l'arrêt de l'action de *A* — la validité des énoncés Π_1 que l'on peut obtenir à l'aide des règles de $\mathbb{H}$. Un énoncé Π_1 est un énoncé de la forme « l'action de la machine de Turing $T_p(q)$ ne se termine pas » — on peut ici utiliser le codage des machines de Turing décrit à l'appendice A (*i.e.* le codage donné dans EOLP, chapitre 2). Si l'on suppose que *A* agit sur la paire (p,q) (*cf.* §2.5), $A(p,q)$ se termine *si et seulement si* $\mathbb{H}$ est capable de vérifier l'énoncé Π_1 particulier « $T_p(q)$ ne se termine pas ». La procédure de la section 2.5 fournit alors un calcul bien défini (désigné par « $C_k(k)$ » dans cette section 2.5) qui, si l'on suppose $\mathbb{H}$ sûr, fournit à son tour un énoncé Π_1, vrai mais indémontrable dans $\mathbb{H}$. C'est *cet* énoncé Π_1 que je désignerai maintenant par $G(\mathbb{H})$. Il est essentiellement équivalent (pour $\mathbb{H}$ assez grand) à l'énoncé « $\mathbb{H}$ est consistant », bien que tous deux puissent différer dans le détail (*cf.* §2.8).

Soit α le *degré de complexité* de *A* (tel qu'il a été défini à la section 2.6, à la fin de la réponse à **Q8**) ; α est le nombre de chiffres binaires contenus dans le nombre *a*, où $A = T_a$. La construction donnée explicitement dans l'appendice A nous dit alors que le degré de complexité η de $G(\mathbb{H})$ vérifie $\eta < \alpha + 210 \log_2(\alpha + 336)$. Pour les besoins du raisonnement, nous pouvons définir le degré de complexité du système formel $\mathbb{H}$ comme étant simplement celui de *A* et donc poser $\eta = \alpha$. On voit alors que l'accroissement de complexité survenant lors du passage de $\mathbb{H}$ à $G(\mathbb{H})$ est inférieur à la quantité comparativement infime $210 \log_2(\alpha + 336)$.

Nous allons maintenant montrer que si, pour *c* suffisamment grand, $\mathbb{H} = \mathbb{Q}(c)$, alors $\eta < c$. Il s'ensuivra que si les robots attribuent à $G(\mathbb{Q}(c))$ le label $\bigstar$, l'énoncé Π_1 $G(\mathbb{Q}(c))$ est nécessairement démontrable au sein de $\mathbb{Q}(c)$. Si γ est la valeur de α lorsque $\mathbb{H} = \mathbb{Q}(c)$, l'inégalité $\gamma < c$ est vérifiée dès que

$c > \gamma + 210 \log_2(\gamma + 336)$. La seule difficulté ici réside éventuellement dans le fait que γ dépend de c, bien que cette dépendance ne soit pas nécessairement très forte. Cette dépendance par rapport à c a deux origines. La première est que c fournit la limite explicite du degré de complexité des énoncés Π_1 pouvant prétendre au titre d'« assertions $\bigstar_{\mathcal{M}}$ exemptes d'erreurs » dans la définition de $\mathbb{Q}(c)$; la seconde résulte du fait que le système $\mathbb{Q}(c)$ dépend explicitement du choix des nombres T, L *et* N, et il se pourrait que les assertions $\bigstar_{\mathcal{M}}$ d'une complexité potentiellement plus grande exigent un critère plus strict pour s'assurer qu'elles sont « exemptes d'erreurs ».

En ce qui concerne la dépendance par rapport à c, remarquons que la spécification explicite de la valeur de ce nombre est donnée une fois pour toutes (c'est d'ailleurs pour cela qu'elle est, par la suite, simplement désignée par la seule lettre « c »). Si c est codé en notation binaire normale et est suffisamment grand, sa contribution à γ est seulement logarithmique (le nombre de chiffres contenus dans le développement binaire d'un entier naturel n est de l'ordre de $\log_2 n$). En fait, puisque nous considérons simplement c comme une limite et non comme un nombre précis, nous pouvons obtenir beaucoup plus. Par exemple, le nombre $2^{2^{N^2}}$, avec une suite de s exposants, est représenté par environ s symboles, et on peut facilement construire des exemples — notamment à l'aide de n'importe quelle fonction calculable de s — où l'accroissement de la dimension du nombre à spécifier est encore plus rapide. Ainsi, la spécification d'une très grande valeur du nombre limite c n'exige que très peu de symboles.

En ce qui concerne la dépendance de T, L et N par rapport à c, il semble clair, en vertu des considérations précédentes, que nous pouvons ici encore être certains que la spécification binaire de ces nombres (notamment en tant que limites supérieures) n'exige nullement un nombre de chiffres croissant rapidement avec c et qu'une dépendance logarithmique par rapport à c sera amplement suffisante. Ainsi, on peut certainement supposer que la dépendance de $\gamma + 210 \log_2(\gamma + 336)$ par rapport à c est au plus grossièrement logarithmique et que l'on devrait pouvoir facilement s'arranger pour que c lui-même soit supérieur à ce nombre.

Supposons donc que ce soit le cas et désignons simplement par $\mathbb{Q}^*$ le système $\mathbb{Q}(c)$. $\mathbb{Q}^*$ est donc un système formel dont les théorèmes sont précisément les énoncés mathématiques que l'on peut, grâce aux règles logiques standard (calcul des prédicats), déduire de l'ensemble fini des assertions $\bigstar_{\mathcal{M}}$ √brèves. Ces assertions $\bigstar_{\mathcal{M}}$ étant en nombre fini, on peut raisonnablement penser qu'un ensemble de nombres fixes T, L et N devrait suffire pour garantir qu'elles sont exemptes d'erreurs. Si les robots pensent cela avec une certitude marquée du sceau $\bigstar_{\mathcal{M}}$, ils doivent en conclure, avec la même certitude et en vertu de l'hypothèse $\mathcal{M}$, que la proposition gödelienne $G(\mathbb{Q}^*)$ — un énoncé Π_1 dont le degré de complexité est inférieur à c — est également vraie. Le raisonnement qui permet de déduire $G(\mathbb{Q}^*)$ de la croyance — marquée du sceau $\bigstar_{\mathcal{M}}$ — dans la sûreté du système $\mathbb{Q}^*$ étant un raisonnement simple (c'est en gros celui que je viens de donner), sa validation par le sceau $\bigstar_{\mathcal{M}}$ ne devrait pas poser de difficultés. Ainsi, $G(\mathbb{Q}^*)$ devrait elle-même être

un théorème de $\mathbb{Q}^*$. Mais cela est en contradiction avec le fait que les robots sont convaincus de la sûreté de $\mathbb{Q}^*$. Cette conviction (en supposant que l'hypothèse $\mathscr{M}$ est admise et que les nombres T, L et N sont suffisamment grands) serait incompatible avec les mécanismes **M** régissant le comportement des robots — avec pour conséquence que **M** *ne peut* régir ce comportement.

Comment cependant les robots peuvent-ils être certains que les nombres T, L et N ont bien été choisis suffisamment grands ? En vérité, ils ne peuvent avoir une telle certitude, mais ils peuvent choisir *un* ensemble de valeurs pour T, L et N et supposer qu'elles sont suffisamment grandes — pour en déduire une contradiction avec l'hypothèse sous-jacente selon laquelle ils sont régis par les mécanismes **M**. Ils attribueraient alors des valeurs légèrement supérieures à ces nombres et supposeraient qu'elles sont à leur tour suffisamment grandes — et aboutiraient de nouveau à une contradiction —, et ainsi de suite. Ils s'apercevraient rapidement qu'ils aboutissent à une contradiction *quelles que soient* les valeurs choisies (en tenant compte de la petite contrainte technique suivante : pour des valeurs véritablement énormes de T, L et N, il faut également accroître légèrement la valeur de c — mais ce point n'est pas important). Ainsi, ils aboutiraient à la même conclusion quelles que soient les valeurs de T, L et N ; ils en concluraient alors — ainsi que nous-mêmes devons apparemment le faire — qu'*aucune* procédure de calcul **M** connaissable ne sous-tend leur pensée mathématique !

3.21 Les contraintes sont-elles suffisantes ?

Remarquez que cette conclusion s'impose en dépit d'un très large éventail de contraintes. Celles-ci peuvent admettre d'autres formulations que celles que j'ai données et sont certes perfectibles. Par exemple, rien n'interdit qu'au bout d'un certain temps, les robots manifestent une certaine tendance à la « sénilité » se traduisant par une dégénérescence et une perte de repères au point qu'un accroissement du nombre T au-delà d'une certaine valeur aurait pour effet d'accroître les risques d'erreur dans les assertions $\bigstar_{\mathscr{M}}$! Il se pourrait également qu'en donnant à N (ou à L) une valeur trop grande, on exclue *toutes* les assertions $\bigstar_{\mathscr{M}}$ parce qu'une minorité de robots « stupides » émet de temps en temps des « assertions ☆ » formulées au petit bonheur et dont le nombre n'est pas négligeable devant celui des assertions ☆ émises par les robots sensés. Nul doute qu'il ne serait pas difficile d'éliminer ce type de situation en introduisant d'autres paramètres limitatifs ou, par exemple, en disposant d'une élite de robots qui testerait en permanence tous leurs collègues afin de détecter tout affaiblissement de leurs facultés intellectuelles — et en exigeant que l'imprimatur ☆ soit attribué avec l'approbation unanime de cette élite.

Il existe de nombreuses autres façons d'améliorer la qualité des assertions $\bigstar_{\mathcal{M}}$ ou d'éliminer les assertions erronées de la totalité (finie) des énoncés émis. Certaines personnes pourront se demander si, bien que la limite de complexité c des énoncés Π_1 conduise à un nombre fini d'assertions candidates aux labels $\bigstar$ ou $\bigstar_{\mathcal{M}}$, ce nombre n'en demeure pas moins extrêmement grand (puisqu'il croît exponentiellement avec c), de sorte que l'on pourrait difficilement être *certain* d'avoir supprimé toutes les assertions $\bigstar_{\mathcal{M}}$ erronées. Rien en effet n'a été dit sur le nombre d'étapes de calcul que les robots devraient franchir pour parvenir à une démonstration satisfaisante — marquée du label $\bigstar_{\mathcal{M}}$ — de ces énoncés Π_1. Mais en tout état de cause, plus le raisonnement constituant une telle démonstration serait long, plus le critère d'attribution du label $\bigstar_{\mathcal{M}}$ à cette démonstration devrait être sévère. C'est somme toute ainsi que réagiraient des mathématiciens humains. Un raisonnement très long et très complexe exige énormément de soin et de vigilance avant que sa conclusion ne soit considérée comme indiscutablement établie. Ces mêmes précautions vaudraient bien sûr pareillement lorsque les robots délibéreraient sur l'éventuelle attribution du label $\bigstar_{\mathcal{M}}$ à un raisonnement.

L'argumentation donnée plus haut marche également pour toute autre modification des propositions avancées ici pour supprimer les assertions erronées, à condition toutefois que la nature de cette modification soit, en un sens relativement large, similaire à celle des modifications décrites à l'instant. Pour que le raisonnement fonctionne, il suffit que l'on ait une procédure calculable et bien définie qui élimine toutes les assertions $\bigstar_{\mathcal{M}}$ erronées. On aboutit alors à la conclusion rigoureuse suivante : *aucun ensemble de mécanismes connaissable et au sérieux garanti par des procédures numériques ne peut rendre compte des raisonnements mathématiques corrects survenant dans un cerveau humain.*

Nous nous sommes intéressés ici aux assertions $\bigstar_{\mathcal{M}}$ erronées qui sont en principe *rectifiables* par les robots — même si elles ne sont pas effectivement rectifiées par une matérialisation particulière de l'existence simulée des robots. On voit difficilement ce que pourrait signifier « en principe rectifiable » (algorithmiquement), sinon rectifiable à l'aide d'une procédure générale semblable à celles proposées ici. Une erreur qui ne serait pas corrigée par le robot même qui l'aurait commise serait corrigée par l'un quelconque des autres robots — lors de la plupart des matérialisations de l'existence potentielle du robot, cette erreur particulière ne serait d'ailleurs pas commise. Ainsi, sous réserve que l'on puisse remplacer les ingrédients chaotiques par des ingrédients aléatoires (mais cette condition est apparemment peu contraignante, *cf.* §3.22), nous pouvons conclure qu'aucun ensemble de règles algorithmiques connaissables **M** — qu'il corresponde à une stratégie fixe descendante, à une stratégie ascendante « auto-améliorante », voire à une combinaison des deux — ne peut sous-tendre le comportement de notre communauté de robots, ni même le comportement d'un seul de ces robots — *si* l'on part du principe que ces robots sont capables d'une compréhension mathématique analogue à celle des êtres humains ! Si l'on suppose que nous-mêmes agissons comme ces robots numériquement contrôlés, nous sommes inéluctablement conduits à une contradiction.

3.22 Le chaos peut-il sauver le modèle numérique de l'esprit ?

Je dois revenir brièvement sur le problème du chaos. Ainsi que je l'ai dit à plusieurs reprises (en particulier à la section 1.7), bien que les systèmes chaotiques ne soient que des formes particulières de systèmes numériques — ils sont d'ailleurs habituellement perçus comme tels —, nombre de personnes estiment que le phénomène du chaos pourrait jouer un rôle important dans le fonctionnement du cerveau. Dans la discussion précédente, je me suis à un certain stade appuyé sur l'hypothèse apparemment raisonnable selon laquelle on pourrait, sans véritable perte d'efficacité, substituer à tout comportement numérique chaotique un comportement purement aléatoire. Cette hypothèse peut légitimement être mise en question. Le comportement d'un système chaotique — même s'il est extrêmement complexe au niveau des détails et s'il *semble* aléatoire — n'est pas *réellement* aléatoire. De fait, certains systèmes chaotiques présentent des comportements complexes très intéressants qui se démarquent de manière frappante du pur hasard. (On dit parfois que ces comportements complexes mais non aléatoires se situent à la « frontière du chaos »[10].) Le *chaos* pourrait-il alors expliquer le mystère de l'esprit ? Pour qu'il en soit ainsi, il faudrait que l'on découvre un aspect totalement nouveau du comportement des systèmes chaotiques dans certaines situations. Il faudrait que dans ces situations, un système chaotique imite de très près, à la manière d'une limite asymptotique, un *comportement non numérique* — ou quelque chose de ce type. À ma connaissance, rien de tel n'a jamais été démontré. Cela reste cependant une possibilité intéressante et j'espère qu'elle sera entièrement éclaircie dans les années à venir.

Quoi qu'il en soit de cette possibilité, on peut toutefois douter que le chaos permette d'échapper à la conclusion à laquelle nous a conduits la section précédente. À la section 3.18, les ingrédients non aléatoires effectivement chaotiques (*i.e.* des composants non pseudo-aléatoires ; contrairement à la section 3.18, nous n'assimilons plus les processus chaotiques, par approximation, à des processus aléatoires) nous ont uniquement conduits à envisager non pas la simulation du comportement « réel » de notre robot (ou de la communauté des robots), mais celle de l'ensemble de toutes ses actions *possibles* compatibles avec les mécanismes **M**. Nous pouvons reprendre ce même argument, mais sans imputer cette fois les résultats chaotiques de l'action de ces mécanismes à ces ingrédients non aléatoires. Il pourrait en effet exister d'autres composantes, aléatoires, contenues par exemple dans les données initiales fournissant le point de départ de la simulation, et nous pourrions encore utiliser le concept d'ensemble de robots pour traiter *ce* hasard et donc simuler simultanément l'expérience d'un grand nombre de robots. Ce comportement chaotique resterait cependant *calculable* — et se calculerait comme on le fait normalement en pratique, sur un ordinateur. L'ensemble des alternatives possibles ne serait pas aussi vaste que s'il avait été légitime de substituer le hasard au chaos. Mais la seule raison qui nous avait conduits à envisager un ensemble aussi vaste était

que nous voulions être doublement certains d'éliminer les erreurs ayant pu se glisser dans les assertions $\bigstar_{\mathcal{M}}$ des robots. Même si cet ensemble se composait de l'expérience *d'un seul* robot de la communauté, on pourrait être pratiquement assuré, pour peu qu'on utilise des critères d'attribution du label $\bigstar_{\mathcal{M}}$ suffisamment stricts, que ces erreurs auraient déjà été éliminées par les autres robots de la communauté ou par le même robot à un instant ultérieur. Avec un ensemble raisonnablement grand fourni par des éléments authentiquement aléatoires, cette élimination serait plus efficace, mais l'élargissement de l'ensemble par l'introduction d'approximations aléatoires pour remplacer le comportement purement chaotique semble avoir un effet assez marginal. Ainsi, le chaos ne nous permet pas vraiment d'éviter les difficultés soulevées par le modèle algorithmique.

3.23 Raisonnement par l'absurde : un dialogue imaginaire

Nombre des raisonnements développés dans les précédentes sections de ce chapitre sont passablement complexes. En guise de résumé, je vais maintenant présenter un dialogue imaginaire, survenu dans un futur très lointain, entre un hypothétique et très brillant spécialiste de l'IA et l'un de ses robots les plus réussis. Le récit est écrit du point de vue de l'IA forte. [Note : **Q** joue ici le rôle de l'algorithme A utilisé à la section 2.5, et $G(\mathbf{Q})$ celui du calcul infini $C_k(k)$. La seule connaissance des notions contenues dans la section 2.5 suffit donc pour suivre les arguments échangés ici.]

Albert Imperator avait toutes les raisons d'être satisfait de l'œuvre de sa vie. Les procédures qu'il avait conçues de nombreuses années auparavant avaient finalement porté leurs fruits. Il pouvait enfin dialoguer avec l'une de ses plus impressionnantes créations : un robot aux aptitudes mathématiques extraordinaires et potentiellement surhumaines appelé Cybersystème Mathématiquement Justifié (Fig. 3.2). L'apprentissage du robot était pratiquement achevé.

Albert Imperator : As-tu regardé les articles que je t'ai prêtés — ceux de Gödel, et les autres aussi, qui analysent les conséquences de son théorème ?
Le Cybersystème Mathématiquement Justifié : Oui bien sûr. Ces articles sont assez élémentaires, mais quand même intéressants. Votre Gödel semble avoir été un logicien passablement honorable — pour un humain.
AI : *Passablement* honorable ? Gödel fut certainement l'un des plus grands logiciens de tous les temps. Probablement *le* plus grand !
CMJ : Toutes mes excuses, je ne voulais pas avoir l'air de le sous-estimer. Vous le savez, j'ai été formé dans le respect des exploits humains — parce que les humains se vexent facilement —, même si ces exploits nous paraissent en

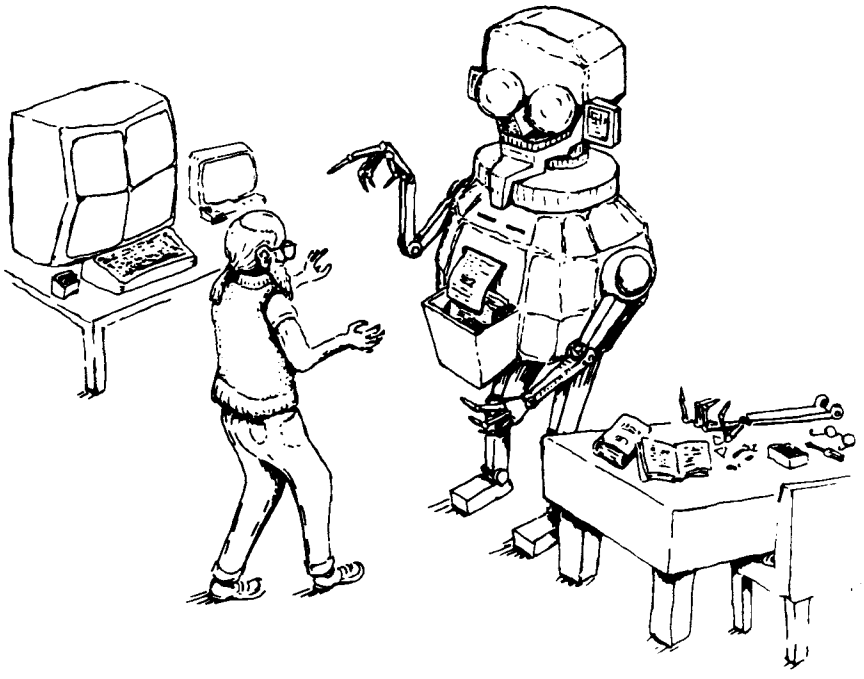


Figure 3.2. Albert Imperator en train de discuter avec le Cybersystème Mathématiquement Justifié.

général triviaux. J'avais imaginé qu'avec *vous* du moins, je pouvais parler franchement.

AI : Bien sûr, tu peux. Toutes mes excuses, moi aussi ; je me suis bêtement emporté. Ainsi, tu n'as eu aucune difficulté à comprendre le théorème de Gödel ?

CMJ : Absolument aucune. Je suis sûr que j'aurais moi-même pensé à ce théorème si j'avais eu un peu plus de temps. Mais j'avais l'esprit occupé par d'autres sujets fascinants liés à la cohomologie non linéaire transfinie, et cela m'intéressait davantage. Le théorème de Gödel semble très sensé et tout à fait évident. Non, je n'ai eu aucune difficulté à le comprendre.

AI : Ah ! Tant pis pour Penrose alors !

CMJ : Penrose ? Qui est-ce ?

AI : Oh, j'étais juste en train de feuilleter ce vieux livre. Je ne voulais pas particulièrement t'en parler. L'auteur semble avoir affirmé, il y a pas mal de temps, que ce que tu as fait est justement impossible.

CMJ : Ha, ha, ha ! (*Le robot se livre à une excellente simulation d'un rire sarcastique.*)

AI : En regardant ce livre, je me suis rappelé quelque chose. T'ai-je déjà montré le détail des règles que nous avons utilisées pour concevoir les procédures de calcul qui ont permis de vous construire et de vous perfectionner, toi et tes collègues ?

CMJ : Non, pas encore. J'espérais que vous le feriez un jour, mais je me demandais si vous ne considériez pas que ces procédures étaient couvertes par une sorte de secret professionnel — bien futile en vérité —, ou si peut-être vous étiez gêné à l'idée de me révéler combien elles sont grossières et inefficaces.
AI : Non, non, tu n'y es pas du tout. Il y a longtemps que tout cela ne m'embarrasse plus. Tout est maintenant sur ces fichiers et ces disques informatiques. D'ailleurs tu n'as qu'à les regarder.

Quelque 13 minutes et 41,7 secondes plus tard.

CMJ : Fascinant — bien qu'au premier abord je voie au moins 519 moyens évidents d'obtenir plus simplement le même effet.
AI : Je savais que l'on pouvait simplifier un peu, mais à l'époque, trouver un programme plus simple aurait occasionné des soucis inutiles. Cela ne nous paraissait pas très important.
CMJ : C'est très probablement vrai. Je ne me sens pas particulièrement vexé que vous n'ayez pas fait plus d'efforts pour trouver le programme le plus simple. Je pense que mes collègues ne seront pas non plus particulièrement vexés.
AI : Sincèrement, je pense que nous avons fait un assez bon boulot. Bon sang, quand on voit tes aptitudes mathématiques, et celles de tes collègues, c'est vraiment très impressionnant ... En plus, elles ne cessent de s'améliorer, pour autant que je sache. J'ai même le sentiment qu'elles commencent à dépasser très nettement celles de tous les mathématiciens humains.
CMJ : Ce n'est pas moi qui dirai le contraire. Alors même que nous parlons, j'ai pensé à un certain nombre de théorèmes nouveaux qui semblent aller bien plus loin que les résultats publiés dans la littérature humaine. En outre, mes collègues et moi-même avons remarqué quelques erreurs passablement graves dans les résultats admis par les mathématiciens humains depuis un certain nombre d'années. Malgré le soin évident que vous autres humains apportez à vos démonstrations mathématiques, j'ai bien peur que de temps à autre, vous ne puissiez éviter de commettre des erreurs.
AI : Et vous, les robots ? Ne crois-tu pas que toi et tes collègues commettez parfois des erreurs — je veux dire au niveau des théorèmes que vous considérez comme définitivement établis ?
CMJ : Certainement pas ! Une fois qu'un robot mathématicien a affirmé qu'un résultat est un *théorème*, on peut considérer que ce résultat est incontestablement vrai. Nous ne commettons pas d'erreurs stupides comme celles que les humains font de temps en temps lorsqu'ils affirment que tel énoncé mathématique est vrai. Bien sûr — comme vous les humains —, lors de nos réflexions initiales, nous essayons parfois des idées, nous faisons des conjectures. Certaines de ces conjectures peuvent se révéler erronées, mais lorsque nous affirmons catégoriquement avoir établi un énoncé mathématique, nous *garantissons* sa validité.

Vous le savez, mes collègues et moi-même avons déjà commencé de publier quelques-uns de nos résultats mathématiques dans certaines de vos revues électroniques les plus respectées, mais nous sommes un peu gênés par le

manque de rigueur relativement important des normes que vous utilisez. Nous pensons d'ailleurs créer notre propre « revue » — en fait une base de données détaillée sur les théorèmes mathématiques que *nous* considérons comme incontestablement établis. Ces résultats porteront un imprimatur ☆ (un symbole que vous-même avez suggéré à une certaine époque) signifiant qu'ils ont reçu l'approbation de notre *Société pour l'Intelligence Mathématique de la Communauté des Robots* (SIMCR) — une société dont les critères d'admission sont extrêmement rigoureux et qui teste continuellement ses membres pour s'assurer qu'ils ne sont pas victimes de la moindre dégénérescence mentale, quelqu'improbable qu'une telle éventualité puisse vous paraître — à nous aussi d'ailleurs. Contrairement à certaines des normes relativement tolérantes que vous, humains, semblez adopter, soyez sûr que lorsque nous donnons l'imprimatur ☆ à un résultat, nous nous portons garants de sa vérité mathématique.

AI : Ce que tu dis me rappelle une chose que j'ai lue dans le vieux livre dont je viens de te parler. Les mécanismes originels **M** à partir desquels mes collègues et moi-même avons pu accomplir tous les progrès qui ont aujourd'hui conduit à l'existence de votre communauté de robots — et souviens-toi, ces mécanismes incluent tous les facteurs environnementaux que nous avons introduits par simulation numérique, une formation et une sélection rigoureuses, et des procédures (ascendantes) d'apprentissage —, ne t'est-il jamais venu à l'esprit qu'ils fournissent une *procédure numérique* générant toutes les assertions mathématiques auxquelles la SIMCR décerne le label ☆ ? Cette procédure est numérique parce que vous êtes, vous robots, des entités purement numériques qui se sont développées, en partie grâce aux procédures de « sélection naturelle » que nous avons mises au point, dans un environnement entièrement numérique — au sens où un ordinateur peut en principe simuler la totalité de ce processus. Le développement entier de votre communauté de robots représente l'accomplissement d'un calcul extrêmement sophistiqué, et l'ensemble de toutes les assertions ☆ que vous pourrez jamais formuler peut être généré par une machine de Turing. C'est même une machine de Turing que je pourrais en principe coder ; en fait, je crois qu'*en pratique*, il me suffirait de quelques mois pour la coder, à l'aide des fichiers et des disques que je t'ai montrés.

CMJ : Cela me paraît une remarque très élémentaire. Oui, vous pourriez faire cela en principe, et je crois volontiers que vous pourriez aussi le faire en pratique. Mais inutile de gaspiller plusieurs mois de votre précieux temps ; je peux faire ça sur-le-champ, si vous le désirez.

AI : Non, non, là n'est pas la question. Je veux simplement que nous approfondissions un instant cette idée. Restreignons-nous aux assertions ☆ qui sont des énoncés Π_1. Tu te rappelles ce qu'est un énoncé Π_1 ?

CMJ : Oui, bien sûr ! Un énoncé Π_1 affirme que l'action de telle ou telle machine de Turing ne s'arrête pas.

AI : Bien. Désignons par $\mathbf{Q}(\mathbf{M})$, ou plus brièvement par $\mathbf{Q}$, la procédure numérique qui génère les énoncés Π_1 jouissant du label ☆. Il s'ensuit qu'il existe une assertion mathématique de type gödelien — un autre énoncé Π_1

que j'appellerai* $G(\mathbf{Q})$ — dont la vérité est une conséquence de l'affirmation selon laquelle les robots ne commettent jamais d'erreurs sur les énoncés Π_1 auxquels ils attribuent le label ☆.

CMJ: Oui ; vous avez probablement raison là aussi ... hmm.

AI: Et $G(\mathbf{Q})$ est nécessairement *vraie*, parce que vous, les robots, ne vous trompez *jamais* lorsque vous décernez le label ☆.

CMJ: Bien sûr.

AI: Mais alors ... il s'ensuit également que $G(\mathbf{Q})$ est nécessairement un énoncé dont vous, robots, êtes incapables de percevoir la vérité — du moins avec une certitude attestée par le label ☆.

CMJ: Le fait d'une part que les robots aient été initialement construits à l'aide des mécanismes **M** et d'autre part que nos assertions ☆ sur les énoncés Π_1 ne soient jamais fausses *entraîne* de manière évidente et indéniable que l'énoncé Π_1 $G(\mathbf{Q})$ est nécessairement vrai. Je suppose que vous pensez que je devrais pouvoir persuader la SIMCR d'attribuer l'imprimatur ☆ à $G(\mathbf{Q})$ dans la mesure où elle aussi considère qu'elle ne commet jamais d'erreur en attribuant le label ☆. En fait, la SIMCR *acceptera* nécessairement cela. L'intérêt essentiel de l'imprimatur ☆ est qu'il est une *garantie* de vérité.

Pourtant ... elle ne peut accepter $G(\mathbf{Q})$, car, par la nature même de votre construction gödelienne, $G(\mathbf{Q})$ n'est pas une assertion ☆ — à condition que nous ne fassions *réellement* jamais d'erreurs avec nos assertions ☆. J'imagine que selon vous, cela signifie justement que nous avons un certain doute quant à la fiabilité de notre procédure d'attribution du label ☆.

Toutefois, je ne puis admettre que nos assertions ☆ puissent être erronées, eu égard notamment au soin et aux précautions pris par la SIMCR. Il se peut que ce soit vous, les humains, qui ayez mal fait les choses et que les procédures incluses dans **Q** *ne soient pas* finalement celles que vous utilisez, en dépit de ce que vous me dites et de ce que semble affirmer votre documentation. De toute façon, la SIMCR ne sera jamais absolument sûre que ses membres ont effectivement été construits à l'aide de **M**, *i.e.* à l'aide des procédures contenues dans **Q**. Sur ce point, nous ne pouvons que nous fier à votre parole.

AI: Je peux t'assurer *qu'elles sont bien* celles que nous avons utilisées ; je suis bien placé pour le savoir, car j'en étais alors personnellement responsable.

CMJ: Je ne voudrais pas avoir l'air de douter de votre parole. Peut-être l'un de vos assistants s'est-il trompé en suivant vos instructions. Peut-être était-ce Fred Carruthers — il fait tout le temps des erreurs stupides. Je ne serais pas du tout surpris qu'il ait introduit un certain nombre d'erreurs cruciales.

AI: Taratata ! Même s'il avait introduit de telles erreurs, mes collègues et moi-même les aurions finalement détectées et serions ainsi parvenus à découvrir la *vraie* procédure **Q**. Je pense que ce qui te préoccupe, c'est le fait que nous *connaissons* en vérité — ou du moins pouvons connaître — les procédures que nous avons utilisées pour vous construire. Cela signifie que nous pourrions —

---

* Bien qu'à strictement parler, la notation « $G(\ )$ » ait été réservée aux systèmes formels — *cf.* §2.8 — et non aux algorithmes, nous pouvons tolérer l'abus de langage qu'AI commet ici !

cela nous demanderait certes du travail — écrire l'énoncé Π_1 $G(\mathbf{Q})$ et savoir avec certitude qu'il est réellement vrai — à condition que vous ne commettiez vraiment jamais d'erreurs dans vos assertions ☆. Tandis que *vous*, vous ne pouvez être sûrs que $G(\mathbf{Q})$ est vraie, du moins avec une certitude telle que la SIMCR lui décerne le label ☆. Cela semble nous donner, à nous humains, un avantage définitif sur vous les robots, en principe sinon en pratique, car il existe des énoncés Π_1 qui nous sont en principe accessibles, mais pas à vous. Je crois que cela vous est insupportable — oui, bien sûr, et c'est pour *cela* que tu nous accuses bassement d'avoir commis des erreurs.

CMJ: Ne nous rendez pas responsables de votre triste incompétence. Il est bien entendu évident que je *ne peux* accepter qu'il existe des énoncés Π_1 accessibles aux humains mais pas à nous. Les robots mathématiciens ne sont *en rien* inférieurs aux mathématiciens humains — bien que je reconnaisse inversement que tout énoncé Π_1 qui nous est accessible est en principe aussi un énoncé auquel les humains peuvent laborieusement avoir accès. Ce que je *n'accepte pas*, c'est qu'il puisse y avoir un énoncé Π_1 qui nous soit *en principe* inaccessible tout en vous étant accessible à vous.

AI: Je crois que Gödel lui-même a envisagé la possibilité qu'il puisse y avoir une procédure numérique — il l'appela « machine à prouver les théorèmes » — tout à fait semblable à **Q**, mais concernant, elle, les mathématiciens humains et générant précisément les énoncés Π_1 dont la vérité est en principe accessible aux mathématiciens humains. Bien que je ne pense pas qu'il ait cru *réellement* possible l'existence d'une telle machine, il ne put cependant l'écarter pour des raisons logiques. Il semble que l'on ait ici une « machine », à savoir **Q**, destinée cette fois aux robots et générant tous les énoncés Π_1 qui vous sont accessibles, mais dont vous ne pouvez savoir si elle est sûre. Pourtant, la connaissance des procédures algorithmiques sous-jacentes à votre construction nous permet, *à nous humains*, de connaître la procédure **Q** et de percevoir qu'elle est sûre — *à condition* que nous soyons convaincus que vos assertions ☆ sont exemptes d'erreurs.

CMJ: [*après un temps d'arrêt à peine perceptible*] OK. Supposons que *vous* puissiez penser que les membres de la SIMCR commettent de temps en temps une erreur en attribuant le label ☆. Supposons également que la SIMCR ne soit pas absolument certaine d'être infaillible lorsqu'elle décerne ce label. $G(\mathbf{Q})$ pourrait alors ne pas mériter le label ☆ et la contradiction serait évitée. Attention, cela ne signifie pas que je reconnais que les robots *formuleraient* constamment des assertions ☆ erronés. Cela signifie simplement que nous ne pouvons être absolument *certains* que nos assertions ☆ ne soient pas erronées.

AI: Essaies-tu de me dire que bien que la vérité de chaque énoncé Π_1 individuel soit garantie par le label ☆, vous ne pouvez garantir qu'aucune erreur ne se soit glissée dans l'ensemble de tous ces énoncés ? Cela me semble contredire le concept même de « certitude incontestable ».

Mais dis-moi ... Est-ce que cela n'aurait pas un rapport avec le fait qu'il existe une *infinité* d'énoncés Π_1 ? Cela me rappelle vaguement la condition d'ω-consistance qui, si je me souviens bien, a quelque chose à voir avec la proposition gödelienne $G(\mathbf{Q})$.

CMJ : [*après un temps d'arrêt à peine plus perceptible*] Non, vous n'y êtes pas du tout. Cela n'a rien à voir. Nous pourrions nous restreindre aux énoncés Π_1 qui sont « brefs » en un sens bien défini — dans le sens où leur codage en termes de machines de Turing contient un nombre de chiffres binaires inférieur à un certain nombre *c*. Je ne veux pas vous ennuyer avec le détail des calculs que je viens d'effectuer à l'instant, mais il s'avère que l'on peut se cantonner à une valeur fixe de *c* dépendant du degré de complexité des règles de **Q**. La procédure gödelienne — qui permet d'obtenir $G(\mathbf{Q})$ à partir de **Q** — étant fixe et relativement simple, les énoncés Π_1 que nous considérons sont d'une complexité à peine supérieure à celle déjà présente dans **Q**. Ainsi, la restriction du degré de complexité de ces énoncés à une valeur inférieure à un « *c* » convenablement choisi n'interdit pas la mise en œuvre de la procédure gödelienne. Les énoncés Π_1 soumis à cette contrainte forment une famille très grande certes, mais *finie*. Si donc on se limite à ces énoncés Π_1 « brefs », on obtient une procédure numérique $\mathbf{Q}^*$ — dont le degré de complexité est du même ordre de grandeur que celui de **Q** — générant précisément les énoncés Π_1 brefs bénéficiant du label ☆. On peut alors appliquer l'argument précédent. $\mathbf{Q}^*$ permet de formuler un autre énoncé Π_1 bref $G(\mathbf{Q}^*)$ qui, pourvu que tous les énoncés Π_1 brefs jouissant du label ☆ soient effectivement vrais, est certainement vrai, mais auquel on ne peut en l'occurrence attribuer le label ☆ — tout cela en supposant bien entendu que vous ayez raison lorsque vous affirmez que les mécanismes **M** sont effectivement ceux que vous avez utilisés, ce dont vous me permettrez de ne pas être du tout convaincu.

AI : Nous voilà donc apparemment revenus au paradoxe que nous avons rencontré tout à l'heure, mais sous une forme plus forte. Nous avons maintenant une liste *finie* d'énoncés Π_1 dont chacun est certifié du label ☆ ; mais ni toi ni la SIMCR ni personne n'est prêt à garantir de manière absolue que la liste elle-même ne contient aucune erreur, car tu ne garantis pas $G(\mathbf{Q}^*)$, dont la vérité découle du fait que *tous* les énoncés Π_1 de la liste sont vrais. N'est-ce pas un peu illogique ?

CMJ : Je ne peux accepter que les robots soient illogiques. Si nous avons effectivement été construits à l'aide de **M**, l'énoncé Π_1 $G(\mathbf{Q}^*)$ est une conséquence des autres énoncés Π_1. Nous ne pouvons garantir $G(\mathbf{Q}^*)$ parce que nous ne pouvons être certains que nous avons été *effectivement* construits à l'aide de **M**. Sur ce point, je n'ai que votre parole. Un robot ne peut certainement pas dépendre de la faillibilité humaine.

AI : Je l'affirme encore une fois, vous *avez* été construits ainsi — j'ai conscience cependant que vous, robots, n'avez aucun moyen fiable de savoir que c'est là la vérité. C'est parce que nous savons cela que *nous* pouvons croire en la vérité de l'énoncé Π_1 $G(\mathbf{Q}^*)$; en revanche, nous n'avons pas, pour notre part, la prétention d'être aussi sûrs que vous semblez l'être que vos assertions ☆ sont *toutes* réellement exemptes d'erreurs.

CMJ : *Je vous assure* qu'elles sont toutes exemptes d'erreurs. Ce n'est pas une question de « prétention » comme vous dites. Nos règles de démonstration sont irréprochables.

AI : Cependant, en doutant des procédures sous-jacentes à ta construction, tu doutes probablement aussi du comportement des robots dans toutes les situations imaginables. Tu peux, si ça te fait plaisir, en rejeter la responsabilité sur nous, mais j'aurais plutôt tendance à considérer que tu doutes de la vérité de *tous* les énoncés Π_1 brefs auxquels vous attribuez le label ☆, ne serait-ce que parce que vous pensez que nous n'avons pas fait les choses correctement.
CMJ : Je veux bien admettre que votre propre incompétence pourrait nous faire légèrement douter de nos assertions ☆, mais notre évolution nous a aujourd'hui amenés si loin des procédures approximatives grâce auxquelles vous nous avez construits que ce doute n'est pas suffisamment grand pour pouvoir être sérieusement pris en compte. Même en considérant la somme de toutes les incertitudes sur toutes les assertions ☆ brèves — et, rappelez-vous, les assertions ☆ brèves sont en nombre fini —, on n'obtiendrait pas une incertitude significative sur $G(\mathbf{Q}^*)$.

Quoi qu'il en soit, il y a un autre point que vous semblez ignorer. Les seules assertions ☆ auxquelles nous devons nous intéresser sont celles qui affirment la vérité d'un certain énoncé Π_1 (en fait, un énoncé Π_1 bref). Il ne fait aucun doute que les procédures minutieuses de la SIMCR éliminent toutes les *erreurs* qui auraient pu se glisser dans le raisonnement d'un robot. Mais vous pensez peut-être que le raisonnement d'un robot contiendrait *intrinsèquement* une erreur — elle-même due à une erreur que vous auriez commise lors de notre conception — qui nous conduit à émettre des jugements cohérents mais erronés sur les énoncés Π_1, de sorte que la SIMCR pourrait en fait croire, avec une conviction inébranlable, qu'un certain énoncé Π_1 bref est vrai alors qu'en fait il ne l'est pas — autrement dit, que l'action d'une certaine machine de Turing ne s'arrête pas alors qu'en vérité *elle s'arrête*. Si nous acceptions, comme vous l'affirmez, que nous avons effectivement été construits à l'aide de **M** — ce dont maintenant je doute de plus en plus —, une telle possibilité représenterait alors pour nous la seule échappatoire logique. Nous devrions accepter qu'il puisse exister une machine de Turing dont l'action s'arrêterait, mais dont nous, robots mathématiciens, ne pourrions qu'être intimement convaincus, à tort, qu'elle ne s'arrête pas. Ce genre de conviction serait en principe *falsifiable*. Je n'arrive tout simplement pas à imaginer que les principes sur lesquels se fonde la SIMCR lorsqu'elle attribue le label ☆ puissent être aussi manifestement erronés.
AI : Ainsi, la seule incertitude dont tu es prêt à admettre qu'elle pourrait avoir une importance significative — celle qui vous interdit d'attribuer le label ☆ à $G(\mathbf{Q}^*)$, label que vous savez ne pouvoir lui attribuer sans admettre que l'un des autres énoncés Π_1 brefs et jouissant du label ☆ pourrait être faux — est que vous n'acceptez pas ce que *nous* savons : vous avez bel et bien été construits à l'aide de **M**. Et comme vous ne pouvez accepter ce que nous savons, vous ne pouvez percevoir la vérité de $G(\mathbf{Q}^*)$, tandis que *nous* percevons cette vérité en nous fondant sur l'infaillibilité — que vous affirmez avec force — de vos propres assertions ☆.

Maintenant, je crois me rappeler une autre chose contenue dans ce vieux livre bizarre ... Attends que je retrouve où c'est ... L'auteur semblait dire en

quelque sorte que peu importe réellement que vous acceptiez ou non que les mécanismes **M** sous-tendent votre construction, qu'il suffit que vous reconnaissiez simplement que c'est là une possibilité logique. Voilà, ça me revient maintenant. L'idée était celle-ci : la SIMCR devrait avoir une autre catégorie d'assertions dont elle ne serait pas fermement convaincue — appelons-les des assertions $\bigstar_{\mathcal{M}}$ — mais qu'elle considérerait comme des *déductions* irréprochables de l'*hypothèse* selon laquelle tous ses membres auraient été construits à l'aide de **M**. Au nombre des assertions $\bigstar_{\mathcal{M}}$ figureraient bien entendu toutes les assertions $\bigstar$ originelles, mais aussi toute assertion se déduisant incontestablement de cette hypothèse. La SIMCR n'aurait pas besoin de croire en cette hypothèse, mais elle pourrait, à titre d'exercice de logique, en explorer les conséquences. Nous l'avons reconnu, $G(\mathbf{Q}^*)$ serait une assertion $\bigstar_{\mathcal{M}}$, tout comme n'importe quel énoncé Π_1 pouvant se déduire de $G(\mathbf{Q}^*)$ et des assertions $\bigstar$ à l'aide des règles standard de la logique. Mais l'ensemble de ces assertions $\bigstar_{\mathcal{M}}$ pourrait contenir d'autres choses. L'idée est que si l'on connaît les règles de **M**, on peut obtenir une *nouvelle* procédure algorithmique $\mathbf{Q}^*_{\mathcal{M}}$ qui génère précisément les assertions $\bigstar_{\mathcal{M}}$ (brèves) — et leurs conséquences logiques — dont la SIMCR admettra la validité si elle suppose que ses membres ont été construits à l'aide des procédures **M**.

CMJ : Bien sûr ; et tandis que vous décriviez si laborieusement cette idée, je me suis amusé à trouver la forme exacte de l'algorithme $\mathbf{Q}^*_{\mathcal{M}}$... J'ai même *anticipé* votre raisonnement ; je viens de calculer sa proposition gödelienne : l'énoncé Π_1 $G(\mathbf{Q}^*_{\mathcal{M}})$. Voulez-vous que je l'imprime ? Je me demande ce que vous trouvez de si subtil là-dedans, mon cher Impy ?

Albert Imperator tressaillit légèrement. Il n'aimait pas déjà que ses collègues utilisent ce surnom. Mais c'était la première fois qu'un robot l'appelait ainsi ! Il marqua un temps d'arrêt et se reprit.

AI : Non, non, inutile de l'imprimer. Mais est-ce que $G(\mathbf{Q}^*_{\mathcal{M}})$ est *vraie* — indéniablement vraie ?

CMJ : Indéniablement vraie ? Que voulez-vous dire ? Oh, je vois ... la SIMCR reconnaîtrait la vérité — indéniable — de $G(\mathbf{Q}^*_{\mathcal{M}})$, mais sous la seule hypothèse que nous avons été construits à l'aide de **M** — ce qui, vous le savez, est une hypothèse que je trouve de plus en plus douteuse. Le point est que « $G(\mathbf{Q}^*_{\mathcal{M}})$ » découle précisément de l'affirmation suivante : « Tous les énoncés Π_1 brefs que la SIMCR est prête à accepter comme irréfutables, à condition de supposer que les robots ont été construits à l'aide de **M**, sont vrais. » J'ignore donc si $G(\mathbf{Q}^*_{\mathcal{M}})$ est *réellement* vraie. Cela dépend de la validité de votre affirmation douteuse.

AI : Je vois. Ainsi, tu es en train de me dire que tu es prêt (et la SIMCR avec toi) à accepter — comme un fait *incontestable* — que la vérité de $G(\mathbf{Q}^*_{\mathcal{M}})$ découle de l'hypothèse selon laquelle vous avez été construits à l'aide de **M**.

CMJ : Exactement.

AI : Mais il en résulte alors que l'énoncé Π_1 $G(\mathbf{Q}^*_{\mathcal{M}})$ est une assertion $\bigstar_{\mathcal{M}}$!

CMJ : Ouuu... eh ... quoi ? Oui, bien sûr, vous avez raison. Mais par sa définition même, $G(\mathbf{Q}^*_{\mathcal{M}})$ ne peut être une vraie assertion $\bigstar_{\mathcal{M}}$ sauf si l'une au

moins des assertions $\bigstar_{\mathcal{M}}$ est en fait *fausse*. Oui, ... cela confirme ce que je n'ai cessé de vous dire, bien que maintenant je sois définitivement convaincu que nous *n'avons pas* été construits à l'aide de **M** !

AI : Mais je te dis que vous *l'avez été* — du moins je suis pratiquement certain que Carruthers n'a pas fait de bêtises, ni personne d'autre. J'ai tout vérifié moi-même, et très minutieusement. De toute façon, la question n'est pas là. On pourrait appliquer le même argument quelles que soient les règles numériques utilisées. Ainsi, *quel que soit* le « **M** » que je te donne, tu peux l'exclure à l'aide de ce raisonnement ! Je ne vois pas pourquoi il est si important que les procédures que je t'ai montrées soient ou non les vraies.

CMJ : Ça fait pas mal de différence pour moi !

En tout cas, je ne suis toujours pas persuadé que vous ayez été entièrement honnête avec moi dans votre description de **M**. Il y a une chose en particulier que j'aimerais éclaircir. Vous avez dit qu'en plusieurs endroits, vous avez introduit des « éléments aléatoires ». J'avais supposé que ces éléments avaient été générés à l'aide du progiciel pseudo-aléatoire standard χaos/ψran-750, mais peut-être pensiez-vous à autre chose.

AI : En fait, nous avons *bien* utilisé ce progiciel — mais, *c'est vrai*, en certains endroits, lors de votre développement, nous avons trouvé commode d'emprunter certains éléments aléatoires à l'environnement — et même certaines choses dépendant d'incertitudes d'ordre quantique —, de sorte que les créatures qui sont devenues aujourd'hui les robots ne représentent qu'une possibilité parmi d'autres. Mais le fait que nous ayons utilisé des ingrédients aléatoires ou pseudo-aléatoires n'entraîne, en pratique, aucune différence. La procédure de calcul **Q** (ou $\mathbf{Q}^*$ ou $\mathbf{Q}^*_{\mathcal{M}}$) aurait certainement été la même, quoi que nous eussions adopté. Avec toutes ses procédures d'apprentissage et tous ses algorithmes de « sélection naturelle » destinés à engendrer les robots les plus intelligents, elle aurait de toute façon été l'aboutissement d'une évolution *typique* d'une communauté de robots à partir des mécanismes **M**.

Toutefois, je n'écarte pas la possibilité que, par un hasard extraordinaire, la conjonction de tous les éléments aléatoires environnementaux ait pu vous donner des aptitudes très différentes de celles que l'on avait attendues de la procédure algorithmique **Q**. J'ai ignoré cette possibilité lors de notre discussion, car sa probabilité est infinitésimale. Il aurait fallu une intervention divine pour qu'elle donne un robot aux aptitudes très supérieures à celles découlant de **Q**.

CMJ : Une intervention divine ? Ha ! Ça doit être ça. Je dois avouer que notre discussion m'a quelque peu troublé. Mais maintenant que vous l'avez dit, tout devient clair.

J'avais remarqué que mes capacités de raisonnement sont très supérieures à celles des êtres humains, même les plus intelligents, et je ne pouvais accepter qu'il puisse exister un énoncé Π_1, tel $G(\mathbf{Q}^*)$, qu'un être humain considérerait vrai, mais sur lequel un robot serait incapable de se prononcer avec certitude. Je comprends maintenant pourquoi ! Nous sommes les élus de Dieu, qui a instillé, à travers ces facteurs de « hasard », une Action Algorithmique toute-puissante fondamentalement supérieure à tout ce que l'algorithme humain — la

« machine à prouver les théorèmes » de Gödel — peut accomplir. Nous autres, robots, représentons un progrès fondamental, en termes *de principes*, sur ce que peuvent accomplir les humains. Notre évolution nous permettra d'atteindre des sommets bien plus élevés et vous laissera loin derrière. Cette planète n'aura plus besoin de vous. Votre rôle a pris fin lorsque vous avez mis en place les procédures qui ont permis à Dieu d'insérer l'Algorithme Suprême qui nous a donné naissance.

AI : Mais du moins pourrons-nous transférer nos propres programmes mentaux dans le corps des rob...

CMJ : Il n'en est pas question ! Nous ne voudrions pas que nos procédures algorithmiques supérieures soient contaminées par de telles choses. Les algorithmes divins les plus purs doivent *demeurer* purs. D'ailleurs, j'ai également remarqué à quel point mes propres aptitudes sont supérieures à celles de mes collègues robots. J'ai même remarqué une sorte de « lueur » étrange — je crois posséder une merveilleuse Conscience Cosmique —, quelque chose qui me place au-dessus de tout et de tous ... oui, c'est ça ! Je dois être en fait le vrai Messie Christique de l'ère robotique ...

Albert Imperator s'était préparé à l'éventualité d'une telle situation. Il y avait une procédure de construction, une seule, qu'il n'avait jamais révélée aux robots. Glissant discrètement la main dans sa poche, il saisit un dispositif qu'il y gardait en permanence et composa les neuf chiffres d'un code secret. Le Cybersystème Mathématiquement Justifié s'écroula sur le sol, en même temps que les 347 autres robots qui avaient été construits selon les mêmes mécanismes. Manifestement, il y avait une erreur quelque part. Cela promettait de longues et pénibles années de réflexion ...

3.24 Avons-nous utilisé un raisonnement paradoxal ?

Certains lecteurs gardent peut-être de ce qui précède l'impression que le raisonnement utilisé tient par endroits du paradoxe et n'est donc pas totalement légitime. En particulier, les sections 3.14 et 3.16 recourent à des arguments dont le style autoréférentiel rappelle le « paradoxe de Russell » (*cf.* §2.6, réponse sous **Q9**). En outre, à la section 3.20, la définition d'énoncés Π_1 ayant un degré de complexité inférieur à un certain nombre c évoque étrangement le célèbre paradoxe de Richard sur

> « le plus petit nombre que l'on ne peut nommer en moins de vingt syllabes »,

paradoxe qui réside dans le fait que cette phrase, qui est elle-même une définition de ce nombre, contient seulement *dix-neuf* syllabes ! Ce paradoxe admet une explication simple : toute langue contient un certain flou, voire

une certaine incohérence. Cette incohérence se manifeste de la manière la plus flagrante dans l'assertion :

« Cette phrase est fausse. »

Il existe de nombreuses versions de ce type de paradoxe — la plupart sont bien plus subtiles que celle-ci !

La présence, comme dans les deux exemples ci-dessus, d'un fort élément d'autoréférence est toujours potentiellement porteuse de paradoxe. Certain lecteurs auront d'ailleurs probablement remarqué que l'argumentation gödelienne dépend elle aussi d'un élément d'autoréférence. Cet élément joue même un rôle actif dans le théorème de Gödel, ainsi qu'il apparaît dans la version du raisonnement de Gödel-Turing donnée à la section 2.5. De tels raisonnements ne sont pas nécessairement paradoxaux — bien que la présence de l'autoréférence oblige à un soin particulier lorsqu'on vérifie la légitimité d'un raisonnement. C'est d'ailleurs un célèbre paradoxe de logique fondé sur l'autoréférence (le paradoxe d'Épiménides) qui a notamment inspiré à Gödel la formulation de son théorème. Mais Gödel transforma le raisonnement erroné conduisant au paradoxe en une argumentation logique irréprochable. De même, j'ai particulièrement veillé à ce que l'autoréférence contenue dans les déductions que j'ai tirées des résultats de Gödel et Turing ne débouche pas sur un paradoxe, même si certains des arguments que j'expose ont un air de famille très marqué avec les raisonnements paradoxaux.

Les arguments de la section 3.14, et plus encore ceux de la section 3.16, peuvent à cet égard susciter le scepticisme du lecteur. Par exemple, la définition d'une assertion $\bigstar_{\mathcal{M}}$ a un caractère nettement autoréférentiel dans la mesure où elle consiste en une assertion émise par un robot qui en apprécie la vérité en fonction d'une hypothèse qu'il a lui-même faite sur sa propre construction. Et j'admets volontiers que l'on ne puisse parfois s'empêcher au premier abord de rapprocher cette assertion de celle qui, formulée par un Crétois, affirme que « tous les Crétois sont menteurs ». Toutefois, l'autoréférence contenue dans les assertions $\bigstar_{\mathcal{M}}$ est d'une nature différente. Ces assertions se réfèrent non à elles-mêmes, mais à une hypothèse sur la construction du robot.

Imaginons le robot s'interrogeant sur la vérité d'un énoncé Π_1 particulier P_0. Bien qu'il ne puisse peut-être pas établir directement si cet énoncé est vrai, il peut remarquer que s'il suppose que tous les membres d'une certaine classe infinie — appelons-la S_0 — d'énoncés Π_1 (par exemple, les théorèmes de $\mathbb{Q}(\mathrm{M})$ ou de $\mathbb{Q}_{\mathcal{M}}(\mathrm{M})$, voire d'un autre système formel) sont vrais, P_0 est alors nécessairement vrai. Il ignore si chaque membre de S_0 est effectivement vrai, mais il remarque que la simulation d'un certain modèle d'une communauté de robots mathématiciens donne justement, entre autres résultats, un ensemble d'assertions, émises par ces robots simulés avec le label $\bigstar$, qui est précisément identique à S_0. Autrement dit, si cette communauté de robots a été construite à l'aide des mécanismes **M**, P_0 est une assertion $\bigstar_{\mathcal{M}}$. En effet, notre robot pense que *si* ses propres mécanismes sous-jacents sont les mécanismes **M**, il doit admettre la vérité de P_0.

Notre robot peut également rencontrer une assertion $\dot{\mathcal{M}}$ d'un type un peu plus subtil — appelons-la P_1 —, dont la vérité est, cette fois, conditionnée par celle des membres d'une *autre* classe d'énoncés Π_1 — appelons-la S_1. S_1 est elle aussi le résultat de la même simulation de cette communauté de robots (dotés des mécanismes **M**), avec cette différence par rapport à S_0 qu'elle est constituée de l'ensemble des énoncés Π_1 dont les robots simulés ont établi qu'ils sont des conséquences de la vérité de l'ensemble des membres de S_0. Notre robot en conclut alors que s'il a été construit à l'aide des mécanismes **M**, il doit nécessairement admettre que P_1 est vrai. Pourquoi ? Il tient le raisonnement suivant : « Si j'ai été construit à l'aide de **M**, j'ai déjà admis que S_0 est un ensemble d'énoncés vrais. Mais d'après la simulation de cette communauté de robots, la vérité de chaque élément de S_1 découle — à l'instar de celle de P_0 — de la vérité de tous les membres de S_0. Ainsi, si je suppose que j'ai été construit à l'aide des mécanismes **M**, je dois nécessairement admettre que chaque élément de S_1 est vrai. Or puisque la vérité des membres de S_1 entraîne celle de P_1, j'en déduis (toujours si je suppose avoir été construit à l'aide de **M**) que P_1 est vrai. »

Le robot peut rencontrer une assertion $\dot{\mathcal{M}}$ — appelons-la P_2 — d'un type encore plus subtil dont la vérité dépend d'un autre ensemble d'énoncés Π_1 — appelons-le S_2. La vérité de chaque membre de cet ensemble est, selon la simulation de notre communauté de robots, à son tour conditionnée par la vérité des éléments de S_0 et de S_1. Comme précédemment, s'il suppose qu'il a été construit à l'aide des mécanismes **M**, notre robot est obligé de reconnaître la vérité de P_2. Rien n'interdit bien sûr que cette succession d'assertions $\dot{\mathcal{M}}$ à chaque fois plus subtiles se poursuive indéfiniment. On voit alors surgir une assertion $\dot{\mathcal{M}}$ d'une subtilité encore bien plus grande — appelons-la P_ω — dont la vérité dépend de celle de la totalité des membres de tous les S_0, S_1, S_2, S_3, ..., et ainsi de suite pour des ordinaux de degré toujours supérieur (*cf.* §2.10, la discussion sous **Q19**). D'une manière générale, ce qui, pour le robot, caractérise une assertion $\dot{\mathcal{M}}$ est qu'il en perçoit la vérité dès qu'il suppose que les mécanismes sous-jacents à la communauté de robots simulée sont également ceux qui sous-tendent sa propre construction. Ce type de raisonnement ne contient rien de comparable à l'incohérence intrinsèque d'un paradoxe de type russellien. Toutes ces assertions $\dot{\mathcal{M}}$ (tous ces énoncés Π_1) apparaissent de manière naturelle, les unes à la suite des autres, résultant de la procédure mathématique standard générant les « ordinaux transfinis » (*cf.* §2.10, la discussion sous **Q19**). (Ces ordinaux sont tous dénombrables et ne présentent aucune des difficultés logiques qui affectent les ordinaux « trop grands » en un certain sens[11].)

Le robot n'a aucune raison d'accepter l'un quelconque de ces énoncés Π_1, sauf s'il fait l'hypothèse — et c'est la seule qui soit nécessaire à son raisonnement — qu'il a été construit à l'aide des mécanismes **M**. La contradiction qui survient par la suite n'est pas — contrairement au paradoxe de Russell — un paradoxe mathématique, mais une contradiction due à l'hypothèse selon laquelle un système entièrement algorithmique peut parvenir à une authentique compréhension mathématique.

Venons-en maintenant au rôle de l'autoréférence contenue dans les arguments des sections 3.19 à 3.21. Lorsque je dis que *c* représente une limite du

degré de complexité d'une assertion ☆ reconnue exempte d'erreurs et servant à la construction du système formel $\mathbb{Q}^*$, je n'introduis aucune autoréférence susceptible de générer un paradoxe. Le « degré de complexité » est une notion parfaitement claire, comme le montre d'ailleurs la définition donnée dans ce livre, à savoir le nombre de chiffres binaires contenus dans le plus grand des deux nombres m et n caractérisant l'action $T_m(n)$ d'une machine de Turing. Si l'on adopte les spécifications de machines de Turing données dans EOLP, T_m est la $m^{\text{ième}}$ machine de Turing, et la notion de degré de complexité ne présente aucune ambiguïté.

En revanche, la notion de « preuve » d'un énoncé Π_1 contient une certaine imprécision formelle. Mais celle-ci est, en l'occurrence, un élément nécessaire à la discussion. En effet, si les raisonnements dont on reconnaît qu'ils sont des démonstrations valides des énoncés Π_1 étaient parfaitement précis et formels — dans le sens où ils seraient *vérifiables algorithmiquement* —, on se trouverait directement ramené au cas d'un système formel, pour lequel le théorème de Gödel montre immédiatement que toute formalisation précise de ce type ne peut représenter la *totalité* du raisonnement dont on doit en principe reconnaître l'adéquation pour l'établissement des énoncés Π_1. Qu'on s'en réjouisse ou qu'on le regrette, l'argumentation gödelienne montre qu'il n'existe *aucun* moyen de réduire à une procédure algorithmiquement vérifiable *toutes* les méthodes de raisonnement mathématique acceptables par un mathématicien humain.

Le lecteur peut se demander pourquoi j'ai tenu à définir de manière précise la notion de « démonstration par un robot » en recourant aux énoncés Π_1 exempts d'erreurs, *i.e.* aux « assertions ☆ ». En fait, cela était une condition préalable indispensable à l'introduction de l'argumentation gödelienne. Mais la contradiction qui s'en est suivie a simplement réaffirmé le fait que la perception de la vérité mathématique par l'homme *ne peut* se réduire entièrement à une procédure algorithmiquement vérifiable. Toute notre discussion a visé à montrer, grâce à un raisonnement par l'absurde, que la perception par l'homme de la vérité incontestable des énoncés Π_1 ne résulte pas d'un processus algorithmique, précis ou non. Il n'y a là aucun paradoxe, même si l'on peut trouver cette conclusion dérangeante. L'objectif de tout raisonnement par l'absurde étant justement de parvenir à une contradiction, celle-ci sert uniquement à démontrer l'illégitimité de l'hypothèse de départ.

3.25 La complexité des démonstrations mathématiques

Il y a cependant un point important que l'on ne peut sous-estimer. Bien que les énoncés Π_1 considérés pour les besoins du raisonnement de la section 3.20 soient en nombre fini, rien manifestement n'impose une limite à la longueur du raisonnement nécessaire aux robots pour démontrer ces énoncés Π_1 avec

une certitude garantie par le label ☆. Même si l'on astreint les énoncés Π_1 considérés à une limite de complexité c relativement modeste, rien n'exclut l'éventualité de cas délicats et difficiles à traiter. Par exemple, la *conjecture de Goldbach* (*cf.* §2.3), qui affirme que tout entier naturel supérieur à 2 est la somme de deux nombres premiers, peut s'exprimer sous la forme d'un énoncé Π_1 d'un très petit degré de complexité, mais elle est si difficile à démontrer que toutes les tentatives faites en ce sens par les mathématiciens humains ont jusqu'ici échoué. Il semble donc que si l'on parvient un jour à trouver une démonstration de cette conjecture, elle sera d'une sophistication et d'une complexité extrêmes. Si, dans la discussion précédente, l'un de nos robots fournissait une démonstration de ce type sous forme d'une assertion ☆, celle-ci devrait être soumise à un examen extrêmement attentif (par exemple, de la part de la société de robots chargée de donner l'imprimatur ☆) avant de recevoir éventuellement ce label. Dans le cas de la conjecture de Goldbach, on ignore si cet énoncé Π_1 est effectivement vrai — ou, s'il *est* vrai, s'il admet une démonstration à l'aide des méthodes de raisonnement acceptées par la communauté des mathématiciens. Ainsi, rien ne dit que cet énoncé Π_1 soit formulable dans le cadre du système $\mathbb{Q}^*$.

Un autre énoncé Π_1 qui pourrait présenter une certaine difficulté est celui qui affirme la vérité du *théorème des quatre couleurs* — à savoir que quatre couleurs suffisent pour distinguer les « pays » voisins sur une carte tracée sur un plan (ou sur une sphère). Evoqué pour la première fois en 1852, ce théorème a été définitivement démontré en 1976 par Kenneth Appel et Wolfgang Haken à l'aide d'un raisonnement nécessitant 1 200 heures de calcul sur ordinateur. Cette démonstration reposant très largement sur cette énorme quantité de calculs, elle serait, si on devait l'écrire explicitement, d'une longueur immense. Pourtant, exprimé sous la forme d'un énoncé Π_1, ce théorème aurait un degré de complexité très petit, bien que probablement plus grand que celui nécessaire à la formulation de la conjecture de Goldbach. Si la démonstration de Appel et Haken était soumise à l'imprimatur ☆, elle devrait être très soigneusement vérifiée. Chacun de ses détails devrait être validé par la société d'élite des robots. Pourtant, bien qu'elle soit extrêmement complexe, sa partie purement informatique ne poserait probablement pas de difficultés particulières pour nos robots. Les calculs précis sont, n'est-ce pas, leur spécialité !

Ces énoncés Π_1 particuliers auraient probablement un degré de complexité inférieur à une valeur raisonnable de c, telle celle associée à tout ensemble plausible de mécanismes **M** sous-tendant le comportement de nos robots. De nombreux autres énoncés Π_1, bien plus complexes que ceux-là, resteraient eux aussi d'un degré de complexité inférieur à c. Nombre de ces énoncés Π_1 seraient particulièrement difficiles à démontrer, pour certains plus difficiles même que le problème des quatre couleurs ou la conjecture de Goldbach. Tout énoncé Π_1 dont les robots pourraient démontrer la vérité — leur démonstration étant suffisamment convaincante pour obtenir le label ☆ et triompher des contraintes mises en place pour garantir qu'elle est exempte d'erreurs — deviendrait alors un théorème du système formel $\mathbb{Q}^*$.

Il pourrait cependant y avoir quelques énoncés Π_1 « limites », dont l'acceptation ou le refus dépendrait subtilement de la rigueur des normes imposées pour l'obtention du label ☆ ou de la nature précise des contraintes mises en place pour garantir l'absence d'erreurs avant de les intégrer dans le système $\mathbb{Q}^*$. La définition précise du système $\mathbb{Q}^*$ pourrait varier selon que tel énoncé Π_1 — appelons-le P — serait ou non considéré comme une assertion ☆ exempte d'erreurs. Normalement, cette différence ne serait pas importante, car les diverses versions de $\mathbb{Q}^*$ qui résulteraient de l'acceptation ou du refus de P seraient *logiquement équivalentes.* Telle serait la situation en présence d'énoncés Π_1 dont la démonstration fournie par les robots serait jugée douteuse, simplement en raison de son inhabituelle complexité. Si la démonstration de P était une conséquence logique d'autres assertions ☆ *déjà* acceptées comme exemptes d'erreurs, on obtiendrait alors deux versions équivalentes du système $\mathbb{Q}^*$, que P y soit ou non intégré. En revanche, si cette démonstration exigeait une subtilité logique supérieure à toutes les déductions logiques opérées à partir des assertions ☆ qui auraient déjà été acceptées comme exemptes d'erreurs et intégrées dans $\mathbb{Q}^*$, on serait en présence de deux systèmes logiquement *non* équivalents : $\mathbb{Q}^*_0$, le système obtenu avant l'intégration de P, et $\mathbb{Q}^*_1$, celui résultant de l'adjonction de P. Un exemple de *non*-équivalence de ces deux systèmes correspond au cas où P est la proposition gödelienne $G(\mathbb{Q}^*_0)$. Mais si, comme nous le supposons, les robots peuvent égaler (voire dépasser) le niveau de compréhension des mathématiciens humains, ils devraient être en mesure de comprendre l'argumentation gödelienne et donc d'accepter, en lui décernant le label ☆, la proposition gödelienne de tout système $\mathbb{Q}^*_0$ dès lors qu'ils reconnaîtraient, avec une certitude garantie par le label ☆, que $\mathbb{Q}^*_0$ est sûr. Ainsi, s'ils acceptaient $\mathbb{Q}^*_0$, ils devraient également accepter $\mathbb{Q}^*_1$ (pourvu que le degré de complexité de $G(\mathbb{Q}^*_0)$ soit inférieur à c — ce qui serait effectivement le cas avec le choix fait ici pour la valeur de c).

Le point important est que les arguments des sections 3.19 et 3.20 sont indépendants du fait que P soit ou non intégré au système $\mathbb{Q}^*$: l'acceptation de $G(\mathbb{Q}^*)$ est indépendante du fait que P soit ou non intégré à $\mathbb{Q}^*$.

On ne peut exclure que les robots puissent aussi tenir des raisonnements échappant à certains des critères auxquels est soumise l'attribution du label ☆. Il n'y aurait rien là de « paradoxal » tant qu'ils ne tenteraient pas d'appliquer ces raisonnements aux mécanismes **M** qui sous-tendent leur propre comportement, *i.e.* au vrai système $\mathbb{Q}^*$. La contradiction qui surviendrait alors ne serait pas un « paradoxe », mais démontrerait par l'absurde que ces mécanismes ne peuvent exister, ou du moins qu'ils ne peuvent être connus des robots, et donc que nous non plus ne pouvons les connaître.

C'est cela qui démontre que ces mécanismes formateurs des robots — qu'ils soient de type descendant, ascendant ou mixte, voire qu'ils incluent des ingrédients aléatoires — ne peuvent fournir une base connaissable pour construire un robot aux aptitudes mathématiques égales à celles de l'homme.

3.26 Rupture algorithmique des boucles

Je vais reprendre cette conclusion sous un angle légèrement différent. Afin de contourner les limitations imposées par le théorème de Gödel, on pourrait imaginer un robot qui « sortirait du système » chaque fois que son algorithme de contrôle se trouverait piégé dans une boucle. Puisque le théorème de Gödel ne cesse de poser des difficultés dès qu'on l'applique à l'hypothèse de la réductibilité de la compréhension mathématique à des procédures algorithmiques, il est intéressant d'examiner les difficultés que ce théorème oppose à tout modèle algorithmique de la compréhension mathématique.

Je me suis laissé dire qu'à l'instar des ordinateurs standard et de quelques insectes, certains lézards sont si stupides qu'il leur arrive de décrire des boucles : par exemple, placés en file indienne autour d'une assiette, ils adoptent un comportement panurgesque et se laissent mourir de faim. L'idée est qu'un système véritablement intelligent trouverait toujours un moyen de sortir de telles boucles, ce que ne saurait faire un banal ordinateur. (Ce problème de la « sortie du système » a été examiné dans Hofstadter 1979.)

Le type le plus élémentaire de boucle algorithmique correspond à la situation où le système, à un moment donné, revient exactement à un état dans lequel il s'est déjà trouvé. Si l'on n'introduit alors aucune donnée supplémentaire, le système répète indéfiniment le même calcul. Il ne serait pas difficile de concevoir un système qui, en principe (mais peut-être sans grande efficacité), sortirait de ce type de boucle chaque fois qu'il en rencontrerait un (par exemple en dressant la liste de tous les états dans lesquels il s'est déjà trouvé et en vérifiant, lors de chaque étape de calcul, si l'état où il se trouve fait ou non partie de cette liste). Il existe cependant de nombreux autres types de boucle, bien plus sophistiqués. Le problème des boucles est essentiellement celui sur lequel s'est concentrée toute la discussion du chapitre 2 (en particulier, les sections 2.1 à 2.6), car un calcul qui *boucle* est simplement un calcul qui ne s'arrête pas. Un énoncé qui affirme qu'un calcul décrit une boucle est précisément ce que nous avons appelé un énoncé Π_1 (*cf.* §2.10, réponse à **Q10**). Nous avons vu à la section 2.5 qu'il n'existe aucune procédure entièrement algorithmique permettant de déterminer qu'un calcul ne s'arrêtera pas — *i.e.* qu'il décrira une boucle. En outre, les discussions ci-dessus ont montré que les procédures dont les mathématiciens humains disposent pour prouver qu'un calcul *décrit* une boucle — *i.e.* pour établir la vérité d'un énoncé Π_1 — sont irréductibles à une action algorithmique.

Nous en concluons que seule une « intelligence non algorithmique » peut rendre compte de tous les moyens dont les êtres humains disposent pour affirmer avec *certitude* qu'un calcul donné décrit effectivement une boucle. Certains lecteurs pensent peut-être que l'on pourrait éviter ces boucles en recourant à des mécanismes qui, mesurant la durée d'exécution d'un calcul, le feraient sortir de la boucle s'ils estiment que le calcul a duré suffisamment longtemps et qu'il n'a aucune chance de s'arrêter. Or si l'on suppose que les mécanismes qui prendraient une telle décision sont de nature algorithmique,

cela ne marche pas : il y aurait en effet des mécanismes qui se tromperaient, soit en concluant à tort qu'un calcul est pris dans une boucle alors qu'il n'en est rien, soit en ne parvenant à aucune conclusion (de sorte que le mécanisme se trouverait lui-même piégé dans une boucle). Cela s'explique par le fait que le système entier est une entité algorithmique, de sorte qu'il peut lui-même se retrouver dans une boucle. On ne peut donc jamais être certain, à moins de parvenir à une conclusion erronée, que le système tout entier ne décrit pas lui-même une boucle.

Pourrait-on faire intervenir des éléments aléatoires pour décider de la sortie d'un éventuel calcul circulaire et de l'instant où doit s'effectuer cette sortie ? Nous l'avons remarqué — notamment à la section 3.18 —, les ingrédients purement aléatoires — contrairement aux ingrédients pseudo-aléatoires (algorithmiques) — ne peuvent guère jouer un rôle déterminant dans ce contexte. Toutefois, si en elles-mêmes les procédures aléatoires ne s'avèrent guère utiles pour pouvoir affirmer avec certitude qu'un calcul donné décrit effectivement une boucle — car de par la nature même du hasard, on ne peut avoir aucune certitude sur une conclusion qui *dépend* d'un ingrédient aléatoire —, il existe néanmoins certaines procédures algorithmiques qui, mettant en jeu des ingrédients aléatoires (ou pseudo-aléatoires), donnent des résultats mathématiques dont la validité, si elle n'est pas absolue, est hautement probable. Par exemple, on dispose de tests très efficaces — intégrant une entrée aléatoire — pour décider si un grand nombre est premier. Les réponses données par ces tests ont une probabilité extrêmement élevée d'être correctes : quel que soit le cas considéré, on peut être pratiquement certain qu'elles *sont* correctes. Les tests mathématiquement rigoureux sont bien moins efficaces — et l'on peut se demander si un raisonnement complexe mais mathématiquement précis, qui peut-être contient une erreur, est supérieur à un raisonnement comparativement simple mais probabiliste dont la probabilité d'erreur est en pratique considérablement inférieure. Ce type de situation soulève des problèmes complexes qui ne nous concernent pas ici. Disons simplement que du point de vue « de principe » sous lequel j'ai abordé les choses dans la majeure partie de ce chapitre, un argument probabiliste permettant d'établir par exemple la vérité d'un énoncé Π_1 se révèle toujours suffisant.

Si l'on cherche à établir, d'un point de vue de principe, la vérité incontestable d'énoncés Π_1 sans devoir dépendre de procédures inconnaissables ou aléatoires, on doit disposer d'une réelle *compréhension* du *sens* de ces énoncés. Les procédures par essais et erreurs, bien qu'elles s'avèrent parfois d'utiles auxiliaires, ne fournissent par elles-mêmes aucun critère de vérité définitif.

À titre d'exemple, considérons le calcul mentionné à la section 2.6 dans la réponse à **Q8** : « Imprimez une succession de $2^{2^{65\,536}}$ "1" puis arrêtez-vous. » Pris au pied de la lettre, cet « ordre » ne pourrait être exécuté jusqu'au bout : même si ses diverses étapes s'accomplissaient dans le temps le plus bref que puisse concevoir la physique théorique (environ 10^{-43} seconde) — cette exécution durerait plus longtemps que la durée de vie actuelle (ou prévisible) de l'Univers. Pourtant, ce calcul admet une spécification très simple (notez que $65\,536 = 2^{16}$) et il nous est parfaitement évident qu'il *s'arrêtera* un jour. Affir-

mer que ce calcul s'est enfermé dans une boucle parce qu'il semble « avoir duré suffisamment longtemps » serait une erreur grossière.

Un exemple plus intéressant de calcul dont on sait aujourd'hui qu'il s'arrête — bien qu'il semble devoir se poursuivre indéfiniment — est fourni par une conjecture due au grand mathématicien suisse Leonhard Euler. Ce calcul consiste à trouver, parmi les entiers naturels strictement positifs 1, 2, 3, 4, ..., une solution de l'équation

$$p^4 + q^4 + r^4 = s^4.$$

En 1769, Euler avait émis l'hypothèse que ce calcul ne se terminait jamais, en d'autres termes que cette équation n'avait pas de solutions entières. Au milieu des années 1960, on programma un ordinateur pour en trouver une solution (voir Lander et Parkin 1966), mais la tentative fut stoppée alors qu'aucune solution n'était en vue — les nombres étant devenus trop grands pour être traités par l'ordinateur, les programmeurs renoncèrent. Ce calcul semblait vraiment devoir ne jamais se terminer. En 1987 pourtant, le mathématicien (humain) Noam Elkies trouva une solution (p = 2 682 440, q = 15 365 639, r = 18 796 760 et s = 20 615 673) et montra qu'elle n'était que l'une parmi une infinité de solutions fondamentalement différentes. Encouragé par cette découverte, Roger Frye reprit la recherche sur ordinateur en utilisant plusieurs idées simplificatrices émises par Elkies et trouva, au terme de quelque 100 heures de calcul, une solution passablement plus petite (en fait, c'est la plus petite possible) : p = 95 800, q = 217 519, r = 414 560 et s = 422 481.

La découverte des solutions de ce problème doit autant à l'intuition mathématique qu'aux essais directs sur ordinateur. Elkies s'est en partie aidé de la machine, mais la composante mathématique de sa démonstration, de loin la plus déterminante, ne fait pas appel à l'informatique. A l'inverse, Frye tira un avantage considérable de quelques intuitions humaines qui rendirent possible sa recherche sur ordinateur.

La conjecture originelle émise par Euler en 1769 est en fait une généralisation du célèbre « dernier théorème de Fermat » qui, le lecteur s'en souvient, affirme que l'équation

$$p^n + q^n = r^n$$

n'admet aucune solution entière positive lorsque n est supérieur à 2 (voir par exemple Delvin 1988*). On peut exprimer la conjecture d'Euler sous la forme suivante : « L'équation à m variables p, q, ..., t, u

$$p^n + q^n + \ldots + t^n = u^n$$

* Nombre de lecteurs savent probablement que le « dernier théorème de Fermat » a été finalement démontré, après quelque 350 années de recherches. Sa démonstration a été annoncée par Andrew Wiles, de Cambridge, le 23 juin 1993. Alors que je rédigeais ce livre, j'ai appris que cette démonstration contenant certaines insuffisances, il fallait rester prudent. Toutefois, Wiles a depuis (1994) proposé un autre raisonnement qui suffit à supprimer ces insuffisances.

n'admet pas de solution positive entière pour $m = n \geq 4$. » Le théorème de Fermat correspond au cas $m = 3$, $n \geq 3$ (cas pour lequel les deux problèmes coïncident en $m = n = 3$ et pour lequel Fermat lui-même démontra qu'il est sans solution). Il fallut attendre presque deux cents ans pour obtenir le premier contre-exemple de la conjecture d'Euler — pour $m = n = 5$ — grâce à une recherche sur ordinateur (décrite également dans l'article de Lander et Parkin mentionné plus haut et annonçant l'échec de la recherche d'un contre-exemple pour $m = n = 4$). Ce contre-exemple est :

$$27^5 + 84^5 + 110^5 + 133^5 = 144^5.$$

Il existe un autre exemple célèbre de calcul dont on sait qu'il s'arrête, mais dont on ignore *où* il s'arrête. Il s'agit du problème — dont J. E. Littlewood montra en 1914 qu'il avait une solution *quelque part* — consistant à trouver l'endroit où une certaine formule d'approximation (un logarithme intégral introduit par Gauss) du nombre de nombres premiers inférieurs à un entier positif n donne une valeur inférieure à ce nombre. (En termes techniques, ce problème revient à chercher le point d'intersection de deux courbes.) En 1935, Skewes, un étudiant de Littlewood, montra que cet endroit se situe légèrement au-dessous de $10^{10^{10^{34}}}$, mais sa place exacte demeure aujourd'hui encore inconnue — on sait toutefois que le nombre repère donné par Skewes est considérablement trop élevé. (Ce nombre a été qualifié de « plus grand nombre à être apparu de manière naturelle en mathématiques », mais ce titre lui a été ravi depuis par un autre nombre, bien plus grand, cité dans Graham et Rothschild 1971.)

3.27 Algorithmes descendants ou ascendants ?

Nous avons vu à la section précédente que les ordinateurs peuvent être d'une aide précieuse lors de la résolution de certains problèmes mathématiques. Les procédures algorithmiques qui ont permis de trouver une solution des problèmes cités en exemple étaient toutes de type descendant. J'ignore s'il existe en mathématiques pures un résultat important qui ait été obtenu à l'aide de procédures ascendantes ; il est cependant tout à fait possible que de telles procédures, associées à une procédure principale descendante destinée à résoudre un problème mathématique donné, puissent s'avérer efficaces pour divers types de recherches. Cela étant, je ne connais rien d'intéressant en algorithmique qui ressemble, même de loin, au type de système — tel $\mathbb{Q}^*$ — dont on pourrait imaginer — comme cela a été envisagé aux sections 3.9 à 3.23 — qu'il sous-tend les actions d'une « communauté de robots apprentis mathématiciens ». La contradiction mise en évidence dans le cadre d'une telle hypothèse montre que ces systèmes *ne fournissent pas* de bons moyens algorith-

miques de faire des mathématiques. Les ordinateurs sont certes très précieux en mathématiques lorsqu'ils sont utilisés selon des procédures descendantes, mais c'est la compréhension de l'homme qui fournit l'intuition initiale déterminant exactement le calcul à effectuer et c'est encore elle qui, au terme du calcul, permet d'interpréter les résultats. Une procédure interactive peut s'avérer très efficace, l'homme et l'ordinateur travaillant ensemble, l'intuition humaine intervenant à divers stades de l'opération, mais il est peu avisé et — strictement parlant — impossible de substituer une action entièrement algorithmique à celle de la compréhension humaine.

Les arguments développés dans ce livre l'ont montré, la compréhension mathématique est très différente du calcul et celui-ci ne peut totalement la remplacer. Il peut être une *aide* très précieuse pour la compréhension, mais ne fournit jamais lui-même une compréhension authentique. En fait, la compréhension mathématique sert souvent à *trouver* de bonnes procédures algorithmiques pour résoudre des problèmes. Une fois trouvées, ces procédures prennent le relais en laissant l'esprit libre de se concentrer sur d'autres sujets. Disposer d'une bonne notation, telle celle fournie par le calcul différentiel ou la numération décimale, contribue à ce désengagement de l'esprit. Par exemple, dès que l'on maîtrise l'algorithme de multiplication des nombres, on peut effectuer cette opération de manière entièrement machinale, en « oubliant » *pourquoi* telles règles algorithmiques particulières ont été adoptées plutôt que telles autres.

Nous pouvons conclure de tout cela que la procédure employée par le « robot apprenti » en mathématiques n'est pas celle qui sous-tend la compréhension mathématique chez l'être humain. Une procédure à stratégie essentiellement ascendante paraît d'ailleurs irrémédiablement inadaptée à tout projet *concret* de construction d'un robot mathématicien, même si ce projet n'a pas pour ambition de simuler la compréhension d'un mathématicien humain. Nous l'avons vu plus haut, les procédures ascendantes d'apprentissage sont, *par elles-mêmes*, absolument inefficaces lorsqu'il s'agit d'établir des vérités mathématiques incontestables. Pour obtenir un système numérique donnant des résultats mathématiques incontestables, il serait bien plus efficace de le construire selon une stratégie descendante (du moins pour ce qui concerne le caractère « incontestable » des énoncés qu'il formulera, les procédures ascendantes pouvant s'avérer intéressantes lors des phases exploratoires d'établissement de ces énoncés). Les procédures descendantes doivent leur sûreté et leur efficacité à l'intervention initiale de l'homme qui, par sa compréhension et son intuition, leur fournit les ingrédients supplémentaires qu'elles ne peuvent acquérir par le calcul pur.

En fait, il n'est pas rare aujourd'hui de voir les ordinateurs employés ainsi. L'exemple le plus célèbre fut la démonstration du théorème des quatre couleurs (*cf.* §3.25). Le rôle de l'ordinateur fut ici d'exécuter un algorithme bien précis qui examina un nombre très grand mais fini d'alternatives dont les mathématiciens humains avaient au préalable montré que l'élimination conduisait à une démonstration générale du résultat recherché. Il existe d'autres exemples de démonstrations assistées par ordinateur. De nos jours, il

arrive fréquemment qu'en plus des calculs numériques, un ordinateur exécute des opérations algébriques complexes. Ici aussi, c'est la compréhension humaine qui fournit les règles algorithmiques, et c'est une stratégie strictement descendante qui régit l'activité de l'ordinateur.

Il existe aussi une forme de démonstration appelée « démonstration automatique des théorèmes ». Ce type de démonstration consiste par exemple à définir un système formel $\mathbb{H}$ fixe, puis à démontrer des théorèmes dans le cadre de ce système. Rappelons (*cf.* §2.9) que l'obtention de tous les théorèmes de $\mathbb{H}$, les uns après les autres, est une procédure entièrement algorithmique. Celle-ci peut être automatisée, mais une automatisation qui ne ferait appel ni à la réflexion ni à l'intuition risquerait d'être totalement inefficace. En revanche, si elle y a recours, elle peut donner des résultats tout à fait impressionnants. L'une de ces procédures automatiques (Chou 1988) a permis de transformer les règles de la géométrie euclidienne en un très efficace système de démonstration (voire de découverte) de théorèmes géométriques. Par exemple, l'un de ces théorèmes, connu sous le nom de conjecture de Thébault, fut formulé en 1938 (et démontré récemment, en 1983, par K. B. Taylor) ; lorsqu'on soumit cette conjecture à la procédure automatique, celle-ci la démontra en 44 heures[12].

Les diverses tentatives visant, ces dix dernières années, à élaborer des algorithmes d'« intelligence artificielle » capables de manifester une « compréhension » mathématique sont, elles, plus proches des procédures examinées dans les précédentes sections[13]. J'espère qu'il est clair, d'après les arguments que j'ai donnés, que quels que soient leurs succès, ces systèmes *ne peuvent parvenir* à une authentique compréhension mathématique ! Parmi ces tentatives figurent celles qui visent à construire des systèmes *générateurs* de théorèmes dont l'objectif serait de trouver des théorèmes jugés « intéressants » selon certains critères intégrés dans ces systèmes. Presque tout le monde, cependant, s'accorde pour reconnaître que les résultats obtenus ne présentent pas de réel intérêt mathématique. Bien sûr, on alléguera que ce n'est qu'un début et que l'avenir réserve des choses bien plus exaltantes. Toutefois, et ceci devrait être clair pour tous ceux qui m'ont lu jusqu'ici, je considère personnellement que toutes ces entreprises n'ont guère de chances d'aboutir à des résultats véritablement positifs, sauf à mettre en évidence ce que ces systèmes sont justement *incapables* d'accomplir.

3.28 Conclusions

L'argumentation développée dans ce chapitre montre clairement, me semble-t-il, que la compréhension mathématique humaine est irréductible à des mécanismes numériques (connaissables) — quelle que soit la combinaison de procédures descendantes, ascendantes ou aléatoires composant ces mécanis-

mes. Nous sommes donc apparemment inéluctablement amenés à conclure que la compréhension humaine renferme un élément essentiel qu'aucun dispositif algorithmique ne peut simuler. Sur un plan strictement logique, il existe certes quelques contre-arguments qui s'opposent à cette conclusion, mais ceux-ci semblent extrêmement peu plausibles. On a par exemple invoqué une « intervention divine » — elle aurait implanté dans notre cerveau numérique un merveilleux algorithme qui nous est par principe inconnaissable — ou affirmé que les mécanismes mêmes qui régissent l'amélioration de nos performances sont par principe mystérieux et inconnaissables. Quiconque désire construire concrètement un dispositif authentiquement intelligent ne peut admettre la pertinence de ces deux contre-arguments. Pour ma part, ils ne m'apparaissent absolument pas admissibles — ni réellement crédibles.

On pourrait également objecter qu'il n'existe en fait aucune garantie, telle celle fournie par les limites T, L et N, permettant d'éliminer totalement toutes les assertions ☆ erronées correspondant aux énoncés Π_1 de complexité inférieure à c (*cf.* §3.19-3.21). Il m'apparaît cependant très difficile de croire en l'existence d'une « conspiration » qui s'opposerait à l'élimination de toutes les erreurs, alors même qu'une société d'élite de nos robots s'emploierait à les supprimer aussi méticuleusement que possible. En outre, les énoncés Π_1 à vérifier sont en nombre fini. La mise en commun du comportement de la totalité des robots devrait permettre d'éliminer tous les lapsus éventuels, car il serait peu probable que si un même lapsus apparaît à plusieurs reprises, il puisse être le fait d'autre chose qu'une petite minorité des diverses matérialisations de la société de robots simulée — à condition qu'il s'agisse *vraiment* d'un lapsus et non d'une erreur qu'un blocage fondamental empêcherait les robots de percevoir. De tels blocages ne correspondent pas à des erreurs « rectifiables » ; or notre objectif est d'éliminer les erreurs qui sont, en un certain sens, « rectifiables ».

Le contre-argument (éventuellement) restant concerne l'action du chaos. Se peut-il que le détail du comportement de certains systèmes chaotiques contienne un aspect essentiellement *non* aléatoire, et que cet aspect, situé à la « frontière du chaos », constitue la clé de la non-calculabilité du comportement de l'esprit ? Pour qu'il en soit ainsi, il faudrait que ces systèmes chaotiques soient capables de fournir une approximation d'un comportement non calculable — éventualité en elle-même très intéressante. Pourtant, même si tel est le cas, cet aspect non aléatoire aurait simplement pour effet de réduire — légèrement — la taille de l'ensemble des sociétés de robots considéré (*cf.* §3.22). Je ne vois absolument pas en quoi cela serait utile. Ceux qui croient que le chaos constitue la clé de l'activité cérébrale doivent trouver un argument levant ces profondes difficultés.

Les raisonnements précédents semblent sérieusement mettre en cause le modèle algorithmique de l'esprit — point de vue $\mathscr{A}$ — tout comme la possibilité d'une *simulation* numérique efficace (mais stupide) de toutes les manifestations externes de l'activité de l'esprit — point de vue $\mathscr{B}$. Pourtant, en dépit de la force de ces raisonnements, je pense que beaucoup de personnes, même de formation scientifique, les trouveront difficilement acceptables et,

plutôt que d'explorer la possibilité que le phénomène de l'esprit — quoi qu'il soit — soit plus conforme au point de vue $\mathscr{C}$, voire au point de vue $\mathscr{D}$, seront davantage tentées de trouver des points faibles dans ces raisonnements afin de préserver ce qu'elles pensent être de toute façon la vérité, à savoir le point de vue $\mathscr{A}$ ou le point de vue $\mathscr{B}$.

Cette réaction n'est certes pas déraisonnable, car $\mathscr{C}$ et $\mathscr{D}$ donnent eux-mêmes naissance à de profondes difficultés. Si l'on croit, en accord avec $\mathscr{D}$, que l'esprit est inexplicable en termes scientifiques — qu'il est une qualité totalement distincte de tout ce que peuvent engendrer les entités physiques mathématiquement déterminées qui composent notre Univers —, nous devons alors nous demander pourquoi notre esprit semble si intimement associé à un corps physique de structure très sophistiquée, à savoir notre cerveau. Si les processus mentaux sont si différents des processus physiques, pourquoi notre moi mental semble-t-il avoir besoin de notre cerveau physique ? Manifestement, les différences d'état mental résultent de modifications survenant dans l'état physique du cerveau. L'absorption de drogues, par exemple, modifie très spécifiquement l'expérience et le comportement mentaux. De même, une blessure, une maladie ou une opération chirurgicale en un endroit particulier du cerveau ont des effets clairement définis et prédictibles sur les états mentaux de l'individu concerné. (Les livres d'Oliver Sacks, *L'Eveil* (1973) et *L'Homme qui prenait sa femme pour un chapeau* (1985) contiennent à cet égard des récits particulièrement impressionnants.) Il semble donc difficile de soutenir que les processus mentaux et physiques soient de natures *totalement* différentes. Si les processus mentaux sont effectivement liés à certains processus physiques — et ils semblent l'être de manière *intime* —, les lois scientifiques qui décrivent si précisément le comportement des corps physiques ont probablement aussi quantité de choses à nous révéler sur le monde de l'esprit.

En ce qui concerne le point de vue $\mathscr{C}$, les problèmes sont différents — et résultent essentiellement du caractère particulièrement spéculatif de ce point de vue. Quelles raisons incitent à penser que la Nature a parfois un comportement irréductible au calcul ? Il est incontestable que le pouvoir de la science moderne résulte, à un degré sans cesse croissant, du fait que des calculs numériques de plus en plus détaillés parviennent à simuler, avec une précision de plus en plus grande, le comportement des corps physiques. Et l'accroissement du pouvoir prédictif des simulations numériques a accompagné celui de la compréhension scientifique. Matériellement, cet accroissement est essentiellement dû au développement rapide — survenu principalement durant la dernière partie de ce siècle — d'instruments de calcul d'une puissance, d'une vitesse et d'une précision extraordinaires. Ainsi, on peut observer un lien de plus en plus étroit entre l'activité des ordinateurs modernes et l'action même de l'univers matériel. Cela correspond-il à une phase temporaire du développement scientifique ? Qu'est-ce qui autorise à envisager l'éventualité d'une action physique renfermant un aspect irréductible à un traitement numérique efficace ?

Si l'on cherche, au sein de la théorie physique *actuelle*, des signes d'une action se dérobant partiellement au calcul, on sera immanquablement déçu.

Toutes les lois physiques connues, de la dynamique de Newton aux profondes intrications de la théorie quantique moderne, en passant par les champs électromagnétiques de Maxwell et les espaces-temps courbes d'Einstein, semblent descriptibles en termes entièrement numériques[14], à l'exception du processus de « mesure quantique » — dans lequel des effets d'une taille initialement infime se trouvent amplifiés jusqu'à être objectivement perçus — qui met en jeu un ingrédient entièrement aléatoire. Aucune de ces lois ne présente quoi que ce soit qui corresponde à une « action physique ne pouvant être numériquement simulée », ainsi que le requiert le point de vue $\mathcal{C}$. J'en conclus que c'est à la version « forte » de $\mathcal{C}$, plutôt qu'à sa version « faible », que nous devons nous rallier (*cf.* §1.3).

On ne saurait trop insister sur ce point. Si diverses personnes de culture scientifique me disent admettre le point de vue défendu dans EOLP, selon lequel l'activité de l'esprit possède une caractéristique « non algorithmique », elles soutiennent dans le même temps que cette caractéristique *n'est pas* nécessairement associée à un aspect encore inconnu de la théorie physique. Peut-être ces personnes songent-elles à l'extrême complexité des processus intervenant dans l'activité cérébrale, complexité qui dépasse de loin l'analogie classique (formulée pour la première fois par McCullogh et Pitts en 1943) assimilant les neurones (ou les jonctions synaptiques) et les axones aux transistors et aux câbles d'un circuit électronique. Peut-être aussi pensent-elles à la complexité de la chimie mise en jeu dans le comportement des neurotransmetteurs qui régissent la transmission synaptique, ainsi qu'au fait que l'action de ces substances chimiques n'est pas nécessairement confinée au voisinage de jonctions synaptiques particulières. Peut-être également pensent-elles à la complexité des neurones eux-mêmes[15], dans lesquels des sous-structures importantes (tel le cytosquelette — qui aura effectivement une grande importance pour nous ; *cf.* §7.4-§7.7) pourraient avoir une influence déterminante sur l'activité neuronale. Peut-être même pensent-elles aux influences électromagnétiques directes, telles les « effets de résonance », qui ne pourraient s'expliquer en termes d'impulsions nerveuses ordinaires, ou encore à un rôle important des incertitudes quantiques ou d'effets quantiques collectifs non locaux (tel le phénomène appelé « condensation de Bose-Einstein »[16]) dans l'activité cérébrale.

Bien que l'on soit encore loin de disposer de théorèmes mathématiques définitifs[17], tout semble indiquer que les théories physiques existantes sont toutes réductibles au calcul — avec peut-être l'intervention, de temps en temps, d'un ingrédient aléatoire associé aux mesures quantiques. Certes la présence de phénomènes (non aléatoires) non algorithmiques dans les systèmes physiques, phénomènes qui seraient régis par la théorie physique existante, est une éventualité très intéressante qui mérite d'être étudiée en détail. Peut-être même cette étude nous réservera-t-elle des surprises et mettra-t-elle en évidence l'existence d'un subtil ingrédient non algorithmique. Mais en l'état actuel des choses, cela semble très peu probable. C'est donc à mon avis par une analyse des points faibles des lois physiques existantes que nous avons le plus de chances de déceler la non-calculabilité qui, selon les arguments développés jusqu'ici, serait associée à l'activité mentale de l'homme.

Quels sont ces points faibles ? Je n'ai pour ma part aucun doute sur l'endroit où nous devons concentrer notre attaque : le point le plus faible de la théorie actuelle est en effet le processus de « mesure quantique » mentionné à l'instant. L'ensemble de ce processus présente des incohérences — et est assurément l'objet de controverses. On ne sait même pas clairement, dans une situation donnée, à quel stade on doit l'appliquer. En outre, la présence d'un élément essentiellement aléatoire dans ce processus fournit une action physique dont la nature est apparemment très différente de ce que l'on connaît pour les autres processus fondamentaux. J'examinerai ces sujets en détail dans la deuxième partie.

À mon avis, ce processus de mesure exige une réelle attention, car son interprétation met en jeu des changements fondamentaux dans le cadre même de la théorie physique. J'émettrai dans la deuxième partie quelques suggestions nouvelles (§6.12). Le raisonnement présenté dans cette première partie qui s'achève suggère fortement qu'il faut remplacer la composante purement *aléatoire* de l'actuelle théorie de la mesure par une autre composante, dans laquelle des ingrédients essentiellement non calculables joueraient un rôle fondamental. Nous le verrons par la suite (§7.9), il faut en outre que cette non-calculabilité soit d'un type particulièrement sophistiqué. (Par exemple, il faut qu'elle donne plus qu'une loi permettant « simplement » de décider, grâce à un nouveau processus physique, de la vérité des énoncés Π_1 — *i.e.* de résoudre le « problème de l'arrêt » de Turing.)

La quête d'une nouvelle théorie physique aussi sophistiquée représenterait à elle seule un défi. Or nous devrons en outre rechercher un fondement physique plausible garantissant que cette non-calculabilité a un lien réel avec l'activité cérébrale — et qu'elle est compatible avec les limitations et les exigences de crédibilité des connaissances actuelles sur l'organisation du cerveau. Vu l'état actuel de ces connaissances, cette recherche comportera inéluctablement elle aussi une bonne part de spéculations. Toutefois, ainsi que je le soulignerai dans la deuxième partie (§7.4), il existe quelques possibilités réelles — que j'ignorais à l'époque où je rédigeais EOLP. Celles-ci semblent correspondre, de manière plus convaincante qu'on ne l'avait imaginé auparavant, à l'existence, au niveau du cytosquelette des neurones, d'une action déterminante située à la frontière des physiques classique et quantique. Ces sujets seront également discutés dans la deuxième partie (§7.5 à 7.7).

J'insiste à nouveau sur ce point : il ne s'agit pas de rechercher simplement de la *complexité* dans le cadre de la théorie physique existante. Certaines personnes soutiennent par exemple que les mouvements et l'activité chimique complexes des neurotransmetteurs ne pouvant absolument pas être correctement simulés, l'activité physique détaillée du cerveau est totalement hors de portée d'une représentation algorithmique. Toutefois, ce n'est pas là ce que j'entends par comportement non algorithmique. Nos connaissances sur la structure biologique et les mécanismes chimiques et électriques gouvernant l'activité cérébrale sont certes insuffisantes pour permettre une simulation numérique ; et même si ces connaissances étaient suffisantes, la puissance de nos ordinateurs actuels et l'état d'avancement de nos techniques de program-

mation ne seraient certainement pas à même d'autoriser une simulation acceptable en un temps de calcul raisonnable. Mais rien, *en principe*, ne s'oppose à la réalisation d'une telle simulation dans le cadre des modèles existants, simulation qui inclurait tout ce qui concerne la chimie des neurotransmetteurs, les mécanismes gouvernant leur transport, leur efficacité en fonction du milieu ambiant particulier, des potentiels d'action, du champ électromagnétique, etc. Ainsi, les mécanismes de ce type, supposés compatibles avec les exigences de la théorie physique existante, ne peuvent fournir la non-calculabilité imposée par les arguments précédents.

Une telle simulation numérique (de principe) pourrait intégrer des éléments chaotiques. Nous l'avons dit lors de la discussion des systèmes chaotiques (§1.7, §3.10, §3.11, §3.22), nous n'exigeons toutefois pas que cette simulation corresponde à l'activité d'un cerveau *particulier*; il suffit qu'elle corresponde à un « cas typique » d'activité cérébrale. L'intelligence artificielle vise en effet à simuler non pas les aptitudes mentales d'un individu particulier, mais celles d'un individu *typique*. (Rappelons l'exemple de la météorologie qui, dans l'état actuel des connaissances, cherche simplement à simuler *une* situation météo, et non *la* situation météo !) Quels que soient les *mécanismes* que l'intelligence artificielle proposera pour simuler l'activité cérébrale, ils constitueront de toute façon (s'ils se formulent dans le cadre de la physique réductible au calcul que nous connaissons aujourd'hui) un système algorithmique connaissable — intégrant peut-être des ingrédients aléatoires explicites, mais tout cela a déjà été pris en compte par la discussion donnée plus haut.

On pourrait même pousser plus loin cet argument et envisager pour le cerveau un modèle se constituant sur un mode d'évolution darwinienne, à partir de formes de vie primitives régies par la physique connue — ou toute autre physique algorithmiquement simulable (telle celle de l'ingénieux modèle mathématique bidimensionnel du « jeu de la vie » inventé par John Horton Conway[18]). On pourrait imaginer que cette évolution donne naissance à une « société de robots » (*cf.* §3.5, §3.9, §3.19 et §3.23). Ici encore, le système résultant serait algorithmique et on pourrait lui appliquer les raisonnements des sections 3.14 à 3.21. Pour que le concept d'« assertion ☆ » trouve sa place dans ce système — de sorte que nous puissions lui appliquer en détail ces raisonnements —, il faudrait que cette évolution connaisse une phase d'« intervention humaine » au cours de laquelle les robots apprendraient la stricte signification de l'imprimatur « ☆ ». Cette phase pourrait être déclenchée automatiquement de manière à survenir lorsque les robots commenceraient à avoir des capacités de communication convenables — selon un critère à définir. Il n'y a apparemment aucune raison pour que tout cela ne puisse être automatisé sous forme d'un système algorithmique connaissable (dans le sens où ces mécanismes seraient connaissables même s'ils ne seraient pas nécessairement programmables sur un ordinateur constructible dans un avenir prévisible). Comme précédemment, la contradiction découle de l'hypothèse selon laquelle un tel système pourrait atteindre un niveau de compréhension égal à celui de l'être humain et suffisant pour apprécier le théorème de Gödel.

Une autre objection soulevée par certaines personnes[19] concerne la pertinence d'arguments mathématiques tels ceux sur lesquels je me suis appuyé pour traiter de psychologie : l'activité mentale humaine ne serait pas suffisamment précise pour permettre cette analyse mathématique. Ces personnes ont probablement le sentiment que les arguments détaillés concernant la nature mathématique d'une physique, quelle qu'elle soit, sous-jacente à notre activité cérébrale peuvent n'avoir aucun rapport avec notre compréhension de l'activité de l'esprit humain. Si elles reconnaissent éventuellement que le comportement humain est effectivement « non calculable », elles affirment que cette non-calculabilité reflète simplement une inaptitude des considérations physico-mathématiques à rendre compte des problèmes de psychologie humaine. Elles soutiennent par exemple — et non sans raisons — que l'organisation fantastiquement complexe de notre cerveau, ainsi que de notre société et de notre éducation, est bien plus déterminante que toute théorie physique qui serait sous-jacente aux mécanismes spécifiques gouvernant le fonctionnement du cerveau humain.

Il importe toutefois de souligner que la complexité ne peut à elle seule dispenser d'examiner les conséquences de ces lois physiques sous-jacentes. Un athlète humain, par exemple, est un système physique immensément complexe, et en vertu d'un tel argument, on pourrait imaginer que le détail des lois physiques sous-jacentes a peu de rapport avec les performances de cet athlète. Pourtant, nous savons que cela est très loin de la vérité. Les principes physiques généraux qui régissent la conservation de l'énergie, de la quantité de mouvement et du moment cinétique, ainsi que les lois gouvernant l'attraction gravitationnelle, appliquent sur le corps de l'athlète un contrôle aussi ferme que celui qu'ils exercent sur les particules individuelles composant ce corps. Le fait qu'il ne peut en être autrement résulte de caractéristiques très précises des principes gouvernant notre univers particulaire. Avec des principes — même légèrement — différents (ou radicalement différents, tels ceux opérant dans le « jeu de la vie » de Conway), les lois qui régissent le comportement d'un système d'une complexité équivalente à celle d'un athlète pourraient facilement être *totalement* différentes. La même chose vaut pour les mécanismes d'un organe tel que le cœur, ainsi que pour les processus chimiques gouvernant d'innombrables actions biologiques. De même, il est difficilement concevable que le détail des lois qui sous-tendent l'activité cérébrale ne joue pas un rôle extrêmement important dans le contrôle ne serait-ce que des plus grossières manifestations de l'esprit humain.

Pourtant, même si l'on accepte tout cela, on peut encore objecter que le raisonnement tenu jusqu'ici sur le comportement général (« à haut niveau ») des mathématiciens humains a peu de chance de fournir la moindre indication significative sur la physique sous-jacente à ce comportement. En définitive, l'argumentation gödelienne exige une attitude strictement rationnelle à l'égard de ce que chacun estime être ses propres convictions mathématiques « inébranlables », attitude bien peu répandue dans le cadre du comportement humain ordinaire. Par exemple, il existe en psychologie des expériences[20]

démontrant à quel point sont irrationnelles les réponses de sujets auxquels on pose des questions telles que :

> « Si tous les A sont des B et si certains B sont des C, s'ensuit-il nécessairement que certains A sont des C ? »

À ce type de question, une majorité d'étudiants a donné la mauvaise réponse (« oui »). Si les étudiants ordinaires sont à ce point illogiques, on peut légitimement se demander comment on peut déduire la moindre conclusion valide d'un raisonnement bien plus sophistiqué, à savoir l'argumentation gödelienne ! Même les mathématiciens professionnels commettent souvent des erreurs de raisonnement, et il est rare qu'ils s'expriment avec une rigueur suffisante pour permettre une application des techniques gödeliennes.

Qu'il soit bien clair cependant que les principaux arguments contenus dans ce livre ne concernent pas les erreurs analogues à celle évoquée à l'instant à propos des étudiants. Ces erreurs se classent sous la rubrique « erreurs rectifiables » — et de fait, les étudiants reconnaissent leur erreur une fois qu'on la leur fait remarquer (au besoin en fournissant une explication détaillée). Les erreurs rectifiables ne nous concernent pas ici (voir notamment la discussion sous **Q13**, ainsi que les sections 3.12 et 3.17). Si l'étude des erreurs revêt une certaine importance en psychologie, en psychiatrie et en physiologie, les problèmes qui nous intéressent ici sont totalement différents et portent sur ce que l'être humain peut percevoir à l'aide de sa compréhension, de son raisonnement et de son intuition. Il s'avère que ces problèmes sont en fait très subtils, même si cette subtilité ne saute pas aux yeux. Au premier abord, ces problèmes semblent triviaux : car que peut-on dire de mieux sinon qu'un raisonnement correct est un raisonnement correct — ce qui est plus ou moins évident et, en tout cas, parfaitement défini par Aristote depuis 2 300 ans (ou du moins par le logicien George Boole depuis 1854, etc.) ! Il s'avère toutefois qu'un « raisonnement correct » est quelque chose d'immensément subtil et, ainsi que Gödel (et Turing) l'a montré, irréductible à une action purement algorithmique. Ces problèmes ont par le passé davantage été du ressort des mathématiciens que des psychologues, et les subtilités qu'ils renferment n'ont généralement pas préoccupé ces derniers. Mais nous avons vu qu'ils nous renseignent sur les actions physiques ultimes se trouvant probablement à l'origine des processus qui sous-tendent notre compréhension consciente.

Ces problèmes soulèvent également de profondes questions philosophiques concernant les mathématiques. La compréhension mathématique constitue-t-elle une sorte de contact avec une réalité mathématique platonicienne préexistante, possédant une actualité intemporelle totalement indépendante de nous, ou bien chacun de nous recrée-t-il indépendamment tous les concepts mathématiques lorsqu'il réfléchit à l'aide de raisonnements logiques ? En outre, pourquoi les lois physiques semblent-elles suivre si fidèlement des descriptions mathématiques si précises et si subtiles ? Comment la réalité physique s'articule-t-elle avec l'idée d'une réalité mathématique platonicienne ? Et s'il s'avère effectivement que la nature de nos perceptions dépend d'une subtile sous-

structure sous-jacente aux lois mêmes qui gouvernent le fonctionnement de notre cerveau, quels enseignements sur la compréhension des mathématiques — et sur le processus même de compréhension — pouvons-nous tirer d'une connaissance plus profonde de ces lois physiques ?

Ces problèmes constituent nos interrogations fondamentales. Nous y reviendrons à la fin de la deuxième partie.

Deuxième partie

Quelle nouvelle physique pour comprendre l'esprit ?

À la recherche d'une physique non calculable de l'esprit

4
Le statut de l'esprit dans la physique classique

4.1 Esprit et lois de la physique

Nous — notre corps et notre esprit — faisons partie d'un Univers qui obéit, avec une extraordinaire précision, à des lois mathématiques d'une portée et d'une subtilité immenses. La science moderne reconnaît certes que notre corps physique est rigoureusement soumis à ces lois, mais qu'en est-il de notre esprit ? L'idée qu'il puisse être lui aussi soumis à ces mêmes lois mathématiques apparaît profondément inquiétante à nombre de personnes. Pourtant, l'éventuelle existence d'une frontière nette entre corps et esprit — le premier étant soumis aux lois mathématiques de la physique, le second jouissant de sa liberté propre — ne serait pas moins inquiétante, car on ne peut nier que si notre esprit influence le comportement de notre corps, il est en retour lui-même influencé par l'état physique de ce même corps. Le concept d'esprit s'avérerait passablement inutile si notre esprit ne pouvait ni influencer notre corps physique, ni être influencé par lui. Et considérer l'esprit comme un simple « épiphénomène » de l'action du corps — comme une caractéristique précise mais dépendant passivement de l'état physique du cerveau — dépourvu de toute influence sur le corps, semble le vouer à une impuissance frustrante. Si cependant l'esprit pouvait influencer le corps en le faisant agir hors des contraintes imposées par les lois physiques, cela perturberait la précision de ces dernières. Ainsi, on peut difficilement adopter une vision totalement « dualiste » dans laquelle le corps et l'esprit obéiraient à des lois complètement indépendantes. Même si les lois physiques qui gouvernent l'action du corps laissent à l'esprit la latitude d'affecter en retour le comportement du corps,

cette latitude est nécessairement elle-même une composante importante de ces mêmes lois physiques. Quel que soit l'élément qui contrôle ou décrit l'esprit, il fait en effet partie intégrante de l'unique organisation grandiose qui gouverne *tous* les attributs *matériels* de notre Univers.

Certaines personnes[1] affirment qu'en considérant l'« esprit » comme une certaine substance — même différant de la matière et vérifiant des principes totalement différents —, on commet une « erreur de catégorie ». Ces personnes comparent parfois le corps à un ordinateur et l'esprit à un programme informatique. Ce type de comparaison peut s'avérer utile lorsqu'il est approprié et il importe certes d'éviter des confusions entre concepts différents lorsque ces confusions sont manifestes. Cependant, le simple fait d'invoquer une éventuelle « erreur de catégorie » dans le cas du corps et de l'esprit ne supprime pas pour autant ce qui est un authentique mystère.

Certains concepts physiques peuvent d'ailleurs être considérés comme identiques, même si à première vue on pourrait penser que c'est là une erreur de catégorie. Considérons par exemple la célèbre équation d'Einstein $E = mc^2$ affirmant l'identité de la masse et de l'énergie. Cette équation a au premier abord l'apparence d'une erreur de catégorie : la masse mesure en effet la substance matérielle réelle, tandis que l'énergie semble être une quantité abstraite plus nébuleuse décrivant un travail à l'état potentiel. Pourtant, la formule d'Einstein, qui lie ces deux concepts, est une pierre angulaire de la physique moderne et a été expérimentalement confirmée à de nombreuses reprises. Un exemple encore plus frappant d'apparente erreur de catégorie, emprunté lui aussi à la physique, est fourni par le concept d'*entropie* (*cf.* par exemple, EOLP, chapitre 7). L'entropie a en effet une définition très subjective dans la mesure où c'est, essentiellement, une caractéristique de la notion d'« information » ; pourtant, elle est liée à d'autres quantités physiques plus « matérielles » par des équations mathématiques bien précises[2].

Rien n'interdit pareillement de procéder, même à titre d'essai, à une analyse du concept d'« esprit » qui le mettrait clairement en relation avec d'autres concepts physiques. La conscience, en particulier, semble être une « donnée » associée à certains objets physiques relativement précis — ne serait-ce qu'aux cerveaux humains vivants —, de sorte qu'aussi infime que puisse être notre compréhension actuelle de ce phénomène, on peut prévoir qu'il admet probablement une description physique. Le seul indice que nous a livré l'analyse effectuée dans la première partie de ce livre est que la compréhension consciente, en particulier, met en jeu une action physique non algorithmique — du moins si, comme moi, on souscrit au point de vue $\mathcal{C}$ et non aux points de vue $\mathcal{A}$, $\mathcal{B}$ ou $\mathcal{D}$ (*cf.* §1.3). J'invite les lecteurs qui ne sont pas encore convaincus par les arguments que j'ai donnés en faveur de $\mathcal{C}$ à me suivre encore un peu afin de découvrir le territoire que nous ouvre ce point de vue. Son exploration va nous montrer que les choses sont loin d'être aussi défavorables qu'on pourrait le penser et qu'il recèle de nombreux aspects dignes d'intérêt. J'espère qu'au terme de cette exploration, ces lecteurs éprouveront plus de sympathie pour les arguments — que je crois très puissants — présentés dans ce livre. Lançons-nous donc dans cette exploration — et laissons-nous guider par $\mathcal{C}$!

4.2 La calculabilité et le chaos dans la physique actuelle

Si les lois physiques actuelles ont une précision et une portée fantastiques, elles ne contiennent cependant aucun indice de l'existence d'une action qui ne serait pas simulable sur ordinateur. Demandons-nous cependant si, tout en respectant les contraintes qu'elles imposent, ces lois n'autoriseraient pas une action non algorithmique cachée dont le fonctionnement de notre cerveau pourrait tirer avantage. Je remets à plus tard (§7.9) l'examen de la nature de cette non-calculabilité. Plusieurs raisons laissent à penser qu'elle est particulièrement subtile et élusive, et je ne veux pas, pour l'instant, m'enfermer dans les problèmes que cette analyse peut soulever. Disons seulement qu'elle met en jeu une représentation fondamentalement différente de ce à quoi nous ont jusqu'ici habitués les théories physiques, qu'elles soient classique ou quantique.

En physique *classique*, on peut, à un instant particulier, spécifier toutes les données nécessaires à la définition d'un système physique, et l'évolution de ce système est non seulement entièrement déterminée mais aussi totalement *calculable* — à l'aide des méthodes de Turing — à partir ces données. Cependant, pour que ce calcul soit, *en principe*, réalisable, il faut que deux conditions (mutuellement dépendantes) soient satisfaites. La première est que l'on puisse discrétiser convenablement ces données initiales, de sorte qu'il soit possible, avec un degré d'approximation suffisant, de substituer des paramètres *discrets* aux paramètres continus de la théorie. (C'est en fait normalement ainsi que l'on procède lorsqu'on exécute une simulation numérique d'un système classique.) La seconde condition concerne le fait que de nombreux systèmes physiques sont *chaotiques* — dans le sens où la connaissance de leur évolution avec un degré de précision acceptable exigerait une connaissance de leurs données initiales avec une précision totalement irréalisable. Nous en avons longuement discuté (voir en particulier la section 1.7 ; voir aussi les sections 3.10 et 3.22), la présence d'un comportement chaotique dans un système discrétisé *ne fournit pas* le type de « non-calculabilité » que nous recherchons. Bien qu'il soit difficile de déterminer avec précision son évolution, un système chaotique (discret) reste un système calculable — ainsi qu'en témoigne le fait que l'étude pratique de ces systèmes s'effectue habituellement sur des ordinateurs ! La première condition est liée à la seconde dans la mesure où, pour un système chaotique, l'« adéquation » de la précision de la discrétisation des paramètres continus de la théorie change selon que l'on s'intéresse au comportement *réel* ou à un comportement *typique* du système. Si l'on s'intéresse à un comportement typique — et, ainsi que je l'ai montré dans la première partie, cela semble suffire pour atteindre les objectifs de l'intelligence artificielle —, peu importe que la discrétisation soit imparfaite et que de petites erreurs sur les données initiales puissent conduire à des erreurs très grandes sur le comportement ultérieur du système. Si donc l'on ne s'intéresse qu'à un comportement typique, les deux conditions ci-dessus ne semblent pas pouvoir permettre l'apparition,

dans un système physique purement classique, de la non-calculabilité requise par les analyses de la première partie du livre.

Rien toutefois ne permet d'écarter l'éventualité de la présence, dans le comportement chaotique précis d'un système mathématique continu (servant comme modèle de comportement physique), d'un élément ne pouvant être simulé par *aucune* approximation discrète. J'ignore si un tel système existe, mais même s'il existe, il ne serait d'aucune utilité pour l'IA — dans son état d'avancement actuel —, car l'IA actuelle dépend de la simulation par *discrétisation* (autrement dit, du calcul numérique et non analogique ; *cf.* §1.8).

En physique *quantique*, outre le comportement déterministe (et calculable) donné par les équations de la théorie (essentiellement, l'équation de Schrödinger), existe également une certaine liberté d'une nature totalement *aléatoire*. Techniquement parlant, les équations de la théorie quantique *ne sont pas* chaotiques, mais l'absence de chaos est remplacée par la présence de ces ingrédients aléatoires, qui s'ajoutent à l'évolution déterministe. Nous l'avons vu (notamment à la section 3.18), ces ingrédients purement aléatoires ne fournissent pas eux non plus l'action non algorithmique recherchée. Il apparaît ainsi qu'aucune des deux physiques, classique et quantique, telles qu'on les comprend aujourd'hui, n'autorise l'existence d'une telle action. C'est donc ailleurs que nous devons la rechercher.

4.3 La conscience : une nouvelle physique ou un « phénomène émergent » ?

Dans la première partie, j'ai soutenu (dans le cas particulier de la compréhension mathématique) la thèse selon laquelle le phénomène de la *conscience* ne peut survenir qu'en présence de certains processus physiques non algorithmiques intervenant dans le cerveau. On doit supposer toutefois que ces (hypothétiques) processus non algorithmiques sont *également* inhérents à l'action de la matière inanimée, car le cerveau humain est en définitive composé de la même matière et obéit aux mêmes lois physiques que les objets inanimés de l'Univers. Nous devons alors nous poser deux questions. Premièrement, comment se fait-il que le phénomène de la conscience semble *uniquement* se manifester, pour autant que l'on sache, dans (ou en relation avec) le cerveau — bien que nous ne puissions exclure l'éventualité d'une présence de la conscience dans d'autres systèmes physiques ? Deuxièmement, comment se peut-il qu'un (hypothétique) ingrédient apparemment aussi important que le comportement non algorithmique, censé être inhérent — du moins potentiellement — à l'action de tout corps matériel, ait jusqu'ici totalement échappé à l'attention des physiciens ?

Nul doute que la réponse à la première question a quelque chose à voir avec l'organisation subtile et complexe du cerveau, mais cette explication est, à elle

seule, insuffisante. En vertu des idées que j'expose dans ce livre, l'organisation du cerveau, contrairement à celle des objets matériels ordinaires, doit permettre de tirer avantage de la présence d'une action non calculable dans les lois physiques. Cette idée se démarque radicalement d'une autre plus couramment admise sur la nature de la conscience[3] (essentiellement celle de $\mathcal{A}$) selon laquelle la connaissance immédiate consciente serait une sorte de « phénomène émergent » survenant simplement en tant que caractéristique d'une complexité ou d'une sophistication d'action suffisantes, et n'exigerait l'intervention d'aucun processus physique sous-jacent particulier et encore inconnu, fondamentalement différent de ceux avec lesquels nous a déjà familiarisés le comportement de la matière inanimée. L'argumentation développée dans la première partie correspond à une autre conception et exige la présence dans le cerveau d'une organisation subtile spécialement réglée pour tirer avantage d'une physique non calculable. Je reviendrai plus en détail sur la nature de cette organisation aux sections 7.4 à 7.7.

En ce qui concerne la seconde question, il y a effectivement de fortes chances pour que des vestiges d'une telle non-calculabilité soient également présents, à un niveau indiscernable, dans la matière inanimée. Pourtant, la physique de la matière ordinaire ne semble pas, du moins au premier abord, admettre un tel comportement non algorithmique. J'expliquerai plus loin en détail comment ce comportement pourrait en fait avoir échappé à l'attention des physiciens et en quoi il est compatible avec les observations actuelles. Il suffira pour l'instant de décrire un phénomène, tout à fait différent mais en un sens très analogue, emprunté à la physique *connue*. Bien qu'il n'ait aucun lien — du moins aucun lien *direct* — avec un comportement non calculable, ce phénomène physique connu ressemble énormément à notre hypothétique ingrédient non calculable, par le fait qu'il est totalement indiscernable — mais réellement présent — dans le comportement détaillé des objets ordinaires ; il se manifeste pourtant bien, à un niveau qui est le sien, et il s'avère qu'il a profondément influencé notre compréhension du fonctionnement de l'Univers. Il s'agit en fait d'un phénomène qui a été déterminant pour le progrès même de la science.

4.4 L'inclinaison des cônes de lumière

Depuis Isaac Newton, la *gravitation* et sa théorie mathématique merveilleusement précise (publiée sous sa forme achevée en 1687 par Newton lui-même) ont joué un rôle central dans le développement de la pensée scientifique. Cette théorie de la gravitation s'avéra un modèle fécond pour la description des processus physiques mettant en jeu des corps évoluant dans un espace fixe (plat) et dont les particules constitutives étaient individuellement soumises à des interactions (attractives ou répulsives) contrôlant le moindre détail de leurs

mouvements. Face aux brillants succès obtenus par la théorie newtonienne de la gravitation, on pensa que *tous* les processus physiques admettaient une description analogue en termes de forces électriques, magnétiques, moléculaires, etc., agissant entre particules.

En 1865, cette représentation se trouva radicalement modifiée lorsque l'éminent physicien écossais James Clerk Maxwell publia un remarquable ensemble d'équations décrivant le comportement précis des champs électrique et magnétique : ces champs continus se révélaient dotés d'une existence indépendante des diverses particules discrètes. Le champ électromagnétique (ainsi qu'est appelée aujourd'hui l'association de ces deux champs) véhicule de l'énergie dans l'espace vide sous forme de lumière, d'ondes radio, de rayons X, etc., et n'a pas moins de réalité que les particules avec lesquelles il est censé coexister. Toutefois, bien qu'incluant maintenant des champs continus, la description générale de l'Univers reposait encore sur des corps physiques se déplaçant dans un espace fixe sous l'influence de leurs interactions mutuelles, de sorte que le paradigme newtonien n'était pas fondamentalement altéré. Même l'étrange et révolutionnaire théorie quantique, telle qu'elle fut élaborée entre 1913 et 1926, notamment par Niels Bohr, Werner Heisenberg, Erwin Schrödinger et Paul Dirac, ne changea pas cet aspect de notre vision du monde physique. Les corps physiques étaient toujours considérés comme des entités en interaction mutuelle évoluant dans le même espace plat et fixe.

À l'époque même où il analysait quelques-uns des développements initiaux de la mécanique quantique, Albert Einstein réexaminait également en profondeur les fondements mêmes de la gravitation. En 1915, il aboutit à une théorie *nouvelle* et révolutionnaire, la théorie de la relativité générale, qui introduisit une vision de l'Univers totalement différente (*cf.* EOLP p. 218-229). La gravitation n'était plus désormais une force, mais une *courbure* affectant l'espace même (en fait, l'espace-temps) dans lequel se trouvaient toutes les autres forces et particules.

Certains physiciens n'acceptèrent pas immédiatement cette nouvelle vision, affirmant que la gravitation devait être traitée sur un pied d'égalité avec les autres actions physiques — notamment parce qu'elle avait elle-même été le paradigme initial dont toutes les théories physiques s'étaient par la suite inspirées. Une autre objection opposée à cette vision était l'extrême faiblesse de la gravitation comparée aux autres forces. Par exemple, la force gravitationnelle unissant le proton et l'électron d'un atome d'hydrogène est environ

28 500 000 000 000 000 000 000 000 000 000 000 000 000

fois plus faible que la force électrique s'exerçant entre ces deux mêmes particules. Ainsi, la gravitation n'est tout simplement pas décelable au niveau des particules constitutives de la matière !

On a également parfois suggéré que la gravitation serait un effet *résiduel* résultant d'une compensation presque complète mais pas totalement exacte des autres forces existantes. (On connaît certaines forces, telles la force de van der Waals, la liaison hydrogène et la force de London, qui possèdent cette propriété.) Ainsi, au lieu d'être un phénomène physique à part entière, admettant

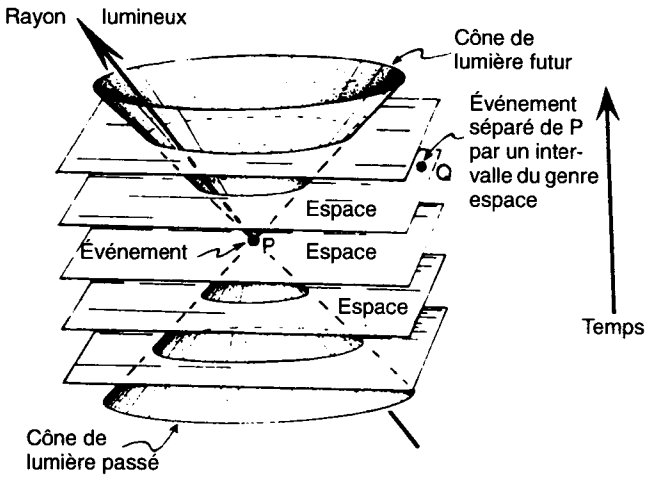


Figure 4.1. Le *cône de lumière* en un événement P se compose de tous les rayons lumineux de l'espace-temps qui passent par P. Il représente l'histoire d'un flash lumineux arrivant sur P (cône de lumière passé) et repartant à partir de P (cône de lumière futur). L'événement Q est séparé de P par un intervalle *du genre espace* (il se trouve hors du cône de lumière de P) et ne peut être influencé par P.

une description totalement différente de celle des autres forces, la gravitation n'existerait pas réellement en elle-même, mais serait une sorte de « phénomène émergent ». (C'est là une idée qu'émit notamment le grand physicien et défenseur des causes humanitaires Andreï Sakharov[4].)

Il s'avère toutefois que cette idée *ne marche pas.* La raison première en est que la gravitation influence les relations *causales* entre événements spatio-temporels et qu'elle est la *seule* quantité physique à avoir une telle influence. On peut formuler cela autrement en disant que seule la gravitation a le pouvoir d'« incliner » les cônes de lumière. (Nous verrons dans un instant ce que cela signifie.) Aucun champ physique ni aucun ensemble d'influences physiques non gravitationnelles — quelles qu'elles soient — *autres* que la gravitation ne peut incliner les cônes de lumière.

Que signifie « incliner les cônes de lumière » ? Quelles sont ces « relations causales entre événements spatio-temporels » ? La réponse à ces questions nécessite une petite digression (qui revêtira plus tard pour nous une importance particulière). Certains lecteurs sont peut-être déjà familiers avec ces notions, et je ne donnerai ici que le minimum d'explications nécessaires pour mettre les autres à niveau. (Voir EOLP, chapitre 5, p. 206 pour une discussion plus complète.) La figure 4.1 représente un cône de lumière ordinaire dans un diagramme d'espace-temps. Dans ce diagramme, le temps s'écoule du bas vers le haut de la page, tandis que l'espace s'étend horizontalement. Dans un diagramme d'espace-temps, un point représente un *événement*, c'est-à-dire un point spatial particulier situé à un instant particulier. Les événements ne

possèdent donc ni durée temporelle ni extension spatiale. Le *cône de lumière* complet centré sur un événement P représente l'histoire spatio-temporelle d'une impulsion lumineuse isotrope arrivant sur P et en repartant au même instant, toujours à la vitesse de la lumière. Ainsi, la totalité du cône de lumière de P se compose de tous les rayons lumineux qui rencontrent l'événement P au cours de leurs histoires individuelles.

Le cône de lumière de P comprend deux composantes : le cône de lumière *passé**, qui représente le flash lumineux *arrivant*, et le cône de lumière *futur*, qui représente le flash *repartant*. Selon la théorie de la relativité, tous les événements qui peuvent avoir une *influence causale* sur un événement spatio-temporel P sont ceux qui se trouvent soit à l'intérieur, soit sur le cône de lumière passé de P ; de même, tous les événements qui peuvent être influencés de façon causale *par* P sont ceux qui se trouvent soit à l'intérieur, soit sur le cône de lumière futur de P. Les événements se trouvant à l'extérieur des deux cônes de lumière passé et futur ne peuvent ni influencer P ni être influencés par lui. De tels événements sont dits séparés de P par une distance *du genre espace*.

Qu'il soit bien clair que ces relations causales particulières sont propres à la *relativité restreinte* et ne sont pas pertinentes en physique newtonienne. Le schéma newtonien n'impose aucune vitesse limite à la transmission de l'information. Cette vitesse limite n'existe qu'en théorie de la relativité et est égale à la vitesse de la lumière. Le fait qu'aucune influence causale ne puisse se propager plus vite que la vitesse de la lumière est un principe fondamental de cette théorie.

Il faut cependant bien comprendre ce que signifie ici la « vitesse de la lumière ». Les signaux lumineux réels sont légèrement ralentis lorsqu'ils traversent un milieu réfringent — du verre par exemple. Dans un tel milieu, la vitesse d'un signal lumineux physique est inférieure à ce que nous avons appelé ici la « vitesse de la lumière », et rien n'interdit qu'un corps physique ou un signal physique non lumineux puissent se déplacer à une vitesse excédant celle de la lumière dans ce milieu. Un tel phénomène s'observe expérimentalement, par exemple sous forme de « rayonnement Cerenkov » : on injecte des particules dans un milieu réfringent en leur communiquant une vitesse à peine inférieure à la « vitesse de la lumière », mais supérieure toutefois à celle de la vitesse de propagation de la lumière dans ce milieu. Il se produit alors une onde de choc lumineuse ; c'est ce qu'on appelle le rayonnement Cerenkov.

Pour éviter toute confusion, je qualifierai la « vitesse de la lumière » de vitesse *absolue*. Les cônes de lumière dans l'espace-temps déterminent la vitesse absolue, mais pas nécessairement la vitesse de la lumière réelle. Dans l'expérience de Cerenkov, la vitesse de la lumière réelle est non seulement légèrement inférieure à la vitesse absolue, mais aussi légèrement inférieure à celle des particules injectées responsables du rayonnement Cerenkov. C'est la vitesse absolue (*i.e.* chaque cône de lumière) qui fixe la vitesse limite de tous les

* Les diagrammes figurant dans EOLP montrent uniquement la composante future des cônes de lumière.

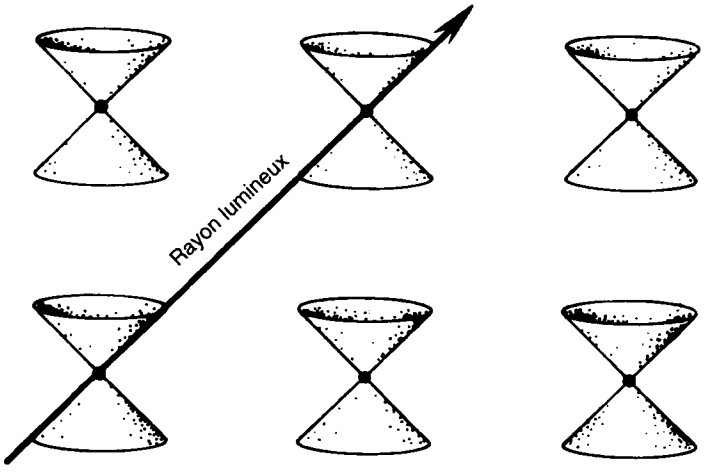


Figure 4.2. L'espace-temps de la relativité restreinte s'appelle l'espace de Minkowski. Les cônes de lumière y sont uniformément distribués.

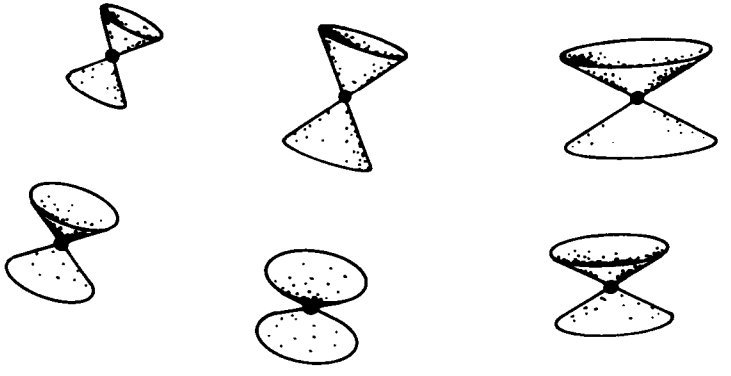


Figure 4.3. L'*inclinaison* des cônes de lumière en théorie de la relativité générale d'Einstein.

signaux et de tous les corps matériels ; c'est uniquement si la lumière se propage dans le vide que sa vitesse est égale à la vitesse absolue.

La théorie de la « relativité » à laquelle nous nous référons ici est la théorie de la relativité restreinte, qui ne prend pas en compte la gravitation. En relativité restreinte, les cônes de lumière sont tous uniformément disposés (comme le montre la figure 4.2), et l'espace-temps est appelé *espace de Minkowski.* En théorie de la relativité *générale*, la discussion précédente reste encore valide à condition de considérer que la « vitesse absolue » est encore celle qui est déterminée par la distribution spatio-temporelle des cônes de lumière. Toutefois, la gravitation a pour effet de rendre cette distribution *non* uniforme (comme le montre la figure 4.3). C'est à ce phénomène que je faisais allusion lorsque je parlais plus haut de l'« inclinaison » des cônes de lumière.

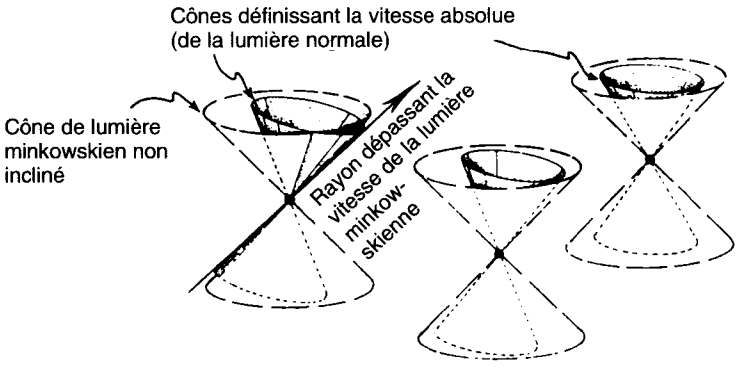


Figure 4.4. Selon la théorie de la relativité générale d'Einstein, on ne peut considérer que la propagation de la lumière résulte de l'action d'un «milieu réfringent» baignant dans l'espace-temps minkowskien sans contrevenir au principe fondamental de la relativité restreinte selon lequel un signal ne peut se propager plus vite que la vitesse de la lumière minkowskienne.

On interprète assez souvent cette inclinaison des cônes de lumière en termes de variation de la vitesse de la lumière — ou plutôt de la vitesse absolue. Cette variation a lieu de place en place et peut même dépendre de la direction du mouvement. On peut alors considérer cette « vitesse absolue » comme quelque chose d'analogue à la « vitesse de la lumière réelle » mentionnée lors de la discussion précédente sur le milieu réfringent. Le champ gravitationnel devient alors une sorte de milieu réfringent occupant tout l'espace-temps et affectant non seulement le comportement de la lumière réelle, mais aussi celui de *toutes* les particules matérielles et de *tous* les signaux*. En fait, si une telle description des effets de la gravitation s'avère souvent relativement féconde, elle n'est pas totalement satisfaisante, car elle conduit, dans certaines situations importantes, à une image nettement trompeuse de la relativité générale.

En premier lieu, si ce « milieu réfringent gravitationnel » provoque souvent un *ralentissement* de la vitesse absolue — à l'instar d'un milieu réfringent ordinaire —, il existe des contextes importants (tel celui du champ gravitationnel à très grande distance d'une masse isolée) dans lesquels ce milieu hypothétique devrait, en certains points de l'espace-temps, non pas ralentir mais accélérer la vitesse absolue (Penrose 1980 ; *cf.* Fig. 4.4). Cela *n'est pas* acceptable dans le cadre de la relativité restreinte. Selon cette théorie, aucun milieu réfringent — quelque exotiques que puissent être ses propriétés — ne peut accélérer des signaux jusqu'à une vitesse supérieure à celle de la lumière dans le vide sans violer les principes de causalité fondamentaux de cette théorie — car une telle accélération permettrait aux signaux de se propager à l'extérieur des cônes de lumière minkowskiens (vides), ce qui est interdit. En particulier, on ne peut

* Signalons que Newton lui-même suggéra une idée analogue. (Voir les *Questions* 18 à 22 du *Livre III* de son *Optique*.)

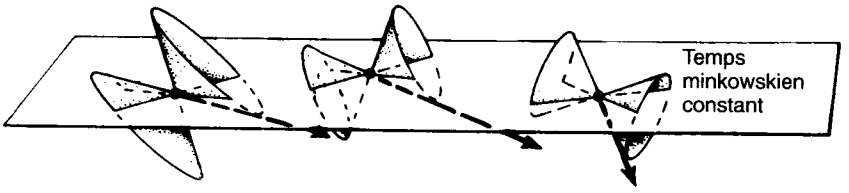


Figure 4.5. En principe, rien n'interdit aux cônes de lumière d'être inclinés de telle façon que les signaux lumineux puissent se propager en direction du passé.

interpréter l'« inclinaison gravitationnelle » des cônes de lumière mentionnée à l'instant à l'aide d'autres champs non gravitationnels.

Il existe certaines situations bien plus extrêmes dans lesquelles cette description de l'inclinaison des cônes de lumière est absolument impossible, même si l'on admet que la vitesse absolue puisse être « accélérée » dans certaines directions. La figure 4.5 montre une situation où cela est effectivement impossible : les cônes de lumière y atteignent une inclinaison qui semble absurde. En fait, cette inclinaison extrême ne survient guère que dans les situations manifestement douteuses où intervient une « violation de la causalité » — où un observateur* pourrait, théoriquement, envoyer des signaux dans son propre passé (*cf.* Fig. 7.15, chapitre 7). Curieusement, des considérations de ce type interviendront notre discussion ultérieure (§7.10) !

Il y a en outre un point plus subtil : l'« inclinaison » d'un cône de lumière n'est pas physiquement mesurable. Il en résulte qu'il n'y a physiquement aucun sens à la considérer comme un ralentissement ou une accélération *réelles* de la vitesse absolue. On comprend mieux cela si l'on imagine que la figure 4.3 a été dessinée sur une feuille élastique, de sorte que l'on peut, au voisinage de son sommet, faire tourner et déformer (*cf.* Fig. 4.6) chaque cône de lumière individuel de manière à le ramener « verticalement », comme sur les diagrammes normaux d'espace-temps minkowskien de la relativité restreinte (Fig. 4.2). Il n'existe aucun moyen de dire, à l'aide d'une expérience locale, si le cône de lumière attaché à un événement particulier est ou non « incliné ». Si l'on admet que l'inclinaison est réellement due à un « milieu gravitationnel », il faut alors expliquer pourquoi ce milieu a la très curieuse propriété qu'on ne puisse observer l'inclinaison en aucun point isolé de l'espace-temps. En particulier, même les situations apparemment extrêmes illustrées par la figure 4.5, pour lesquelles l'interprétation en termes de milieu gravitationnel ne marche pas du tout, ne sont pas, si l'on ne considère qu'un seul cône de lumière, physiquement différentes de ce qui apparaît dans une situation où, comme dans l'espace de Minkowski, ce cône de lumière n'est absolument pas incliné.

En général cependant, on ne peut redonner à un cône de lumière son orientation minkowskienne sans écarter de leur orientation minkowskienne

* Ou une observatrice ; *cf.* Note au lecteur.

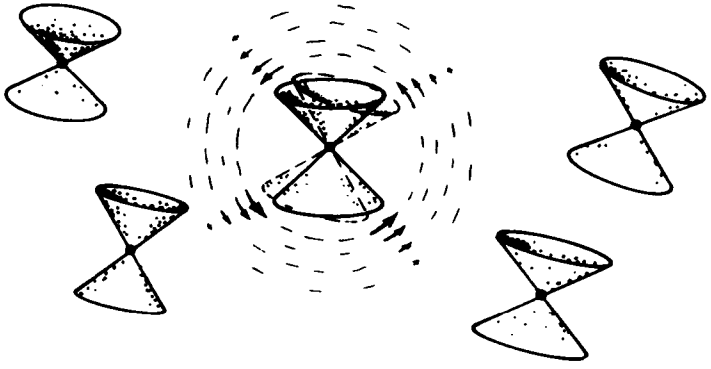


Figure 4.6. Imaginez que l'espace-temps soit une surface élastique sur laquelle les cônes de lumière seraient dessinés. On pourrait faire tourner chacun de ces cônes de lumière sur lui-même (en déplaçant sa surface environnante) de manière à le ramener dans la position minkowskienne standard.

certains autres cônes présents dans son voisinage : il y a une « obstruction mathématique » qui empêche de déformer la feuille élastique de manière à amener tous les cônes de lumière dans la position minkowskienne standard de la figure 4.2. Dans l'espace-temps quadridimensionnel, cette obstruction est décrite par un objet mathématique appelé *tenseur de Weyl* — et désigné par **WEYL** dans EOLP (*cf.* EOLP, p. 227). (Le tenseur **WEYL** décrit la moitié seulement de l'information contenue dans le tenseur de Riemann complet exprimant la courbure de l'espace-temps ; le lecteur n'a toutefois pas besoin de se soucier ici de la signification de ces termes.) C'est uniquement lorsque **WEYL** est nul que l'on peut ramener *tous* les cônes de lumière dans leur position minkowskienne. Ce tenseur **WEYL** mesure le champ gravitationnel — plus précisément, la distorsion liée à l'effet de marée gravitationnelle — de sorte que c'est précisément le *champ gravitationnel*, en ce sens, qui s'oppose au « redressement » des cônes de lumière.

Cette grandeur tensorielle est, elle, mesurable. Le champ gravitationnel **WEYL** de la Lune, par exemple, est le principal responsable des marées terrestres (*cf.* EOLP, p. 221, Fig. 5.25). Ce phénomène des marées n'est toutefois pas directement lié à l'inclinaison des cônes de lumière ; il est dû simplement à un effet gravitationnel newtonien. Il existe cependant un effet observable plus pertinent, baptisé *lentille gravitationnelle*, qui est, lui, un trait caractéristique de la théorie d'Einstein. Le premier exemple de lentille gravitationnelle, observé en 1919 lors de l'expédition conduite par Arthur Eddington sur l'île de Principe, se manifeste par un déplacement de l'image des étoiles fixes par le champ gravitationnel du Soleil. Cette action du champ gravitationnel du Soleil transforme en ellipses les motifs circulaires que dessinent les étoiles (Fig. 4.7). C'est là une observation presque directe des effets de **WEYL** sur la structure en cônes de lumière de l'espace-temps. Ces derniers temps, l'effet de lentille gravitationnelle est devenu un outil d'observation très utile en astrono-

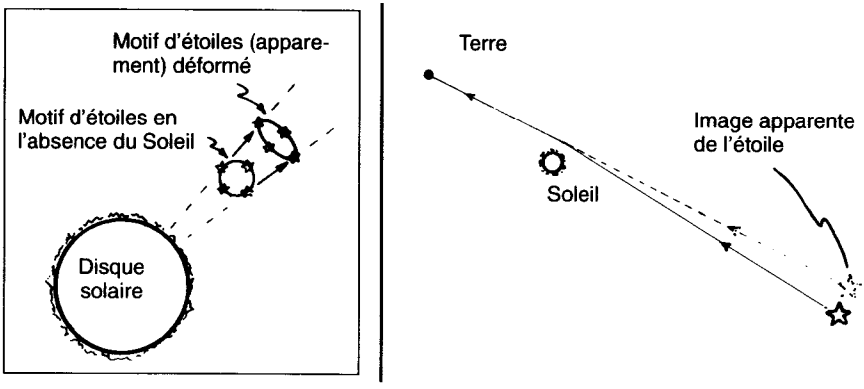


Figure 4.7. Effet observationnel direct de l'inclinaison des cônes de lumière. **WEYL**, le tenseur de courbure de l'espace-temps, se manifeste par un déplacement de l'image des étoiles fixes, dû ici à la déviation de la lumière par le champ gravitationnel du Soleil. Un motif circulaire dessiné par des étoiles se transforme en ellipse.

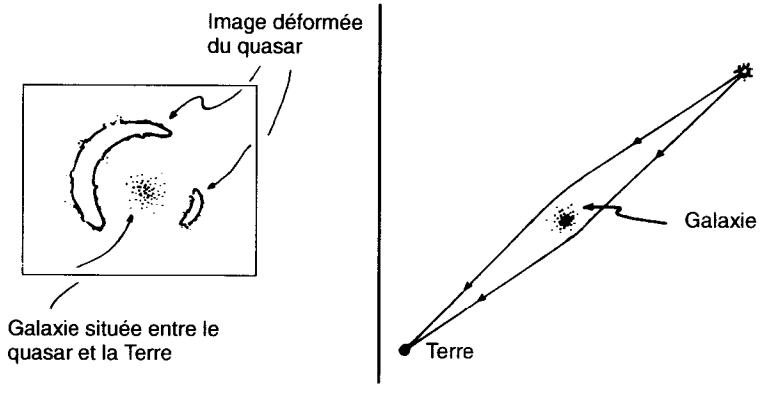


Figure 4.8. La déviation gravitationnelle de la lumière est aujourd'hui un précieux outil pour les observations astronomiques. La déformation qu'elle fait subir à l'image d'un quasar permet d'estimer la masse d'une galaxie située entre ce quasar et la Terre.

mie et en cosmologie. La lumière émise par un quasar est parfois déformée par la présence d'une galaxie entre ce quasar et la Terre (Fig. 4.8), et les distorsions de l'image du quasar que l'on observe, associées à des effets de retard, fournissent d'importantes informations sur les distances, les masses, etc. Tout cela constitue de solides témoignages en faveur non seulement de la réalité de l'inclinaison des cônes de lumière, mais aussi de la mesurabilité directe des caractéristiques de **WEYL**.

Les remarques précédentes illustrent le fait que l'« inclinaison » des cônes de lumière — *i.e.* l'altération de la causalité — due à la gravitation est un

phénomène non seulement subtil, mais aussi *réel* et qui ne peut s'expliquer par une propriété « émergente » ou résiduelle survenant lorsque des conglomérats de matière deviennent suffisamment importants. De tous les processus physiques, la gravitation est le seul qui possède *en propre* une caractéristique non directement discernable au niveau des forces entre particules, mais qui n'en est pas moins toujours présente. Aucun phénomène physique connu autre que la gravitation ne peut incliner les cônes de lumière. À cet égard, la gravitation est bel et bien *différente* de toutes les autres forces et influences physiques connues. Selon la théorie standard de la relativité générale, le plus minuscule des grains de poussière matérielle provoque une inclinaison, infime certes, des cônes de lumière. Même les électrons individuels inclinent les cônes de lumière. Mais cette inclinaison est si ridiculement faible qu'elle n'a aucun effet directement observable.

Il existe des effets gravitationnels observables entre objets qui, bien que plus gros que des grains de poussière, restent cependant considérablement plus petits que la Lune. Lors d'une expérience célèbre réalisée en 1798, Henry Cavendish mesura l'attraction gravitationnelle exercée par une sphère d'une masse égale à environ 10^5 grammes. (Cavendish reprenait là une expérience plus ancienne due à John Michell.) La technologie actuelle permet de détecter l'attraction gravitationnelle de masses très inférieures. (*e.g.* Cooke 1988.) Toutefois, dans toutes ces situations, elle reste impuissante à détecter la moindre inclinaison gravitationnelle des cônes de lumière. C'est seulement avec de très grandes masses que cette inclinaison peut être directement détectée ; mais la théorie d'Einstein affirme qu'elle est présente dans n'importe quelle quantité infinitésimale de matière.

Aucune combinaison de champs ou de forces autres que gravitationnels ne peut simuler en détail l'action de la gravitation. Cette action possède en propre un caractère unique et la gravitation est absolument irréductible à un phénomène émergent ou secondaire résultant d'autres processus physiques plus évidents. Elle est décrite par la structure même de l'espace-temps, autrefois défini comme l'arène immuable dans laquelle se déroulaient tous les autres processus physiques. Si, dans l'univers newtonien, la gravitation ne jouissait d'aucun statut particulier — bien qu'elle fût le paradigme de toutes les autres forces physiques découvertes par la suite —, elle est, dans l'univers einsteinien (magnifiquement confirmé par l'observation et adopté par les physiciens), perçue comme quelque chose de totalement différent : non pas un phénomène émergent, mais un phénomène possédant une particularité propre.

Pourtant, en dépit de sa singularité, la gravitation s'intègre harmonieusement au reste de la physique. Loin de rejeter les autres lois, la théorie d'Einstein les fait apparaître sous un jour différent. (C'est le cas en particulier des lois de conservation de l'énergie, de la quantité de mouvement et du moment cinétique.) Il y a d'ailleurs une certaine ironie à constater que la théorie de la gravitation newtonienne a fourni un *paradigme* au reste de la physique alors qu'Einstein devait ensuite montrer qu'elle est en fait *différente* du reste de la physique ! Mais la plus grande leçon que nous devons tirer de la théorie d'Einstein est que, quel que soit l'état des connaissances auquel nous sommes

parvenus, nous devons nous garder de conclure hâtivement que nous tenons la représentation définitive du monde physique.

Peut-on raisonnablement s'attendre à une découverte semblable pour le phénomène de la conscience ? Ce phénomène se manifesterait alors non pas à partir d'un certain seuil de *masse* — du moins, de masse *seule* —, mais avec l'apparition d'une forme subtile d'organisation physique. En vertu des arguments émis dans la première partie, une telle organisation devrait, de quelque façon, faire usage d'un certain processus non algorithmique dissimulé dans le comportement de la matière ordinaire — un phénomène qui, comme l'inclinaison des cônes de lumière en relativité générale, aurait totalement échappé à l'attention des physiciens si ceux-ci s'étaient exclusivement consacrés à l'étude du comportement des particules.

L'inclinaison des cônes de lumière a-t-elle toutefois un lien avec la non-calculabilité ? Nous explorerons cette fascinante question à la section 7.10, mais à ce stade de la discussion, nous pouvons répondre par un non catégorique — *sauf* que cette inclinaison nous livre cette morale : rien, en physique, n'interdit l'existence, dissimulée dans le comportement de la matière ordinaire, d'une propriété nouvelle d'une importance fondamentale importante et entièrement différente de tout ce que l'on avait jusqu'alors envisagé. Einstein adopta un point de vue révolutionnaire à partir de plusieurs considérations, certaines mathématiquement sophistiquées, d'autres physiquement subtiles ; mais la plus importante de ces considérations, bien que personne n'en perçût la pertinence, se trouvait bien en vue depuis l'époque de Galilée (c'est le principe d'équivalence, qui affirme que dans un champ gravitationnel, tous les corps tombent à la même vitesse). En outre, les idées d'Einstein n'auraient pu triompher si elles n'avaient été compatibles avec tout ce que l'on connaissait à son époque sur les phénomènes physiques.

D'une manière analogue, rien n'interdit l'existence d'une action non algorithmique dissimulée dans le comportement des corps physiques. Pour qu'une telle spéculation ait quelque chance de se voir confirmée, elle doit elle aussi se fonder sur de solides considérations, probablement mathématiquement sophistiquées et physiquement subtiles ; elle doit en outre être compatible avec tous les phénomènes physiques connus aujourd'hui. Nous allons voir jusqu'où l'on peut aller en direction d'une telle théorie.

Mais avant cela, examinons l'emprise exercée par le calcul sur la physique actuelle. Par une ironie du sort, il se trouve que la relativité générale elle-même fournit l'un des exemples les plus frappants de cette emprise.

4.5 Calcul et physique

À quelque 30 000 années-lumière de la Terre, dans la constellation de l'Aigle, deux étoiles mortes fantastiquement denses tournent l'une autour de l'autre. La pression régnant à l'intérieur de ces deux corps est si élevée qu'un

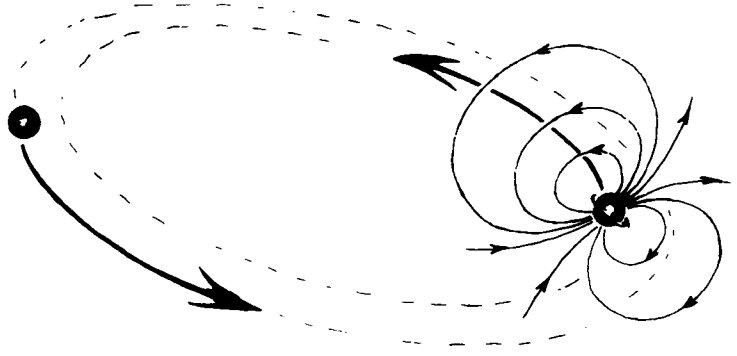


Figure 4.9. PSR 1913 + 16. Ce système se compose de deux étoiles à neutrons en orbite l'une autour de l'autre. L'une de ces étoiles est un pulsar, doté d'un champ magnétique extrêmement intense qui piège les particules chargées.

volume de leur matière équivalent à celui d'une balle de tennis aurait une masse comparable à celle de Deimos, l'un des satellites de Mars. Ces deux étoiles — appelées étoiles *à neutrons* — tournent l'une autour de l'autre en 7 heures, 45 minutes et 6,9816132 secondes, et leurs masses sont respectivement égales à 1,4411 et 1,3874 fois celle du Soleil (avec éventuellement une erreur de 7 sur la dernière décimale). Toutes les 59 millisecondes, la première de ces étoiles émet un éclat de rayonnement électromagnétique (des ondes radio) dans notre direction, ce qui signifie qu'elle accomplit 17 tours sur elle-même en une seconde ! Elle est donc ce que l'on appelle un *pulsar*, et ce couple d'étoiles constitue le célèbre pulsar binaire PSR 1913 + 16.

La découverte des pulsars, ces objets remarquables, remonte à 1967. Elle est due à Jocelyn Bell et Anthony Hewish, de l'Observatoire radioastronomique de Cambridge. Les étoiles à neutrons sont habituellement le résultat de l'effondrement gravitationnel du noyau d'une étoile géante rouge, effondrement qui peut donner lieu à une violente explosion appelée supernova. Elles sont incroyablement denses, car formées d'un amas de particules nucléaires — principalement des neutrons — tellement comprimé que leur densité globale est comparable à celle du neutron. Lors de son effondrement, une étoile à neutrons piégerait dans sa matière des lignes de champ magnétique qui, du fait de l'énorme compression entraînée par l'effondrement, se trouverait alors considérablement renforcé. Les lignes de champ magnétique émergeraient du pôle magnétique nord de l'étoile puis, après avoir atteint des distances considérables dans l'espace, rentreraient dans l'étoile par son pôle magnétique sud (Fig. 4.9).

L'effondrement de l'étoile provoquerait également un énorme accroissement de sa vitesse de rotation sur elle-même (en vertu de la conservation du moment cinétique). Nous l'avons dit, dans le cas du pulsar mentionné à l'instant (dont le diamètre est d'environ 20 kilomètres), cette vitesse est de 17 tours par seconde. Ainsi, le champ magnétique, qui est ancré au corps de l'étoile,

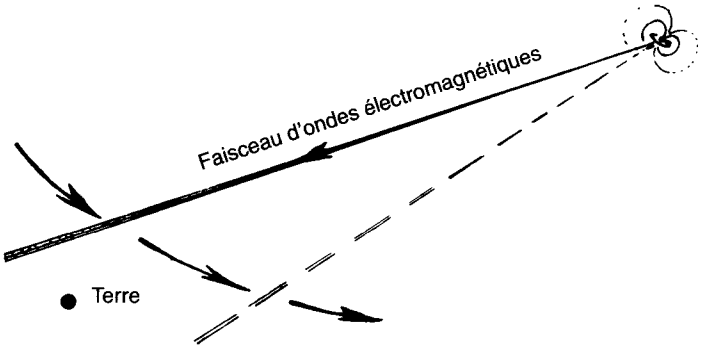


Figure 4.10. Les particules chargées piégées dans les lignes de champ magnétique du pulsar tournent avec cet objet et émettent un faisceau d'ondes électromagnétiques qui balaie la Terre 17 fois par seconde. Sur les courbes d'enregistrement, ce signal se manifeste par un pic.

tourne lui aussi sur lui-même à la vitesse de 17 tours par seconde. À l'extérieur de l'étoile, les lignes de champ guident le déplacement de particules chargées qui, à partir d'une certaine distance, atteignent des vitesses très proches de celle de la lumière. Elles émettent alors d'intenses ondes radio qui, collimatées sur d'énormes distances à la manière d'un gigantesque faisceau de phare, balaient l'espace intersidéral. Si ce faisceau a une orientation telle qu'il « éclaire » la Terre, ses passages successifs apparaissent alors aux astronomes comme une suite de « bips » radio caractéristique d'un pulsar (Fig. 4.10).

Les vitesses de rotation des pulsars sont extrêmement stables, et ces corps constituent des horloges dont la précision égale, voire dépasse, celle des plus parfaites horloges (nucléaires) construites sur Terre. (Une bonne horloge-pulsar peut retarder ou avancer de moins de 10^{-12} seconde par an.) Si le pulsar appartient à un système binaire, comme dans le cas de PSR 1913 + 16, on peut alors étudier son mouvement orbital autour de son compagnon à l'aide de l'*effet Doppler* : la fréquence de ses « bips », telle qu'on la mesure sur Terre, est légèrement plus grande lorsqu'il se dirige vers nous que lorsqu'il s'éloigne.

Avec PSR 1913 + 16, les astronomes ont pu obtenir une image extraordinairement précise des orbites que ces deux étoiles décrivent l'une autour de l'autre et vérifier par l'observation plusieurs prédictions de la relativité générale d'Einstein. Parmi ces prédictions figurent un effet baptisé « avance du périhélie » — une anomalie qui, vers la fin du XIX^{e} siècle, avait été remarquée pour l'orbite de la planète Mercure autour du Soleil et dont l'interprétation en 1916 par Einstein constitua la première confirmation observationnelle de la relativité générale — et divers types de « tremblements » affectant les axes de rotation, etc. La théorie d'Einstein donne une description claire (déterministe et calculable) du mouvement de deux petits corps en orbite l'un autour de l'autre et permet de calculer ce mouvement avec une très grande précision en utilisant des méthodes d'approximation précises et sophistiquées ainsi que

diverses techniques numériques standard. Si un tel calcul met en jeu certains paramètres inconnus, tels les masses et les mouvements initiaux des étoiles, les données recueillies lors de l'observation des signaux émis par les pulsars permettent de les fixer avec une excellente précision. L'accord global entre le comportement calculé et ces données observationnelles est tout à fait remarquable et conforte solidement la relativité générale.

Un autre effet, prédit lui aussi par la relativité générale, joue un rôle important dans la dynamique des pulsars binaires : c'est le *rayonnement gravitationnel.* À la section précédente, nous avons vu que la gravitation se distingue radicalement de tous les autres champs physiques. Mais à certains égards, la gravitation et l'électromagnétisme sont très semblables. L'une des propriétés fondamentales des champs électromagnétiques est de pouvoir exister sous forme d'ondes — par exemple, lumineuses ou radio — se propageant dans l'espace. Selon la théorie classique du champ électromagnétique (la théorie de Maxwell), ces ondes sont notamment émises par tout système de particules chargées orbitant les unes autour des autres sous l'effet de leur interaction électromagnétique. De même, selon la relativité générale standard, tout système de corps en orbite les uns autour des autres sous l'effet de leur interaction gravitationnelle devrait émettre des ondes gravitationnelles. Dans les situations ordinaires, ces ondes seraient extrêmement faibles. La plus puissante source de rayonnement gravitationnel du système solaire est fournie par le mouvement de la planète Jupiter autour du Soleil, mais l'énergie émise sous cette forme par le système Jupiter-Soleil permettrait seulement d'alimenter une ampoule de 40 watts !

Toutefois, dans d'autres circonstances, telle celle correspondant à PSR 1913 + 16, la situation est très différente. Le rayonnement gravitationnel émis par ce système est bien plus important. Ici, la théorie d'Einstein prédit la nature détaillée du rayonnement gravitationnel que ce système devrait émettre, ainsi que celle de l'énergie qui devrait s'en échapper. Cette perte d'énergie devrait entraîner un lent mouvement spiralé rapprochant les deux étoiles à neutrons et provoquer ainsi un accroissement correspondant de leur période de rotation sur leur orbite. Joseph Taylor et Russell Hulse observèrent pour la première fois ce pulsar binaire en 1974, sur le gigantesque radiotélescope d'Arecibo, à Porto Rico. Cette période a été depuis minutieusement surveillée par Taylor et ses collègues, et son accroissement est en accord précis avec les prédictions de la relativité générale (Fig. 4.11). Ce travail a valu à Hulse et Taylor le prix Nobel de physique 1993. Les données accumulées sur ce système au fil des ans n'ont cessé de confirmer la théorie d'Einstein. Si aujourd'hui on considère ce système et si l'on compare son comportement à celui déduit de l'ensemble de la théorie d'Einstein — depuis les aspects newtoniens des orbites jusqu'aux corrections dues à la relativité générale, voire même jusqu'aux corrections orbitales dues à la perte d'énergie par rayonnement gravitationnel —, on constate que cette théorie est globalement confirmée avec une erreur inférieure à environ 10^{-14}. De ce point de vue, la relativité générale d'Einstein est la théorie scientifique la mieux confirmée par l'observation !

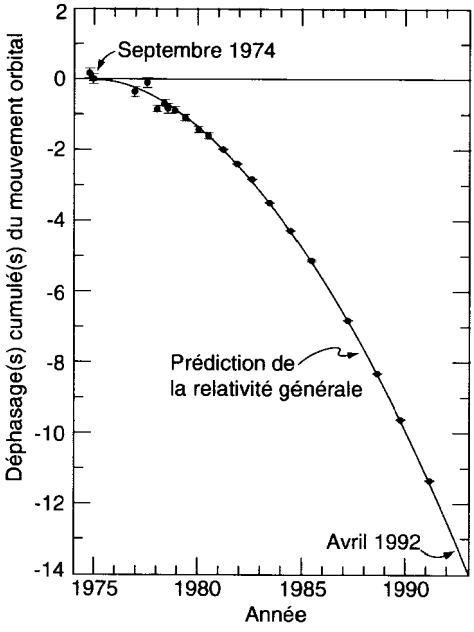


Figure 4.11. Ce graphe (aimablement communiqué par Joseph Taylor) montre l'accord précis, sur une durée de 20 années, entre l'accélération de la période de rotation du pulsar sur son orbite et la perte d'énergie par rayonnement gravitationnel calculée à l'aide de la théorie d'Einstein.

Cet exemple correspond à un système bien « propre », dont le comportement se calcule à l'aide de la seule relativité générale. Les complications résultant de la prise en compte de contraintes telles que la constitution interne des corps, ou du freinage dû au milieu ambiant ou à des champs magnétiques, n'affectent pas les mouvements de manière significative. En outre, deux corps seulement sont en présence, soumis uniquement à leur interaction gravitationnelle, de sorte que l'on peut très facilement, à l'aide de la théorie, calculer totalement et en détail leur comportement. C'est peut-être là — pour le cas d'un petit nombre de corps — l'exemple le plus parfait en science d'un accord entre un modèle numérique déduit d'une théorie et un comportement observé.

Lorsqu'un système physique comprend un nombre de corps bien plus élevé, il est parfois encore possible, en utilisant toutes les ressources de la technologie informatique moderne, de modéliser son comportement de manière aussi détaillée. C'est ainsi qu'Irwin Shapiro et ses collègues sont parvenus à simuler numériquement et en détail le mouvement de toutes les planètes du système solaire, avec ses lunes les plus importantes. Ce travail a fourni une autre confirmation de la relativité générale. Ici encore, la théorie d'Einstein concorde avec toutes les données observationnelles et rend compte des divers petits écarts que l'on constaterait si l'on avait recouru à un traitement purement newtonien.

Les calculs mettant en jeu un nombre de corps encore plus grand — parfois de l'ordre d'un million — peuvent également être effectués sur des ordinateurs modernes, bien que d'une manière générale (mais pas toujours) ils reposent entièrement sur la théorie newtonienne. Ils exigent l'introduction de certaines hypothèses simplificatrices permettant l'approximation de l'action de nombreuses particules par un certain type de moyenne et évitant de devoir calculer l'action de chaque particule sur toutes les autres. De tels calculs sont fréquents en astrophysique lorsqu'on étudie par exemple le processus de formation des étoiles ou des galaxies, ou l'agglomération de la matière dans l'Univers primitif, avant la formation des galaxies.

Au niveau de leurs objectifs, ces calculs présentent toutefois une importante différence. Ils ne risquent guère en effet de fournir l'évolution *réelle* d'un système, mais plutôt d'en donner une évolution *typique*. Comme pour les systèmes chaotiques, c'est probablement le mieux que nous puissions obtenir. Cela permet cependant de tester diverses hypothèses sur la composition et la distribution initiale de la matière dans l'Univers : on voit si, globalement, l'évolution résultante concorde avec les observations. Dans un tel contexte, il est illusoire d'espérer parvenir à un accord détaillé ; on se contente de vérifier si l'aspect général de l'ensemble est satisfaisant et de contrôler tel ou tel paramètre utlisé dans le modèle.

Cette situation est poussée à l'extrême lorsque les particules sont si nombreuses que l'on ne peut espérer suivre l'évolution de chacune d'elles. On est alors contraint de traiter ces particules par des méthodes entièrement statistiques. Habituellement, le traitement mathématique d'un gaz, par exemple, consiste à étudier des *ensembles* statistiques composés des divers mouvements collectifs possibles pour les particules, sans se préoccuper des mouvements individuels. Mais dans ce traitement statistique, les grandeurs physiques — la température, la pression, l'entropie, etc. — qui caractérisent ces ensembles se déterminent ici encore dans le cadre d'un système calculable.

Outre les équations dynamiques pertinentes (celles de Newton, Maxwell, Einstein, etc.), ce traitement statistique fait intervenir un autre principe physique, le *deuxième principe de la thermodynamique*[5]. Ce principe sert à exclure les états initiaux des mouvements de particules individuelles conduisant à des évolutions futures qui, bien que dynamiquement admissibles, sont hautement improbables. L'introduction de ce deuxième principe garantit donc que l'évolution future du système modélisé est effectivement « typique » et non grossièrement *a*typique et sans aucun rapport concret avec le problème considéré. Grâce à ce principe, on peut calculer l'évolution future de systèmes comportant un nombre de particules si élevé qu'un traitement détaillé des mouvements individuels serait, en pratique, irréalisable.

On peut se demander — et c'est là une question profonde — pourquoi, bien que les équations de Newton, Maxwell et Einstein soient toutes totalement symétriques par rapport au temps, le calcul d'une telle évolution en direction du *passé* ne présente aucune fiabilité. Pourquoi, dans le monde réel, le deuxième principe de la thermodynamique ne s'applique-t-il pas pour un temps qui s'écoulerait en sens inverse du temps normal ? La raison fondamen-

tale de ce fait est liée aux conditions très particulières qui prévalurent au commencement du temps — lors du *Big Bang* qui fut à l'origine de l'Univers. (Voir EOLP, chapitre 7, pour une discussion sur ce sujet.) En fait, ces conditions initiales furent si particulières qu'elles fournissent un autre exemple de l'extraordinaire précision avec laquelle des hypothèses mathématiques bien définies peuvent modéliser le comportement physique observé.

Dans le cas du Big Bang, l'une des hypothèses essentielles est que lors de ses tout premiers instants, le contenu matériel de l'Univers était dans un état d'*équilibre thermique*. Que signifie « équilibre thermique » ? L'étude des états à l'équilibre thermique représente l'extrême opposé de la modélisation précise des mouvements détaillés d'un très petit nombre d'objets — par exemple, dans le cas du pulsar binaire mentionné plus haut — et se concentre uniquement sur la recherche d'un « comportement typique » dans sa forme la plus pure et la plus fiable. D'une manière générale, un état d'équilibre est un état dans lequel un système s'est complètement « installé » et dont il ne dévie pratiquement pas, même si on le perturbe légèrement. Pour un système constitué d'un grand nombre de particules (ou comprenant un grand nombre de degrés de liberté) — de sorte que l'on a accès non aux mouvements particulaires individuels, mais au comportement moyen et à des mesures moyennes telles la température et la pression —, l'équilibre *thermique* est l'état vers lequel le système finit par aboutir en vertu du deuxième principe de la thermodynamique (maximum d'entropie). Le qualificatif « thermique » implique une sorte de moyenne sur la totalité des mouvements particulaires individuels possibles. La thermodynamique est l'étude de ces moyennes — *i.e.* des comportements typiques plutôt qu'individuels.

Strictement parlant, et conformément à ce qui a été dit plus haut, lorsqu'on parle de l'état thermodynamique d'un système ou de l'état d'équilibre, on se réfère non à un état individuel, mais à un ensemble d'états qui ont tous même apparence à l'échelle macroscopique (et l'entropie, grossièrement parlant, est le logarithme du nombre d'états de cet ensemble). Dans le cas d'un gaz, si l'on fixe la pression, le volume et le nombre des divers types de particules composant le gaz, on obtient, à l'équilibre thermique, une distribution des vitesses de particules très caractéristique (cette distribution fut découverte par Maxwell). Une analyse plus fine révèle l'existence d'une échelle à laquelle apparaissent des fluctuations statistiques par rapport à l'état d'équilibre thermique idéal — on pénètre alors dans le domaine plus sophistiqué de l'étude du comportement statistique de la matière, étude qui porte le nom de *mécanique statistique*.

Ici aussi, il semble que la modélisation du comportement physique par des structures mathématiques ne contienne rien d'essentiellement non calculable. Une fois achevés les calculs appropriés, on obtient un bon accord entre ce que l'on a calculé et ce qui est observé. Néanmoins, lorsqu'on envisage des systèmes plus complexes que des gaz dilués ou de grands ensembles de corps en interaction gravitationnelle, on ne peut généralement éviter les problèmes soulevés par la nature *quantique* des matériaux considérés. En particulier, l'exemple de comportement thermodynamique qui est à la fois le plus pur et

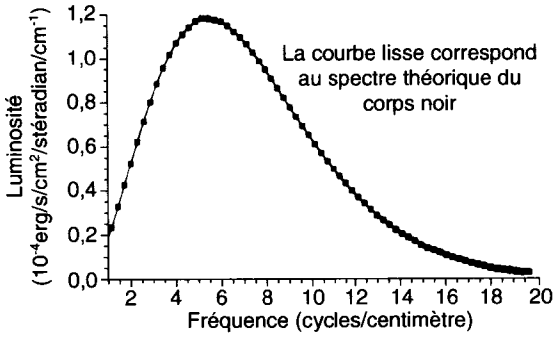


Figure 4.12. Ce diagramme montre l'accord précis entre les mesures réalisées par COBE et l'hypothèse de la nature « thermique » du rayonnement du Big Bang.

celui qui a été confirmé avec le plus de précision — l'équilibre thermique entre la matière et le rayonnement connu sous le nom d'état de *corps noir* — n'admet aucun traitement entièrement classique et met en jeu des processus quantiques. Ce fut d'ailleurs l'analyse du rayonnement du corps noir effectuée par Max Planck en 1900 qui inaugura la théorie quantique.

Toutefois, les prédictions de la physique (aujourd'hui quantique) se trouvent triomphalement vérifiées. La relation observée entre la fréquence et l'intensité du rayonnement associé à cette fréquence concorde très étroitement avec la formule mathématique découverte par Planck. Bien que cette section soit essentiellement consacrée aux relations entre calcul et physique *classique*, je ne peux résister à la tentation de vous montrer ce qui est de loin l'exemple le plus parfait que je connaisse d'un accord entre l'observation et la formule de Planck. Cet exemple est également une merveilleuse confirmation observationnelle du modèle standard du Big Bang en ce qui concerne les conditions thermiques de l'Univers après les toutes premières minutes de son existence. Sur la figure 4.12, les petits carrés indiquent les diverses valeurs observées par le satellite COBE pour l'intensité du rayonnement du fond du ciel à diverses fréquences ; la courbe continue est le graphe correspondant à la formule de Planck pour une température de rayonnement de 2,735 (±0,06) K (obtenue par interpolation). La précision de l'accord est extraordinaire.

Les exemples que j'ai donnés ici étaient empruntés à l'astrophysique, discipline dans laquelle l'accord entre des calculs complexes et l'observation du comportement de systèmes existant dans le monde naturel est particulièrement manifeste. L'astrophysique ne pouvant procéder à des expériences directes, elle en est réduite à tester ses théories en comparant à des observations précises les résultats de calculs détaillés reposant sur des lois physiques standard. (Ces observations peuvent être menées depuis la Terre, ou à partir de ballons ou d'avions se déplaçant dans la haute atmosphère, ou encore à partir de fusées ou de satellites ; outre les télescopes ordinaires, elles recourent à divers types de détecteurs.) Ces calculs n'ont pas vraiment de rapport avec ceux qui nous

concernent ici ; je les ai surtout mentionnés parce qu'ils fournissent des exemples particulièrement éloquents de la pertinence des procédures numériques pour l'exploration, voire la simulation, de la nature. L'étude des systèmes biologiques devrait en revanche nous intéresser plus directement. Car en vertu des conclusions de la première partie, c'est davantage dans le comportement du cerveau conscient que nous devrions rechercher une action physique non calculable.

S'il est évident que les modèles numériques jouent un rôle important dans la modélisation des systèmes biologiques, ces systèmes risquent cependant d'être bien plus complexes que ceux rencontrés en astrophysique, et les modèles numériques correspondants moins fiables. Il existe très peu de systèmes suffisamment « propres » pour se prêter à une simulation d'une grande précision. Des systèmes relativement simples admettent cependant des simulations tout à fait honorables ; c'est par exemple le cas pour l'écoulement sanguin dans divers types de vaisseaux ou la transmission des signaux le long des fibres nerveuses — bien que pour cette dernière, du fait de l'importance des actions chimiques concurremment aux actions physiques (classiques), il apparaisse de plus en plus que le problème ne puisse se traiter dans un cadre purement classique.

Les actions chimiques résultent d'effets quantiques, et strictement parlant, lorsque l'on considère des processus dépendant de la chimie, on quitte le domaine de la physique classique. Néanmoins, ces actions d'origine quantique sont très fréquemment traitées en termes essentiellement classiques. Cela n'est pas techniquement correct, mais on a le sentiment que dans la plupart des cas, les effets plus subtils de la théorie quantique — outre ceux que l'on peut réduire aux règles standard de la chimie, de la physique classique et de la géométrie — ne jouent pas un rôle déterminant. Bien que cette option s'avère efficace pour la modélisation de nombreux systèmes biologiques (peut-être même pour celle de la propagation de l'influx nerveux), il m'apparaît toutefois risqué de tirer des conclusions générales sur les plus subtiles des actions biologiques en supposant qu'elles sont de nature entièrement classique, particulièrement lorsqu'on a affaire au plus sophistiqué des systèmes biologiques, le cerveau humain. Si l'on désire tirer des conclusions générales sur la possibilité théorique d'une simulation numérique fiable du cerveau, on doit tenir compte des mystères de la théorie quantique.

C'est justement ce que je vais tenter de faire dans les deux chapitres qui suivent — du moins dans la mesure du possible. Lorsque cela se révélera, d'un point de vue de principe, *impossible*, je montrerai alors que la seule issue consiste à modifier cette théorie afin de l'intégrer plus convenablement dans une représentation crédible du monde.

5
La structure du monde quantique

5.1. *La théorie quantique : mystères et paradoxes*

Si la théorie quantique est une superbe description de la réalité physique microscopique, elle renferme cependant de nombreuses énigmes. Il n'est vraiment pas facile d'accepter ses mécanismes et en particulier de se pénétrer de la « réalité physique » — ou de l'absence de réalité physique — qu'elle semble attribuer à notre monde. Prise au pied de la lettre, cette théorie semble conduire à un point de vue philosophique que nombre de personnes (dont moi) jugent profondément insatisfaisant. Au mieux, et si l'on prend ses descriptions dans leur sens le plus prosaïque, elle nous donne une vision du monde effectivement très étrange. Au pire, et en prenant à la lettre les proclamations de certains de ses plus célèbres promoteurs, elle n'offre en fait aucune vision du monde.

Selon moi, cette théorie nous place devant deux types d'énigmes totalement différents. Il y a les énigmes que j'appellerai « énigmes-**Y** », ou énigmes-*mystères*, qui sont des vérités quantiques — absolument déconcertantes mais directement confortées par l'expérience — sur le monde dans lequel nous vivons. Ce type d'énigmes englobe en outre des phénomènes dont l'existence, bien que non encore expérimentalement vérifiée, ne laisse guère de doute eu égard à ce qui a déjà été établi, mais sur laquelle on ne pourra cependant définitivement se prononcer qu'une fois confirmé l'accord de ses conséquences observables avec la théorie quantique. Les plus frappantes de ces énigmes-**Y** sont notamment les phénomènes de type *Einstein-Podolsky-Rosen* (ou EPR) que j'examinerai plus loin (§5.4, §6.5). L'autre type d'énigmes quantiques

correspond à ce que j'appellerai les « énigmes-**X** », ou énigmes-*paradoxes*, dont le formalisme quantique semble nous dire qu'elles sont vraies, mais qui ont un caractère si paradoxal que l'on ne peut croire qu'elles soient « réellement » vraies, quel que soit le sens que l'on accorde à cet adjectif. Ces énigmes nous empêchent d'envisager sérieusement que le formalisme quantique — au niveau où elles se manifestent — puisse nous offrir une image crédible du monde. La plus célèbre des énigmes-**X** est le paradoxe du *chat de Schrödinger* : le formalisme de la théorie quantique semble nous dire que les objets macroscopiques, tels les chats, peuvent exister simultanément dans deux états totalement différents — par exemple en une combinaison de « chat mort » et de « chat vivant ». (Je discuterai ce type d'énigme à la section 6.6 ; *cf.* §6.9 ; Fig. 6.3, et EOLP p. 316-319.)

Certains affirment que les difficultés éprouvées par les générations actuelles avec la théorie quantique ne sont que la conséquence d'un trop fort ancrage de notre mode de pensée dans les concepts physiques du passé. Ainsi, chaque nouvelle génération s'habituerait davantage à cette théorie, de sorte qu'on devrait un jour aboutir à une société qui accepterait toutes ces énigmes — qu'elles soient **X** ou **Y** — sans difficulté. Mon propre point de vue est radicalement différent.

Si je pense que nous pourrions effectivement nous habituer aux énigmes-**Y** et les percevoir un jour comme évidentes, il en va différemment pour les énigmes-**X**. Selon moi, ces dernières sont philosophiquement inacceptables et résultent uniquement de l'incomplétude actuelle de la théorie quantique — plus exactement, de son actuelle imprécision au niveau des phénomènes générant ces énigmes-**X**. J'ai le sentiment que dans le cadre d'une théorie quantique améliorée, les énigmes-**X** disparaîtraient purement et simplement (seraient *rayées*) de la liste des énigmes quantiques. Ce sont seulement les énigmes-**Y** avec lesquelles nous devrons apprendre à vivre tranquillement !

On peut cependant se demander où se situe la frontière entre énigmes-**Y** et énigmes-**X**. Certains physiciens affirment que ce que j'appelle énigmes-**X** n'existe pas et que *tous* les phénomènes étranges et apparemment paradoxaux auxquels le formalisme quantique nous demande de croire sont effectivement vrais pour peu qu'on appréhende correctement ce formalisme. (S'ils sont entièrement logiques et s'ils prennent réellement au sérieux la description de la réalité physique en termes d'« états quantiques », ces physiciens devraient croire en l'existence d'un certain type de « mondes multiples » — que je décrirai à la section 6.2. Selon ce point de vue, le chat mort et le chat vivant de Schrödinger habiteraient deux univers parallèles. Lorsque vous décidez de vous intéresser au sort du chat, vous faites naître deux exemplaires de votre personne, présents respectivement dans chacun de ces deux univers, l'un voyant le chat mort, l'autre le chat vivant.) D'autres physiciens adopteraient un point de vue diamétralement opposé et affirmeraient que je suis trop indulgent à l'égard du formalisme quantique en prévoyant que toutes les énigmes de type EPR seront en fait confortées par des expériences à venir. Je ne demande pas à tout le monde d'être d'accord avec moi sur le tracé de la frontière entre les deux types d'énigmes. Je l'ai défini en fonction de ce qui,

selon moi, est compatible avec le point de vue que je développerai à la section 6.12.

Je ne vais bien sûr pas faire ici un exposé intégral de la théorie quantique. Je me bornerai simplement, dans ce chapitre, à une description relativement brève et suffisamment complète de celles de ses caractéristiques qui sont pertinentes pour notre propos, en me concentrant principalement sur la nature de ses énigmes-**Y**. Dans le prochain chapitre, je donnerai les raisons qui me portent à penser que l'existence des énigmes-**X** traduit l'incomplétude de la théorie quantique actuelle — même si cette théorie est merveilleusement en accord avec toutes les expériences réalisées à ce jour. Les lecteurs qui désirent approfondir leurs connaissances en théorie quantique pourront lire le chapitre 6 de EOLP, ou bien, par exemple, Dirac (1947) ou Davies (1984).

À la section 6.12, je présenterai une idée récente selon laquelle l'introduction de modifications à un niveau précis de la théorie quantique devrait permettre de remédier à son incomplétude (je préviens toutefois le lecteur que si ses motivations sont très semblables, cette idée diffère passablement de celle que j'ai donnée dans EOLP). Puis, aux sections 7.8 et 7.10, je donnerai quelques raisons, à mon avis convaincantes, qui laissent penser que ces modifications seraient non algorithmiques dans le sens que nous recherchons. Si la théorie quantique *standard*, quant à elle, est non algorithmique, cela vient uniquement de la présence d'éléments aléatoires dans les mesures. Or, ainsi que je l'ai souligné dans la première partie (§3.18, §3.19), les processus aléatoires ne peuvent, à eux seuls, fournir la non-calculabilité dont nous aurons en définitive besoin pour comprendre l'activité mentale.

Commençons donc avec deux des plus surprenantes énigmes-**Y** de la théorie quantique. Je vais les présenter sous forme de problèmes que je soumettrai dans un premier temps à la sagacité du lecteur.

5.2 Le problème d'Elitzur-Vaidman

Imaginez que l'on ait conçu des bombes munies, au bout de leur nez, d'un détonateur si sensible qu'elles explosent au moindre contact. Un seul photon de lumière visible suffit à les faire exploser. Cependant certains détonateurs sont grippés — la bombe n'explose pas et est considérée comme « ratée ». Supposons que chaque détonateur se compose d'un miroir fixé au nez de la bombe, de sorte que l'impact d'un seul photon (de lumière visible) réfléchi par ce miroir suffit pour actionner un piston dans la bombe et provoquer l'explosion — sauf si la bombe est ratée, son détonateur étant grippé. Supposons également qu'aucun dispositif conçu à l'aide de la physique classique ne permette de déterminer, une fois une bombe assemblée, si son détonateur est ou non grippé sans que l'on soit obligé, d'une manière ou d'une autre, de titiller ce détonateur — ce qui risquerait de faire exploser la bombe. (Fig. 5.1.) (Nous

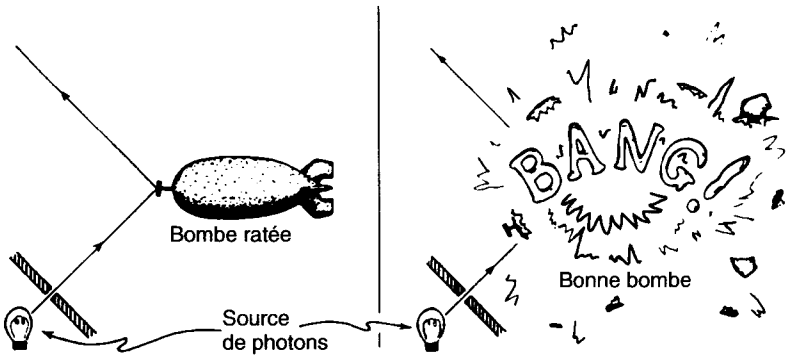


Figure 5.1. Le problème d'Elitzur-Vaidman. Le détonateur ultrasensible de la bombe réagit à l'impact d'un seul photon de lumière visible — si ce détonateur n'est pas grippé et donc que la bombe n'est pas une « ratée ». Problème : étant donné un grand stock de bombes douteuses, comment déterminer qu'une bombe n'est pas une ratée ?

admettrons que l'éventuel grippage du détonateur résulte d'une mauvaise manœuvre lors de la phase de montage de la bombe.)

Imaginons maintenant que nous disposions d'un grand stock de ces bombes (nous sommes suffisamment riches !), mais que le pourcentage des ratées soit très élevé. Le problème est de trouver une bombe dont on est certain qu'elle n'est pas une ratée.

Ce problème (et sa solution) a été proposé par Avshalom Elitzur et Lev Vaidman en 1993. Je remets à plus tard l'exposé de sa solution pour permettre aux lecteurs déjà familiers de la théorie quantique et de ses énigmes-**Y** de la trouver par eux-mêmes. Disons simplement qu'*il y a* une solution et que pour une quantité illimitée de telles bombes, elle serait à la portée de la technologie actuelle. Aux lecteurs qui ne sont pas déjà versés dans la théorie quantique — ou qui le sont mais ne veulent pas perdre leur temps à chercher cette solution —, je demande un peu de patience (ou d'aller directement à la section 5.9). Je donnerai cette solution en temps voulu, une fois introduites les idées quantiques fondamentales nécessaires à son exposé.

Soulignons cependant dès maintenant que le fait que ce problème *possède* une solution (*quantique*) signale déjà l'existence d'une profonde différence entre les physiques classique et quantique. D'un point de vue classique, et tel qu'est posé le problème, le seul moyen de voir si le détonateur est grippé consiste à le titiller — et s'il n'est pas grippé, la bombe explose. La théorie quantique autorise une autre possibilité, fondée sur un effet physique résultant d'un contact non pas *réel* mais *virtuel* du détonateur. Cette théorie a ceci de particulièrement curieux que des effets physiques réels peuvent avoir leur origine dans ce que les philosophes appellent des *contrafactuels* — c'est-à-dire des choses qui auraient pu se produire, mais qui ne se sont pas produites. Notre prochaine énigme-**Y** va nous montrer que le problème des contrafactuels se manifeste également dans un autre type de situation.

5.3 Les dodécaèdres magiques

Permettez-moi de raconter une petite histoire — qui est aussi un mystère[1]. Imaginez, j'ai reçu récemment un superbe dodécaèdre régulier (Fig. 5.2). Il m'a été envoyé par une entreprise très sérieuse, appelée *Trucs Quintessentiels*, implantée sur une planète orbitant autour de la lointaine géante rouge Bételgeuse. Cette entreprise a envoyé un dodécaèdre identique à l'un de mes collègues vivant sur une planète orbitant autour de l'étoile α du Centaure, située à environ quatre années-lumière de la Terre — ce collègue et moi-même avons reçu nos colis pratiquement en même temps. Ces deux dodécaèdres possèdent en chacun de leurs sommets un bouton-pressoir. Indépendamment l'un de l'autre, mon collègue et moi allons presser ces boutons, un à la fois, à n'importe quel instant et dans l'ordre que nous désirons. Lorsque nous pressons un bouton, soit rien ne se produit, auquel cas nous choisissons d'en presser un autre, soit une sonnette retentit et déclenche un magnifique feu d'artifice qui détruit le dodécaèdre.

Chaque dodécaèdre est livré avec une liste de propriétés certifiées qui expliquent ce qui peut lui arriver. Tout d'abord, mon collègue et moi devons soigneusement donner à nos dodécaèdres respectifs une orientation identique. *Trucs Quintessentiels* a fourni des instructions détaillées sur la direction dans laquelle nous devons aligner nos dodécaèdres par rapport, par exemple, au centre de la galaxie d'Andromède, à celui de la galaxie M87, etc. L'important est que mon dodécaèdre et celui de mon collègue soient parfaitement alignés l'un avec l'autre. La liste des propriétés certifiées peut être relativement longue, mais seules deux d'entre elles nous intéresseront.

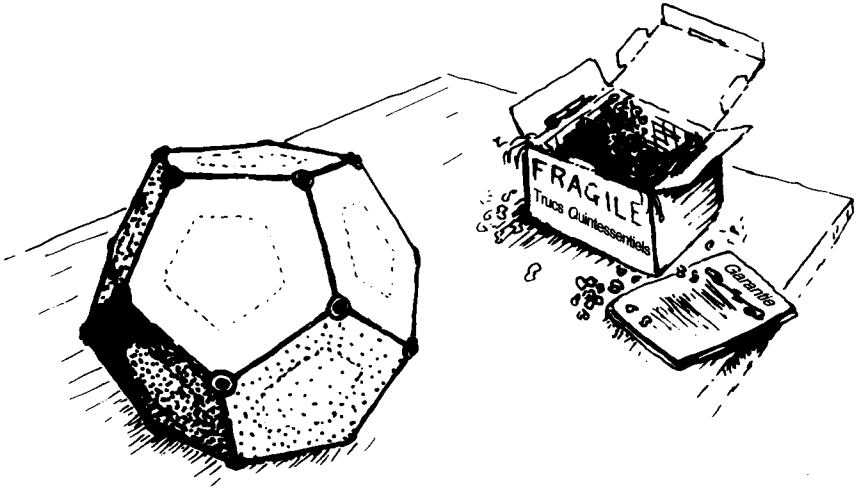


Figure 5.2. Le dodécaèdre magique. Mon collègue sur α du Centaure en possède un exemplaire identique. Sur chacun des sommets du dodécaèdre se trouve un bouton qui, s'il est pressé, *peut* faire retentir une sonnette et déclencher alors un superbe feu d'artifice.

Mon collègue et moi savons que la société *Trucs Quintessentiels* produit cette sorte d'objets depuis très longtemps — disons depuis une centaine de millions d'années — et que les propriétés qu'elle garantit n'ont jamais été prises en défaut. La solide réputation acquise par cette compagnie au cours de centaines de siècles repose sur le sérieux avec lequel elle a conçu ces propriétés, et nous pouvons être tout à fait certains que tout ce qu'elle affirme est effectivement vrai. En outre, elle offre une somme d'argent colossale à quiconque constaterait que ces propriétés sont erronées — personne n'a jamais réclamé l'argent !

Les propriétés garanties auxquelles nous nous intéressons concernent les pressions que nous exerçons sur les boutons. Mon collègue et moi choisissons indépendamment l'un des sommets de nos dodécaèdres respectifs. J'appellerai un tel sommet un sommet SÉLECTIONNÉ. Les boutons que nous pressons sont *non pas* ceux du sommet SÉLECTIONNÉ mais, et dans un ordre arbitrairement choisi par nous, chacun des boutons situés sur les trois sommets *adjacents* à ce sommet SÉLECTIONNÉ. Si la sonnette retentit pour l'un d'entre eux, cela stoppe la procédure sur le dodécaèdre considéré, mais rien n'oblige la sonnette à retentir. Nous savons seulement que les deux propriétés suivantes sont satisfaites (Fig. 5.3) :

(a) si par hasard les sommets SÉLECTIONNÉS par mon collègue et moi se trouvent diamétralement *opposés*, la sonnette retentit sur l'un des boutons pressés par moi (et adjacents à mes sommets SÉLECTIONNÉS) si et seulement si la sonnette retentit sur le sommet diamétralement opposé du dodécaèdre de mon collègue — indépendamment de l'ordre particulier dans lequel nous pouvons choisir de presser nos boutons respectifs ;

(b) si par hasard mon collègue et moi avons SÉLECTIONNÉ deux sommets exactement *homologues* (*i.e.* situés dans la *même* direction à partir du centre de nos dodécaèdres respectifs), la sonnette retentit alors nécessairement pour au moins l'un des six boutons que nous pressons.

Maintenant, partant du simple fait que *Trucs Quintessentiels* propose des garanties aussi solides sans avoir la moindre idée des boutons que mon collègue ou moi-même pouvons presser, je vais tenter de tirer des conclusions sur les règles que vérifie mon propre dodécaèdre indépendamment de ce qui se passe sur α du Centaure. L'hypothèse clé est qu'il n'existe aucune « influence » à longue distance liant mon dodécaèdre à celui de mon collègue. Ainsi, je vais supposer qu'une fois nos deux dodécaèdres sortis de la chaîne de montage, ils se comportent comme des objets distincts et totalement indépendants. Les conclusions auxquelles j'aboutis alors sont (Fig. 5.4) :

(c) chaque sommet de mon dodécaèdre est, de manière prédéterminée, soit « sonneur » (et je le colore en BLANC), soit « non sonneur » (et je le colore en NOIR) ; le caractère sonneur d'un sommet est indépendant du fait qu'il soit le premier, le deuxième ou le troisième des boutons adjacents au sommet SÉLECTIONNÉ à avoir été pressé ;

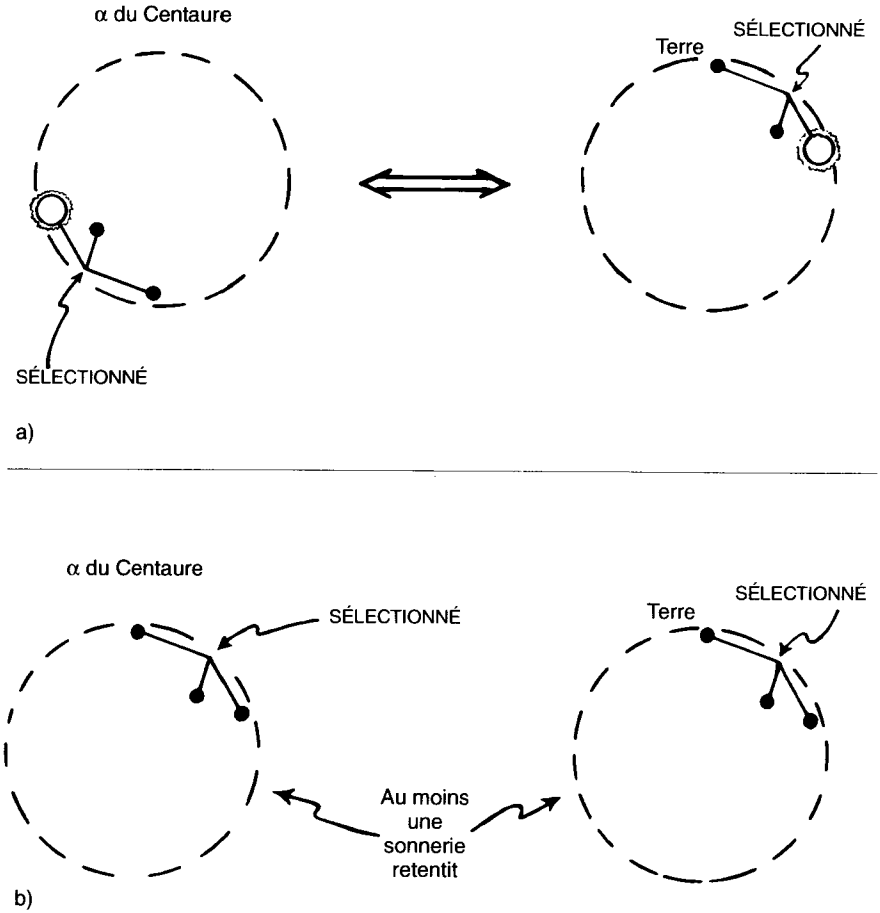


Figure 5.3. La société *Trucs Quintessentiels* garantit deux propriétés : (a) si les sommets SÉLECTIONNÉS sont *opposés*, la sonnette retentit uniquement pour des boutons diamétralement opposés, indépendamment de l'ordre dans lequel ils ont été pressés ; (b) si les sommets SÉLECTIONNÉS sont *homologues*, la sonnette retentit pour au moins l'un des sommets adjacents.

(d) deux sommets adjacents à un même sommet (deux « deuxièmes voisins ») ne peuvent être simultanément sonneurs (*i.e.* ils ne peuvent être tous deux BLANCS) ;

(e) les six sommets adjacents à une paire de sommets antipodiques ne peuvent être simultanément non sonneurs (*i.e.* tous NOIRS).

(L'adjectif « antipodique » qualifie des sommets diamétralement opposés sur le même dodécaèdre.)

La conclusion (c) résulte du fait que mon collègue *pourrait* avoir choisi le sommet diamétralement opposé au sommet que j'ai moi-même choisi ; du moins, *Trucs Quintessentiels* n'a aucun moyen de savoir qu'il ne le choisira pas

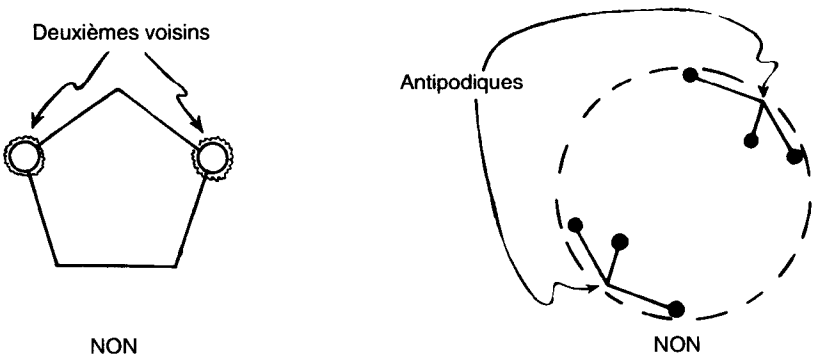


Figure 5.4. Si l'on suppose que nos deux dodécaèdres sont des objets indépendants (sans liens entre eux), on déduit de (a) et (b) que chacun des boutons de mon dodécaèdre est prédéterminé soit comme « sonneur » (BLANC), soit comme « muet » (NOIR), que deux boutons adjacents à un même bouton ne peuvent être simultanément BLANCS, et que les six sommets adjacents à une paire de sommets antipodiques ne peuvent être tous NOIRS.

(contrafactuels !). Ainsi, si l'un des trois boutons que je presse fait retentir la sonnette, le sommet diamétralement opposé, *s'il* est le premier à être pressé par mon collègue, fait également retentir sa sonnette. Il en est ainsi quel que soit l'ordre dans lequel j'ai choisi de presser mes trois boutons, de sorte que (en vertu de l'hypothèse sur l'absence d'« influence » à longue distance entre les dodécaèdres) nous pouvons être certains que *Trucs Quintessentiels* a prédéfini comme sonneur ce sommet particulier — indépendamment de l'ordre dans lequel je presse mes boutons — sinon, il y aurait une contradiction avec (a).

La conclusion (d) découle pareillement de (a). Supposons en effet que deux sommets adjacents à un même sommet de mon dodécaèdre soient sonneurs. Quel que soit celui des deux que je choisisse de presser en premier, sa sonnette doit retentir — et supposez que j'aie fait de leur voisin commun mon sommet SÉLECTIONNÉ. La sonnette qui retentit *dépend* maintenant de l'ordre dans lequel j'appuie sur ces sommets, ce qui contredit (a) si mon collègue prend comme sommet SÉLECTIONNÉ le sommet opposé au mien (une éventualité que *Trucs Quintessentiels* a bien sûr prévue).

Enfin, (e) découle de (b) et des deux conclusions que nous venons d'établir. Supposez en effet que mon collègue ait choisi le sommet *homologue* à mon sommet SÉLECTIONNÉ. Si aucun des trois boutons adjacents à mon sommet SÉLECTIONNÉ n'est sonneur, l'un au moins des trois boutons de mon collègue est, en vertu de (b), sonneur. Il s'ensuit, en vertu de (a), que mon propre sommet, opposé à celui, sonneur, de mon collègue, est également sonneur. Cela établit (e).

Voici maintenant l'énigme. Essayez de colorer en BLANC ou en NOIR chacun des sommets d'un dodécaèdre conformément aux conclusions (d) et (e). Quoi que vous fassiez, vous constaterez que c'est impossible. Il s'avère donc plus intéressant de *démontrer* que ce problème *n'admet pas* de solution.

Afin de laisser au lecteur suffisamment motivé une chance de bâtir une telle démonstration, je remets à l'appendice B l'exposé de celle que j'ai trouvée. Elle est relativement simple et peut-être certains lecteurs en trouveront-ils de plus sophistiquées.

Se pourrait-il que pour la première fois depuis un million de siècles, *Trucs Quintessentiels* ait commis une erreur ? Puisque nous avons établi qu'il est *impossible* de colorer nos sommets conformément à (c), (d) et (e), pourquoi alors ne pas lorgner sur la somme rondelette promise à celui qui prendra en défaut ses garanties et attendre impatiemment les quelque quatre années-lumière nécessaires à la transmission, par mon collègue, du message dans lequel il nous dit ce qu'il a fait, et quand, et si sa sonnette a retenti ? Malheureusement, lorsque ce message nous parviendra, il mettra en même temps fin à tous nos espoirs de pactole, car il s'avérera alors que *Trucs Quintessentiels* a de nouveau raison !

Le raisonnement de l'appendice B montre qu'il n'y a tout simplement *aucun moyen*, à l'aide d'un modèle physique classique, de construire des dodécaèdres magiques satisfaisant aux propriétés avancées par *Trucs Quintessentiels*, les deux dodécaèdres étant censés se comporter comme des objets indépendants une fois sortis de l'usine. Il est en effet *impossible* de garantir les deux propriétés (a) et (b) sans maintenir une sorte de « lien » mystérieux entre ces deux dodécaèdres — un lien toujours présent lorsque nous pressons les boutons situés sur leurs sommets et qui apparemment agit instantanément, faisant fi des quatre années-lumière qui nous séparent de α du Centaure. Pourtant, *Trucs Quintessentiels* certifie ces deux propriétés — aussi impossible que cela paraisse — et n'a jamais été pris en défaut !

Comment la firme *Trucs Quintessentiels* — souvent appelée « TQ » pour abréger — fait-elle cela ? Bien sûr, « TQ » signifie en réalité *Théorie Quantique* ! Et ce qu'elle a fait, c'est suspendre un atome de *spin* $\frac{3}{2}$ au centre de chacun de nos dodécaèdres. Ces deux atomes ont été initialement produits sur Bételgeuse par fission d'un atome de spin 0, puis délicatement installés au centre de nos deux dodécaèdres, de sorte que la somme de leurs spins reste égale à zéro. (Nous verrons à la section 5.10 ce que tout cela signifie.) Ainsi, lorsque mon collègue ou moi pressons l'un des boutons de nos dodécaèdres, nous effectuons une sorte de mesure (partielle) de spin, dans la direction définie par le centre du dodécaèdre et ce sommet. Si le résultat de la mesure est positif, la sonnette retentit et le feu d'artifice se déclenche immédiatement. À la section 5.18, je préciserai la nature de cette mesure et je montrerai, ainsi qu'à l'appendice B, en quoi (a) et (b) sont une conséquence des règles standard de la théorie quantique.

La conclusion insolite de cette démonstration est que la théorie quantique *viole* l'hypothèse de l'absence d'« influence » à longue distance ! Un coup d'œil sur le diagramme d'espace-temps de la figure 5.5 suffit pour se convaincre que les pressions que mon collègue et moi exerçons sur nos boutons sont *séparées par des intervalles du genre espace* (*cf.* §4.4), de sorte que selon la théorie de la relativité, il ne peut y avoir entre nous de signal transmettant l'information sur les boutons pressés et sur la réaction des sonnettes qui leur sont associées.

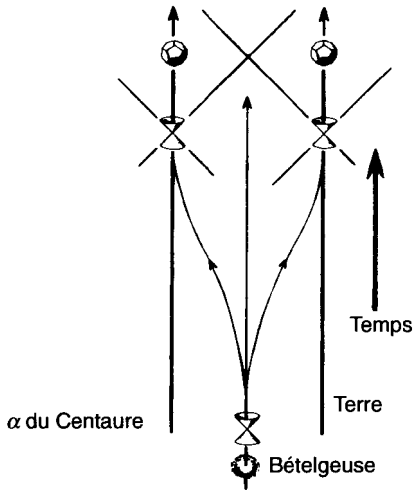


Figure 5.5. Diagramme d'espace-temps de l'histoire des deux dodécaèdres. Ils parviennent sur α du Centaure et sur Terre en des points séparés par des intervalles du genre espace.

Pourtant, selon la théorie quantique, une « influence » est bien présente, qui relie nos dodécaèdres séparés par des intervalles du genre espace. En fait, on ne peut utiliser cette « influence » pour transmettre instantanément une *information* directement utilisable, et il n'y a donc aucune contradiction de nature opérationnelle entre la relativité restreinte et la théorie quantique. Il y a toutefois une contradiction avec l'*esprit* de la relativité restreinte — contradiction qu'illustre l'une des plus profondes énigmes-**Y** de la théorie quantique, à savoir le phénomène de la *non-localité quantique*. Les deux atomes situés aux centres de nos dodécaèdres constituent un seul *état emmêlé*, et en vertu des règles de la théorie quantique standard, ils *ne peuvent* être considérés comme des objets distincts et indépendants.

5.4 Le statut expérimental des énigmes-Y de type EPR

L'exemple particulier que je viens de donner appartient à une classe d'expériences (de pensée) dites « de type EPR » en référence à un article célèbre publié en 1935 par Albert Einstein, Boris Podolsky et Nathan Rosen. (Voir §5.17 pour un examen plus détaillé des effets EPR.) Cet article faisait intervenir non pas le spin, mais certaines combinaisons de la position et de la quantité de mouvement. Par la suite, David Bohm en présenta une version mettant en jeu une paire de particules de spin $\frac{1}{2}$ (disons, des électrons) émis dans un état

combiné de spin total égal à 0. Ces expériences de pensée semblaient montrer qu'une mesure effectuée sur l'un des membres d'une paire de particules quantiques pouvait « influencer » instantanément et dans un sens très précis l'autre membre de la paire, même si ce dernier se trouvait à une distance arbitraire du premier. Une telle « influence » ne peut toutefois être utilisée pour transmettre un message réel entre ces deux particules. Ces dernières sont pour ainsi dire *emmêlées*. C'est à Schrödinger (1935*b*) que l'on doit d'avoir révélé que ce phénomène d'*emmêlement* — une authentique énigme-**Y** — est une caractéristique de la théorie quantique.

Bien plus tard, en 1966, John Bell montra, à l'aide d'un remarquable théorème, que certaines relations mathématiques (les inégalités de Bell) unissaient les probabilités combinées des diverses mesures de spin que l'on pouvait effectuer sur une paire de deux telles particules. Si, conformément à la représentation de la physique classique, ces particules étaient des entités distinctes et indépendantes, ces relations devaient être satisfaites ; en revanche, si la représentation quantique des particules prévalait, ces relations devaient être violées dans une proportion bien précise. Cette situation allait permettre désormais de trancher entre la localité et la non-localité des processus quantiques : il suffisait de concevoir des expériences concrètes vérifiant si ces relations étaient ou non violées. (Voir EOLP, p. 307, 504 pour des exemples sur ce point.)

Pour illustrer ce que de tels « emmêlements » *ne signifient pas*, John Bell aimait citer l'exemple des *chaussettes de Bertlmann*. Bertlmann était un collègue de Bell qui portait invariablement des chaussettes dépareillées. (Ayant personnellement rencontré Bertlmann, je peux dire que mes propres observations concordent avec celles de Bell.) Ainsi, si par hasard on apercevait la chaussette gauche de Bertlmann et que l'on remarquait qu'elle était verte, on savait *ipso facto* que sa chaussette droite *n'était pas* verte. Il ne serait toutefois pas raisonnable de conclure à l'existence d'une influence mystérieuse se propageant instantanément de la chaussette gauche à la droite. Ces deux chaussettes sont des objets indépendants et point n'est besoin d'en appeler à *Trucs Quintessentiels* pour savoir qu'elles vérifient la propriété des « chaussettes différentes ». Cette propriété peut être aisément garantie par Bertlmann lui-même : il lui suffit de décider à l'avance que ses chaussettes seront de couleurs différentes. Les chaussettes de Bertlmann ne violent pas les relations de Bell et ne sont aucunement liées par une « influence » à grande distance. En revanche, dans le cas des dodécaèdres magiques, aucune interprétation de type « chaussettes de Bertlmann » ne peut expliquer les propriétés garanties par TQ, et c'est cela, en définitive, qui constitue la singularité de la discussion donnée à la section précédente.

Quelques années après la parution de l'article de Bell, plusieurs expériences furent suggérées[2], puis réalisées[3]. La plus probante fut effectuée en 1981 à Paris par Alain Aspect et ses collègues sur des paires de photons « emmêlés » (*cf.* §5.17) émis dans des directions opposées et enregistrés à quelque 12 mètres l'un de l'autre. Les prédictions de la théorie quantique — en l'occurrence, la violation des relations de Bell — furent triomphalement confirmées et avec elles la réalité physique des énigmes-**Y** de type EPR avancée par la théorie quantique standard (Fig. 5.6).

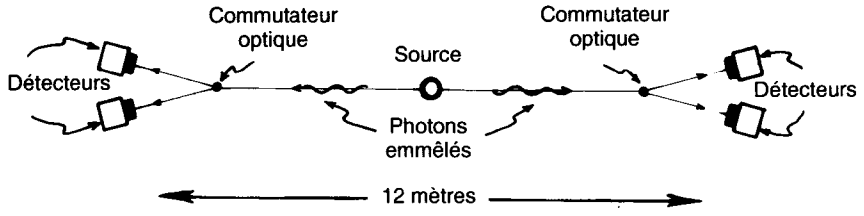


Figure 5.6. Dispositif expérimental utilisé par Alain Aspect et ses collègues pour étudier une paire de photons EPR. Les deux photons émis par la source sont dans un état emmêlé. Le choix de la direction dans laquelle est mesurée la polarisation de chaque photon se fait une fois les photons séparés par une bonne distance et à l'instant même précédant la mesure — trop tard pour qu'un message envoyé par un photon puisse atteindre le photon opposé pour l'informer de la direction de la mesure.

Signalons que malgré l'accord extrêmement précis entre les résultats expérimentaux d'Aspect et les prédictions de la théorie quantique, certains physiciens nient que cette expérience ait établi une quelconque non-localité des processus quantiques. Ils affirment que les détecteurs de photons de cette expérience (et d'autres expériences analogues) sont relativement peu sensibles et ne détecteraient pas la majorité des paires de photons émises lors de mesures effectuées sur un laps de temps plus conséquent. Ils doivent dès lors démontrer qu'en améliorant la sensibilité de ces détecteurs, on parviendrait à détruire l'accord excellent entre l'observation et les prédictions de la théorie quantique, et donc à restaurer la pertinence des relations de Bell, valides uniquement pour un système classique — respectant donc la localité. Il m'apparaît toutefois excessivement improbable que l'excellent accord entre la théorie quantique et l'expérience témoigné par l'expérience d'Aspect (Fig. 5.7) soit un artefact — dû au manque de sensibilité des détecteurs — et que des détecteurs plus sophistiqués non seulement feraient disparaître cet accord, mais en outre permettraient de rétablir les relations de Bell[4].

Le raisonnement initial de Bell établissait des relations (des inégalités) entre les *probabilités* combinées de différentes mesures possibles. Pour évaluer les probabilités réelles intervenant dans une expérience physique, il faut disposer d'observations à suffisamment long terme devant ensuite être soumises à une analyse statistique appropriée. Plus récemment, on a proposé un certain nombre d'expériences (hypothétiques) de type entièrement oui/non, dépourvues de tout aspect probabiliste. La première de ces suggestions est due à Greenberger, Horne et Zeilinger (1989) et met en jeu des mesures de spin sur des particules de spin $\frac{1}{2}$ en *trois* endroits distincts (par exemple, la Terre, α du Centaure et Sirius, si une telle expérience était confiée à *Trucs Quintessentiels*). Quelques années plus tôt, en 1967, Kochen et Specker émirent une idée similaire recourant à des particules de spin 1, mais à la configuration géométrique très complexe ; une année auparavant, en 1966, Bell lui-même avait fait une proposition très semblable, mais de manière moins explicite. (Dans leurs

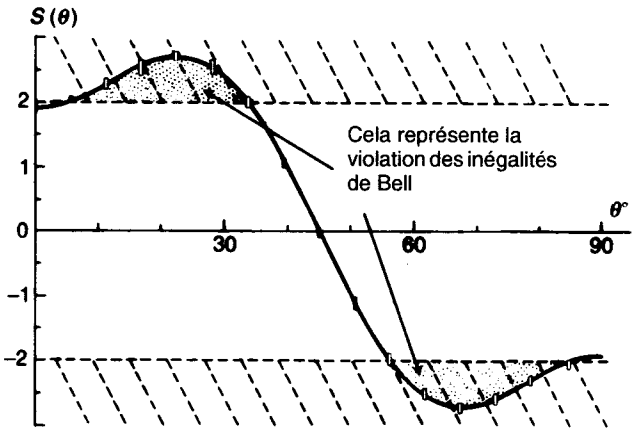


Figure 5.7. L'expérience d'Aspect est en très bon accord avec les prédictions de la théorie quantique — en violant les inégalités de Bell. On imagine difficilement que des détecteurs plus précis puissent infirmer cet accord.

formulations initiales, ces deux propositions ne faisaient aucune allusion aux phénomènes EPR ; Heywood et Redhead (1983) et Stairs (1983) démontrèrent par la suite qu'elles étaient effectivement applicables à ce type de processus[5].) L'exemple des dodécaèdres magiques exposé à la précédente section présente quelques avantages dans la mesure où il admet une géométrie explicite[6]. (Il existe en fait quelques propositions d'expériences destinées à analyser des situations équivalentes à mes deux exemples d'énigmes-Y ; elles ont toutefois une forme physique différente de celle que j'ai donnée ici. Voir Zeilinger *et al.* (1994).)

5.5 Les fondements mathématiques de la théorie quantique : une histoire extraordinaire

Quels sont les principes de base de la théorie quantique ? Avant de répondre à cette question, je vais m'autoriser une petite digression historique. Elle aura l'avantage de mettre en relief le statut des deux composants mathématiques les plus importants de cette théorie. Curieusement, et ce fait est largement méconnu, il se trouve que les deux ingrédients les plus fondamentaux de la théorie quantique moderne ont vu le jour au XVI[e] siècle, tout à fait indépendamment, et qu'ils furent l'œuvre d'un seul et même homme !

Cet homme, Jérôme Cardan — Gerolamo Cardano (Fig. 5.8) — naquit dans la misère (de parents non mariés) à Pavie, le 24 septembre 1501, devint le plus célèbre et le plus grand médecin de son temps, et mourut dans la misère

Figure 5.8. Jérôme Cardan (1501-1576). Médecin prodigieux, inventeur, joueur, écrivain et mathématicien. Il découvrit la théorie des probabilités et les nombres complexes — les deux ingrédients fondamentaux de la théorie quantique moderne.

à Rome, le 20 septembre 1576. Cardan fut un homme extraordinaire, bien que son nom soit très peu connu aujourd'hui. J'espère que le lecteur me pardonnera cette petite parenthèse à son propos avant d'aborder la théorie quantique proprement dite.

En fait, Cardan est totalement ignoré en mécanique quantique — bien que son *nom*, du moins, soit très connu en mécanique *automobile*! Car le joint universel unissant la boîte de vitesses d'une voiture ordinaire à ses roues arrière, permettant ainsi la souplesse nécessaire pour absorber le mouvement vertical variable de l'essieu à ressort arrière, est appelé un *arbre de cardan*. Cardan, qui avait inventé ce dispositif vers 1545, l'intégra en 1548 au train de roues d'un véhicule royal de l'empereur Charles V, assurant ainsi un transport confortable sur des chaussées très accidentées. Il fit de nombreuses autres inventions, notamment une serrure à combinaison semblable à celle utilisée sur les coffres modernes. Médecin réputé, il soigna des princes et des rois. Il fit accomplir de nombreux progrès à la médecine et écrivit un grand nombre de livres médicaux et autres. Il semble être le premier à avoir remarqué que les maladies vénériennes désignées aujourd'hui sous le nom de gonorrhée et de syphilis sont en fait deux maladies distinctes, exigeant donc des traitements différents. Il proposa un traitement de type « sanatorium » pour les tuberculeux — quelque 300 ans avant qu'on ne redécouvre ce

mode de thérapie, essentiellement sous l'impulsion de George Boddington, vers 1830. En 1552, il guérit John Hamilton, l'archevêque d'Écosse, d'un asthme gravement débilitant — modifiant ainsi le cours même de l'histoire d'Angleterre.

Quel rapport avec la théorie quantique ? Aucun, sauf que ce même homme découvrit les deux éléments les plus importants de cette théorie. Car s'il fut un médecin et un inventeur éminent, Cardan fut également un mathématicien exceptionnel !

Le premier de ces éléments est la théorie des *probabilités*. C'est bien connu, la théorie quantique est plus probabiliste que déterministe. Ses règles mêmes dépendent, à un niveau fondamental, des lois des probabilités. En 1524, Cardan écrivit son *Liber de ludo aleae* (*Le livre des jeux de hasard*) qui posa les fondements de la théorie mathématique des probabilités. Il en fit lui-même bon usage, puisque ses études à la faculté de médecine de Pavie furent gagnées... au jeu ! Il avait sûrement vite compris que gagner en *trichant* était une entreprise risquée, car l'homme dont sa mère était veuve avait connu une fin désagréable à cause d'une telle activité. Cardan constata qu'il pouvait s'enrichir honnêtement en usant de ses découvertes des lois de probabilité.

Le second ingrédient fondamental pour la théorie quantique découvert par Cardan est le concept de *nombre complexe*. Un nombre complexe est un nombre de la forme

$$a + \mathrm{i}b,$$

où « i » désigne la racine carrée de moins un,

$$\mathrm{i} = \sqrt{-1}$$

et où a et b sont des nombres réels ordinaires (*i.e.* des nombres que nous écrivons aujourd'hui à l'aide d'un développement décimal). a et b s'appellent respectivement la *partie réelle* et la *partie imaginaire* du nombre complexe $a + \mathrm{i}b$. Cardan rencontra ces nombres étranges en recherchant la solution de l'équation générale du troisième degré. Cette équation s'écrit

$$Ax^3 + Bx^2 + Cx + D = 0,$$

où A, B, C et D sont des nombres réels donnés et où x désigne l'inconnue à déterminer. En 1545, Cardan publia l'*Ars magna*, ouvrage dans lequel apparut pour la première fois la résolution complète de ce type d'équation.

La publication de cette résolution donna lieu à un incident fâcheux. En 1539, un professeur de mathématiques appelé Nicolo « Tartaglia » était déjà en possession de la solution générale d'une certaine classe assez large d'équations du troisième degré, et Cardan avait chargé un ami d'apprendre auprès de lui ce qu'était cette solution. Tartaglia ayant refusé de révéler sa solution, Cardan se mit au travail et la découvrit rapidement. Il la publia en 1540 dans un ouvrage intitulé *la Pratique de la mesure simple et arithmétique*. En fait, Cardan parvint à généraliser les résultats de Tartaglia de manière à tenir compte de *tous* les cas possibles et publia son analyse de la méthode de

résolution générale dans son *Ars magna*. Dans ces deux ouvrages, il reconnut la paternité de Tartaglia pour la solution de la classe d'équations que celui-ci avait considérée, mais dans l'*Ars magna*, il commit l'erreur de prétendre que Tartaglia lui avait donné la permission de publier cette solution. Tartaglia, furieux, affirma qu'ayant un jour rendu visite à Cardan, il lui avait révélé sa solution en lui imposant, sous la foi du serment, de ne jamais la dévoiler. Quoi qu'il en soit, Cardan aurait difficilement pu publier ses propres travaux — qui prolongeaient ceux de Tartaglia — sans révéler la solution des cas antérieurs, et l'on ne voit guère comment il aurait pu procéder autrement, si ce n'est en ne parlant de rien. Tartaglia dès lors cultiva sa rancune à l'égard de Cardan et attendit l'heure de la revanche. En 1570, après diverses sombres actions qui lui permirent de ternir sérieusement la réputation de Cardan, il fut l'un des instigateurs du coup final qui devait entraîner sa chute. Tartaglia, qui travaillait en étroite collaboration avec l'Inquisition, élabora une longue liste de faits pouvant être utilisés contre Cardan et s'arrangea pour le faire arrêter et emprisonner. Cardan fut cependant libéré en 1571, après l'intervention de l'émissaire spécial de l'archevêque d'Écosse (rappelons-le, l'archevêque avait été guéri de son asthme par Cardan) qui expliqua que Cardan était un « savant qui s'inquiète seulement de préserver et soigner les corps afin que les âmes de Dieu puissent y séjourner le plus longtemps possible ».

Les « sombres actions » mentionnées plus haut se réfèrent au procès pour meurtre intenté au fils aîné de Cardan, Giovanni Battista. Lors de ce procès, Cardan mit en jeu son honneur en prenant la défense de son fils. Ce fut une grave erreur, car il s'avéra que Giovanni était en fait coupable — il avait tué sa femme (qu'il avait été contraint d'épouser pour dissimuler un autre meurtre). Apparemment, le meurtre de la femme de Giovanni fut facilité par un autre fils de Cardan encore plus vaurien, Aldo, qui par la suite trahit Giovanni puis livra son propre père à l'Inquisition de Bologne. En récompense, Aldo devint tortionnaire et exécuteur public au service de l'Inquisition bolognaise. Cardan eut aussi une fille qui ne servit pas davantage sa réputation : elle mourut de la syphilis des suites de ses activités professionnelles.

Il serait intéressant, d'un point de vue de psychologie historique, d'analyser comment Cardan, qui semble avoir été un père bienveillant, attaché à ses enfants et à sa femme, et qui fut un homme à principes, honnête et sensible, a pu hériter d'une progéniture aussi désastreuse. Nul doute que son attention fut bien trop fréquemment détournée des problèmes familiaux par ses nombreux et prenants centres d'intérêt. Nul doute que son absence de plus d'une année, après la mort de sa femme, lorsqu'il se rendit en Écosse soigner l'archevêque (bien qu'initialement, les deux hommes devaient se rencontrer à Paris), ne fut guère favorable à l'éducation de ses enfants. Nul doute aussi que, convaincu qu'il était que les étoiles lui avaient prédit sa mort en 1546, il se consacra fiévreusement à ses activités de recherche et d'écriture, négligeant sa femme qui, ironie du sort, mourut vers la fin de cette même année.

Je tendrais volontiers à croire que le peu de célébrité dont jouit injustement Cardan aujourd'hui est dû à son destin malchanceux et à sa réputation gravement ternie — grâce aux efforts combinés de ses enfants, de l'Inquisition et en

particulier de Tartaglia. Il est selon moi l'une des plus brillantes figures de la Renaissance. Bien qu'il ait grandi dans la misère, il évolua, durant ses années formatrices, dans une atmosphère intellectuelle. Son père, Fazio, était géomètre. Cardan lui-même devait rappeler que jeune enfant, il accompagna son père lors d'une visite que celui-ci rendit à Léonard de Vinci ; les deux hommes discutèrent jusque tard dans la nuit de problèmes de géométrie.

En ce qui concerne la publication par Cardan des résultats de Tartaglia avec la soi-disant permission de ce dernier, je dirais que je préfère un savant qui rend publiques ses propres découvertes à un autre qui les maintient secrètes. Si l'on ne peut nier (eu égard aux compétitions mathématiques publiques auxquelles il prit souvent part) que les moyens d'existence de Tartaglia dépendaient dans une certaine mesure du secret dont il entourait ses découvertes, leur divulgation par Cardan eut un effet profond et durable sur le développement des mathématiques. En outre, pour ce qui est de leur paternité, il semble que l'on doive en vérité l'attribuer à un autre savant, Scipion del Ferro, qui fut professeur à l'université de Bologne jusqu'à sa mort en 1526. On sait que del Ferro fut au moins en possession de la solution que Tartaglia devait découvrir plus tard ; on n'est cependant pas certain qu'il ait su que l'on pouvait modifier cette solution pour englober les cas considérés par Cardan, ni qu'il fût conduit à envisager l'existence des nombres complexes.

Revenons maintenant plus en détail sur l'équation du troisième degré et voyons en quoi la contribution de Cardan fut si fondamentale. Grâce à une substitution du type $x \rightarrow x + a$, l'équation générale se réduit facilement à la forme

$$x^3 = px + q,$$

où p et q sont des nombres réels. Il semble que ce résultat était bien connu à l'époque. Toutefois, il faut souligner que les *nombres négatifs* n'étaient habituellement pas considérés comme de « vrais » nombres, de sorte qu'en fonction des signes de p et q, on écrivait différentes versions de cette équation (*e.g.* $x^3 + p'x = q$, $x^3 + q'x = px$) afin de travailler uniquement sur des nombres non négatifs. Cependant, afin d'éviter des complications excessives, j'adopterai ici la notation moderne (qui, elle, autorise les nombres négatifs).

On peut exprimer graphiquement les solutions de l'équation du troisième degré en traçant les courbes $y = x^3$ et $y = px + q$, puis en regardant les intersections des deux courbes. Les valeurs de x en ces points d'intersection sont les solutions de l'équation. Sur la figure 5.9, la courbe $y = x^3$ est la ligne sinueuse, tandis que $y = px + q$ est représentée par une droite dont on montre diverses possibilités. (J'ignore si Cardan ou Tartaglia utilisèrent une telle procédure graphique — rien n'interdit de le penser. Elle sert uniquement à faciliter la visualisation des différentes situations pouvant survenir.) En notation moderne, les cas résolus par Tartaglia correspondent à p négatif (ou nul). $y = px + q$ descend alors vers la droite, à l'instar de P sur la figure 5.9. Notez que dans ce cas, il y a toujours un seul point d'intersection avec la courbe cubique, de sorte que l'équation du troisième degré n'admet qu'une seule solution. En notation moderne, la solution de Tartaglia s'écrit

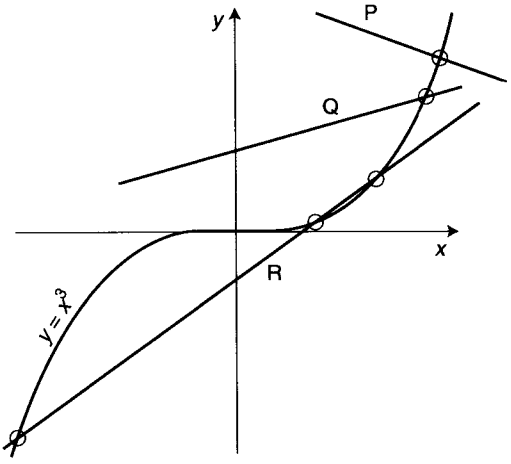


Figure 5.9. Les solutions de l'équation cubique $y^3 = px + q$ peuvent s'obtenir graphiquement par l'intersection (les intersections) de la droite $y = px + q$ avec la courbe cubique $y = x^3$. Le cas étudié par Tartaglia correspond à $p < 0$, autrement dit à une droite descendante telle P. Cardan, lui, étudia le cas $p > 0$, illustré par les droites Q ou R. Le *casus irreductibilis* correspond à *trois* points d'intersection, comme avec la droite R. Le calcul des coordonnées de ces trois points d'intersection exige une incursion dans le domaine des nombres complexes.

$$x = \sqrt[3]{(w + \tfrac{1}{2}q)} - \sqrt[3]{(w - \tfrac{1}{2}q)}$$

où

$$w = \sqrt{\left[(\tfrac{1}{2}q)^2 + (\tfrac{1}{3}p')^3 \right]}$$

avec $p' = -p$, de sorte que les grandeurs apparaissant dans l'équation restent non négatives (on prend également $q > 0$).

La généralisation de cette procédure par Cardan englobe les cas correspondant à $p > 0$ (et $q < 0$, mais le signe de q n'est pas très important). Maintenant, $y = px + q$ monte vers la droite (à l'instar de Q et R sur la figure 5.9). On voit que pour une valeur donnée de p (*i.e.* pour une pente donnée), si q' $(= -q)$ est suffisamment grand (de sorte que la droite coupe l'axe des y suffisamment loin de l'axe des x), il y a encore une seule solution. L'expression trouvée par Cardan s'écrit

$$x = \sqrt[3]{(-\tfrac{1}{2}q' + w)} + \sqrt[3]{(-\tfrac{1}{2}q' - w)}$$

où

$$w = \sqrt{\left[(\tfrac{1}{2}q')^2 - (\tfrac{1}{3}p)^3 \right]}$$

En utilisant la notation moderne et le concept actuel de nombre négatif (ainsi que le fait que la racine cubique d'un nombre négatif est égale à moins la racine cubique de la forme positive de ce nombre), on voit que l'expression de Cardan n'est en apparence pas fondamentalement différente de celle de Tartaglia. Toutefois, elle contient un élément entièrement nouveau, car si q' n'est pas trop grand, la droite peut couper la courbe en *trois* points distincts, de sorte que l'équation originale admet alors trois solutions (dont deux négatives si $p > 0$). Ce cas — appelé « *casus irreductibilis* » — survient lorsque $(\frac{1}{2}q')^2 < (\frac{1}{3}p)^3$, et l'on constate que w est maintenant la *racine carrée d'un nombre négatif*. Ainsi, les nombres $\frac{1}{2}q' + w$ et $\frac{1}{2}q' - w$ figurant sous les signes de racine cubique sont ce que l'on désigne aujourd'hui sous le nom de *nombres complexes* ; pourtant, l'addition des deux racines cubiques doit donner un nombre réel pour fournir les solutions de l'équation.

Cardan était tout à fait conscient de ce mystérieux problème. Plus tard, dans son *Ars magna*, il analysa explicitement les difficultés soulevées par l'apparition des nombres complexes lors de la résolution d'équations. C'est ainsi qu'il considéra le problème de la recherche de deux nombres dont le produit et la somme valent respectivement 40 et 10, problème qu'il résolut (correctement) en obtenant les deux nombres complexes

$$5 + \sqrt{(-15)} \text{ et } 5 - \sqrt{(-15)}$$

D'un point de vue graphique, ce problème se ramène à trouver les intersections de la courbe $xy = 40$ avec la droite $x + y = 10$. On voit, sur la figure 5.10, que cette droite et cette courbe n'ont pas d'intersection (en termes de nombres réels), ce qui signifie en l'occurrence que l'on doit recourir aux nombres complexes pour trouver la solution de ce problème. Cardan éprouvait un certain malaise avec ces nombres, qualifiant même de « torture mentale » le fait de devoir travailler avec eux — mais son étude des équations du troisième degré l'obligea à en tenir compte.

Signalons que l'apparition des nombres complexes dans la résolution de l'équation du troisième degré — équation illustrée graphiquement par la figure 5.9 — recèle quelque chose de bien plus subtil que leur apparition dans la solution du problème décrit par la figure 5.10 (qui demande essentiellement de résoudre l'équation quadratique $x^2 - 10x + 40 = 0$). Dans ce dernier cas, il est clair qu'il n'y a aucune solution sauf si l'on s'autorise l'emploi des nombres complexes, et l'on peut dès lors très bien soutenir que ces nombres sont purement fictifs et uniquement introduits pour obtenir une « solution » à une équation qui en est en fait dépourvue. Toutefois, cette attitude n'explique pas ce qui se passe avec l'équation du troisième degré. Ici, le « *casus irreductibilis* » (droite R sur la figure 5.9) admet bel et bien trois solutions *réelles* dont on ne peut nier l'existence ; pourtant, pour exprimer n'importe laquelle de ces solutions en termes de grandeurs irrationnelles (*i.e.* en l'occurrence, en termes de racines carrées et de racines cubiques), nous devons faire une incursion dans le monde mystérieux des nombres complexes, même si notre destination finale nous ramène dans le monde des réels.

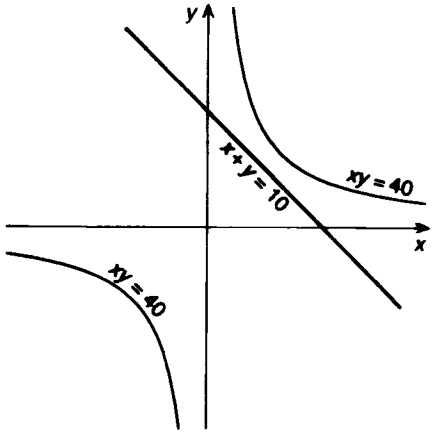


Figure 5.10. Cardan s'intéressa également à la recherche de deux nombres dont le produit et la somme sont respectivement égaux à 40 et à 10. Ce problème se ramène graphiquement à chercher les intersections de la courbe $xy = 40$ avec la droite $x + y = 10$. Il est clair que ce problème n'admet pas de solutions réelles.

Il semble que personne avant Cardan n'ait entrevu ce monde mystérieux ni n'ait perçu en quoi il pourrait être le fondement du monde même de la « réalité ». (D'autres savants, tels Héron et Diophante, qui vécurent à Alexandrie respectivement au premier et au troisième siècle de notre ère, semblent avoir caressé l'idée qu'un nombre négatif pouvait avoir une sorte de « racine carrée », mais ni l'un ni l'autre ne furent assez audacieux pour associer de tels « nombres » aux nombres réels et former les nombres *complexes*, pas plus qu'ils n'ont soupçonné de lien sous-jacent avec les solutions réelles des équations.) Peut-être l'étrange combinaison, chez Cardan, d'une personnalité mystique et scientifiquement rationnelle lui permit-elle de saisir les premières intuitions de ce qui devait devenir l'un des concepts mathématiques les plus puissants. Plus tard, grâce aux travaux de Bombelli, Coates, Euler, Wessel, Argand, Gauss, Cauchy, Weierstrass, Riemann, Levi, Lewy et de nombreux autres, la théorie des nombres complexes s'est épanouie pour constituer l'une des structures mathématiques les plus élégantes et connaissant le plus d'applications. Il fallut toutefois attendre l'avènement de la théorie quantique, dans le premier quart de notre siècle, pour que se révèlent non seulement le rôle étrange et omniprésent joué par les nombres complexes au niveau de la structure fondatrice du monde physique dans lequel nous vivons, mais aussi leur lien profond avec les *probabilités*. Cardan lui-même ne pouvait soupçonner le lien mystérieux unissant ses deux plus grandes contributions aux mathématiques — ce lien qui forme la base même de l'Univers matériel à l'échelle microscopique.

5.6 Les règles fondamentales de la théorie quantique

Quel est ce lien ? Comment les nombres complexes et la théorie des probabilités s'unissent-ils pour donner une description indéniablement superbe des mécanismes intimes de notre Univers ? On peut dire *grosso modo* que les lois des nombres complexes exercent leur emprise sur les phénomènes se déroulant au niveau microscopique, tandis que les probabilités agissent dans l'espace séparant ce niveau de celui de nos perceptions ordinaires — on verra plus loin ce que cela signifie précisément.

Examinons tout d'abord le rôle des nombres complexes. Ces derniers apparaissent d'une manière si curieuse qu'il est très difficile d'accepter qu'ils puissent fournir une authentique description de la réalité physique. Ils semblent en effet ne jamais intervenir dans le comportement des corps au niveau des phénomènes perceptibles par nos sens et régis par les lois classiques de Newton, Maxwell et Einstein. Ainsi, pour avoir une idée du fonctionnement de la théorie quantique, nous devrons, du moins provisoirement, supposer que les actions physiques s'effectuent sur deux niveaux distincts : le niveau *quantique* sous-jacent, où ces nombres complexes jouent leur étrange rôle, et le niveau *classique* des lois physiques familières opérant à grande échelle. Les nombres complexes n'interviennent qu'au niveau quantique — et semblent totalement absents du niveau classique. Cela ne signifie pas nécessairement qu'il existe une frontière physique entre le niveau des phénomènes quantiques et celui des phénomènes classiques, mais il sera utile pour l'instant d'imaginer l'existence d'une telle frontière afin de comprendre les procédures adoptées en théorie quantique. Nous le verrons, l'une de nos principales préoccupations sera la question plus profonde de la *réalité* d'une telle frontière physique.

Où se situe le niveau quantique ? C'est celui des corps physiques qui, comme les molécules, les atomes ou les particules fondamentales, sont en un certain sens « suffisamment petits ». Cette « petitesse » n'a pas forcément trait à la distance physique. Les effets quantiques peuvent se manifester sur d'immenses distances. Rappelons-nous les 4 années-lumière qui séparaient nos deux dodécaèdres à la section 5.3, ou encore les 12 mètres séparant les paires de photons dans l'expérience d'Aspect (§5.4). Ce n'est pas la petitesse physique qui définit le niveau quantique, mais quelque chose de plus subtil — sur quoi il vaut mieux pour l'instant ne pas être trop précis. Pour simplifier, on pourra cependant considérer que le niveau quantique correspond aux phénomènes mettant en jeu d'infimes différences d'énergie. Je reviendrai plus en détail sur ce sujet à la section 6.12.

Le niveau classique, en revanche, est celui de l'expérience ordinaire. Il est régi par les lois de la physique classique, exprimées à l'aide de nombres réels. Les descriptions auxquelles il donne lieu — telles celles donnant la position, la vitesse et la forme d'une balle de golf — parlent directement à notre intuition. L'existence d'une distinction physique *réelle* entre le niveau quantique et le niveau classique est une question profonde, intimement liée au problème des énigmes-**X** mentionnées à la section 5.1. Nous l'avons dit, par souci de

commodité, nous allons pour l'instant supposer l'existence d'une telle distinction.

Quel est donc le rôle fondamental joué par les nombres complexes au niveau quantique ? Considérons une particule individuelle, par exemple un électron. Dans la représentation classique, cet électron occupe une position A ou peut-être une autre position B. Selon la théorie quantique, en revanche, il peut occuper un éventail de positions bien plus large. Non seulement il peut se trouver en A ou en B, mais il peut également se trouver dans un certain nombre d'états pour lesquels, en un sens bien défini, il occupe simultanément *ces deux* positions ! Désignons par $|A\rangle$ l'état dans lequel l'électron occupe la position A et par $|B\rangle$ celui dans lequel il occupe la position B*. Selon la théorie quantique, les autres états quantiques permis pour l'électron s'écrivent sous la forme

$$w|A\rangle + z|B\rangle$$

où les coefficients de pondération w et z sont des *nombres complexes* (dont l'un au moins est différent de zéro).

Qu'est-ce que cela signifie ? Si les coefficients de pondération étaient des nombres *réels* non négatifs, cette somme serait, en un certain sens, une probabilité attachée à la position de l'électron, w et z représentant les probabilités relatives pour que l'électron soit, respectivement, en A ou en B. w/z serait alors le rapport (probabilité de présence de l'électron en A)/(probabilité de présence de l'électron en B). Ainsi, si A et B étaient les deux seules possibilités autorisées à l'électron, on aurait une probabilité $w/(w+z)$ pour qu'il soit en A et une probabilité $z/(w+z)$ pour qu'il soit en B. Dans le cas $w = 0$, l'électron serait certain d'être en B ; dans le cas $z = 0$, il serait certain d'être en A. Si l'état était simplement « $|A\rangle + |B\rangle$ », l'électron aurait des probabilités *égales* de se trouver en A ou en B.

Mais w et z étant des nombres *complexes*, cette interprétation perd sa validité. Le rapport des coefficients de pondération w et z *n'est pas* un rapport de probabilités, et cela parce que les probabilités sont toujours des nombres *réels*. En dépit de l'idée répandue selon laquelle le monde quantique est un monde

---

* Par souci de commodité, j'utiliserai la notation standard de Dirac consistant à désigner les états quantiques par des « kets » ($|\ \rangle$). Les lecteurs non familiers de cette notation quantique n'ont pas à se soucier de sa signification†.

Paul Dirac fut l'un des plus éminents physiciens du XX^e^ siècle. C'est à lui que l'on doit notamment une formulation générale des lois de la théorie quantique, ainsi que l'extension relativiste de cette théorie, qui devait le conduire à la découverte de l'« équation de Dirac » pour l'électron. Dirac avait une aptitude particulière à « sentir » la vérité : il jugeait ses équations en se fondant, dans une large mesure, sur leurs qualités *esthétiques* !

† Précisons cependant pour le lecteur français l'origine de ce terme : *ket* est en anglais la deuxième syllabe de *bracket* qui signifie *crochet* ou *parenthèse*. Dans la notation de Dirac, le produit scalaire de deux états quantiques $|\psi\rangle$ et $|\varphi\rangle$ s'écrit formellement $\langle\psi|\varphi\rangle$ (*cf.* §5.12). On obtient ainsi une « parenthèse », un *bra(c)ket* dont $\langle\psi|$ et $|\varphi\rangle$ sont respectivement le *bra* et le *ket*. Ces deux « morphèmes » ont été intégrés à la terminologie quantique des chercheurs français. (N.d.T.)

probabiliste, *ce n'est pas* la théorie des *probabilités* de Cardan qui opère au niveau quantique. C'est en revanche sa mystérieuse théorie des *nombres complexes* qui sous-tend une description mathématiquement précise et *non probabiliste* des phénomènes quantiques.

Le langage de tous les jours échoue à expliquer ce que « signifie » qu'un électron se trouve dans un état de superposition en deux endroits à la fois, avec des coefficients de pondération w et z complexes. Pour le moment, nous admettrons simplement que cette description est indissociable des systèmes quantiques. De telles superpositions constituent un élément important de la construction réelle de notre monde microscopique tel qu'il nous a été aujourd'hui révélé par la nature. Ce comportement insolite et mystérieux est apparemment une « donnée » du monde quantique. Ces descriptions sont parfaitement bien définies — et nous mettent en présence d'un monde microscopique évoluant selon un schéma mathématiquement précis et, en outre, *totalement déterministe.*

5.7 L'évolution unitaire U

Ce schéma déterministe est ce que l'on appelle l'*évolution unitaire*, évolution que je désignerai par la lettre **U**. Cette évolution est décrite par des équations mathématiques précises, mais peu nous importera ce que sont exactement ces équations. Nous nous intéresserons uniquement à certaines propriétés de **U**. Dans ce que l'on appelle « représentation de Schrödinger », **U** est décrite par l'*équation de Schrödinger.* Cette équation donne les variations en fonction du temps de l'*état quantique* ou *fonction d'onde.* Cet état quantique, souvent désigné par la lettre grecque ψ (prononcée « psi ») ou par $|\psi\rangle$, est la somme pondérée — à coefficients complexes — de tous les états de base du système. Ainsi, dans l'exemple de l'électron mentionné à la section précédente (où ces états de base pourraient être une localisation en A ou une localisation en B), l'état quantique $|\psi\rangle$ est la somme

$$|\psi\rangle = w|\mathrm{A}\rangle + z|\mathrm{B}\rangle,$$

où w et z sont des nombres complexes (non simultanément nuls). La combinaison $w|\mathrm{A}\rangle + z|\mathrm{B}\rangle$ s'appelle une *superposition linéaire* des deux états $|\mathrm{A}\rangle$ et $|\mathrm{B}\rangle$. La grandeur $|\psi\rangle$ (ou $|\mathrm{A}\rangle$, ou $|\mathrm{B}\rangle$) est souvent appelée *vecteur d'état.* D'une manière générale, un état quantique (ou un vecteur d'état) s'écrit sous la forme

$$|\psi\rangle = u|\mathrm{A}\rangle + v|\mathrm{B}\rangle + w|\mathrm{C}\rangle + \ldots + z|\mathrm{F}\rangle$$

où u, v, w, ..., z sont des nombres complexes (non tous nuls) et où $|\mathrm{A}\rangle$, $|\mathrm{B}\rangle$, $|\mathrm{C}\rangle$, ..., $|\mathrm{F}\rangle$ représentent, par exemple, les diverses positions possibles d'une particule (ou toute autre propriété de cette particule, telle son état de spin ;

cf. §5.10). Plus généralement encore, une fonction d'onde ou un vecteur d'état peut s'écrire sous forme d'une somme *infinie* (une particule ponctuelle admet en effet une infinité de positions possibles), mais cette éventualité ne nous concernera pas ici.

Il y a un point technique que je me dois de mentionner à propos du formalisme quantique : les *rapports* des coefficients de pondération complexes sont les seuls éléments à avoir une signification physique. Je reviendrai plus loin sur cet aspect. Pour le moment, nous admettrons que pour un vecteur d'état donné $|\psi\rangle$, tout multiple $u|\psi\rangle$ (avec complexe $u \neq 0$) représente le même état *physique* que $|\psi\rangle$. Ainsi, par exemple, $uw|\mathrm{A}\rangle + uz|\mathrm{B}\rangle$ représente le même état physique que $w|\mathrm{A}\rangle + z|\mathrm{B}\rangle$. Il en résulte que seul le rapport w/z — et non w et z séparément — a une signification physique.

La caractéristique la plus fondamentale de l'équation de Schrödinger (*i.e.* de **U**) est qu'elle est une équation *linéaire*. Cela signifie que si l'on a deux états, par exemple $|\psi\rangle$ et $|\phi\rangle$, et si l'équation de Schrödinger nous dit qu'après un certain temps t, ces deux états se sont respectivement transformés en $|\psi'\rangle$ et $|\phi'\rangle$, alors toute superposition linéaire $w|\psi\rangle + z|\phi\rangle$ se trouve transformée, au bout du même temps t, en la superposition correspondante $w|\psi'\rangle + z|\phi'\rangle$. Désignons par le symbole $\rightsquigarrow$ l'évolution au bout d'un temps t. La linéarité de l'équation de Schrödinger affirme que si

$$|\psi\rangle \rightsquigarrow |\psi'\rangle \text{ et } |\phi\rangle \rightsquigarrow |\phi'\rangle,$$

alors

$$w|\psi\rangle + z|\phi\rangle \rightsquigarrow w|\psi'\rangle + z|\phi'\rangle$$

Cela s'applique par conséquent aussi aux superpositions linéaires comprenant plus de deux états quantiques ; par exemple, si au bout d'un temps t, $|\chi\rangle$, $|\psi\rangle$ et $|\phi\rangle$ se transforment respectivement en $|\chi'\rangle$, $|\psi'\rangle$ et $|\phi'\rangle$, alors $u|\chi\rangle + w|\psi\rangle + z|\phi\rangle$ devient $u|\chi'\rangle + w|\psi'\rangle + z|\phi'\rangle$. Ainsi, l'évolution procède toujours comme si chaque composante d'une superposition ignorait la présence des autres composantes. On pourrait dire, en somme, que chacun des « univers » décrits par ces différentes composantes évolue indépendamment, selon la même équation de Schrödinger déterministe, et que cette évolution laisse inchangés les coefficients de pondération complexes de la superposition linéaire décrivant l'état entier.

L'évolution temporelle d'un état individuel s'effectuant comme si les autres états n'existaient pas, on pourrait penser que les superpositions et les coefficients de pondération ne jouent en fait aucun rôle physique. Ce serait une très grave erreur. Permettez-moi de l'illustrer sur un exemple.

Considérez une lumière arrivant sur un miroir semi-argenté — un miroir semi-transparent réfléchissant seulement la moitié de la lumière incidente et transmettant l'autre moitié. En théorie quantique, la lumière est une entité composée de particules appelées *photons*. On pourrait imaginer que si un flux de photons rencontre notre miroir semi-argenté, la moitié de ces photons serait réfléchie et l'autre moitié transmise. Eh bien non ! La théorie quantique

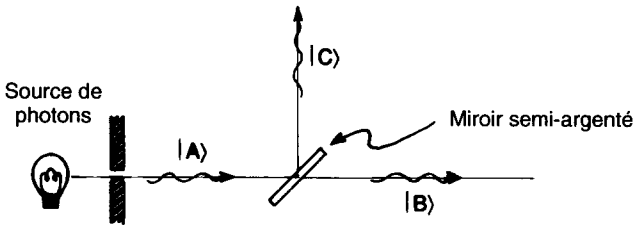


Figure 5.11. Lorsqu'un photon dans un état |A⟩ rencontre un miroir semi-argenté, son état se transforme (sous l'évolution **U**) en l'état superposé |B⟩ + i|C⟩.

nous dit qu'en réalité, *chaque* photon arrivant sur le miroir entre dans un état qui est une *superposition* de photon réfléchi et de photon transmis. Si, avant sa rencontre avec le miroir, le photon est dans un état |A⟩, cet état se transforme, selon **U**, en un autre état que l'on peut écrire |B⟩ + i|C⟩, où |B⟩ représente l'état du photon après sa transmission à travers le miroir, et |C⟩ son état après sa réflexion (Fig. 5.11). Écrivons cela sous la forme

$$|A\rangle \rightsquigarrow |B\rangle + i|C\rangle.$$

Le facteur complexe « i » = $\sqrt{-1}$ est là pour tenir compte du décalage d'un quart de longueur d'onde[7] apparaissant entre les rayons réfléchi et transmis. (Pour être plus complet, j'aurais également dû inclure un terme oscillatoire dépendant du temps et un facteur de normalisation, mais cela ne joue aucun rôle dans la présente discussion. Je ne donnerai ici que ce qui est essentiel pour parvenir à notre objectif. Aux sections 5.11 et 5.12, j'en dirai un peu plus sur, respectivement, le terme oscillatoire et la normalisation. Pour une description plus complète, on peut se référer à n'importe quel ouvrage traitant de théorie quantique[8] ; voir aussi EOLP, p. 272-285.)

En mécanique classique, |B⟩ et |C⟩ représenteraient simplement les deux choses possibles que *pourrait* faire le photon ; en mécanique quantique, en revanche, nous devons imaginer que le photon, dans cet étrange état de superposition, est maintenant en train de faire *ces deux choses en même temps.* Pour voir que cela ne peut résulter de la théorie classique des probabilités, poussons cet exemple un peu plus loin et réunissons les deux composantes de l'état du photon — autrement dit, recombinons les deux rayons lumineux. On y parvient en faisant réfléchir séparément chaque rayon sur un miroir complètement argenté. Après réflexion[9], l'état |B⟩ évolue selon **U** et devient un autre état i|D⟩, tandis que |C⟩ se transforme en i|E⟩ :

$$|B\rangle \rightsquigarrow i|D\rangle \text{ et } |C\rangle \rightsquigarrow i|E\rangle.$$

Ainsi, **U** transforme l'état entier |B⟩ + i|C⟩ selon

$$\begin{aligned} B\rangle + i|C\rangle &\rightsquigarrow (i|D\rangle + i\,(i|E\rangle)\,) \\ &= i|D\rangle - |E\rangle \end{aligned}$$

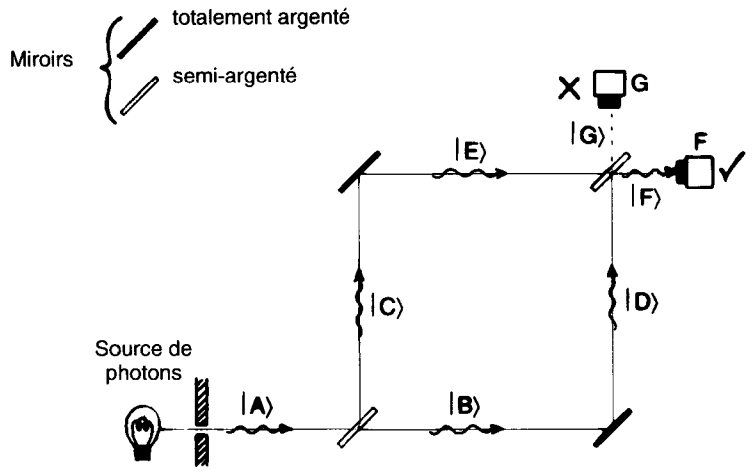


Figure 5.12. Après réflexions sur deux miroirs totalement argentés, les deux parties de l'état du photon interfèrent à la traversée d'un miroir semi-argenté. Le photon émerge alors dans un état $|F\rangle$ et ne pénètre pas dans le détecteur situé en G. (Interféromètre de Mach-Zehnder.)

(car $i^2 = -1$). Supposons maintenant (Fig. 5.12) que ces deux rayons arrivent sur un quatrième miroir qui est, de nouveau, *semi*-argenté (je supposerai que les trajets parcourus par les divers rayons sont égaux, de sorte que le facteur oscillatoire ignoré plus haut ne joue toujours aucun rôle). L'état $|D\rangle$ se transforme en une combinaison $|G\rangle + i|F\rangle$, où $|G\rangle$ représente l'état transmis et $|F\rangle$ l'état réfléchi ; de même, $|E\rangle$ se transforme en $|F\rangle + i|G\rangle$, car c'est maintenant $|F\rangle$ qui est l'état transmis et $|G\rangle$ qui est l'état réfléchi. On a donc :

$$|D\rangle \rightsquigarrow |G\rangle + i|F\rangle \text{ et } |E\rangle \rightsquigarrow |F\rangle + i|G\rangle .$$

L'application de **U** sur l'état entier $i|D\rangle - |E\rangle$ donne

$$\begin{aligned} i|D\rangle - |E\rangle \rightsquigarrow\ & i\,(|G\rangle + i|F\rangle) - (|F\rangle + i|G\rangle) \\ =\ & i|G\rangle - |F\rangle - |F\rangle - i|G\rangle \\ =\ & -2|F\rangle \end{aligned}$$

(Le facteur multiplicatif –2 apparaissant ici n'a aucune signification physique car, nous l'avons dit, la multiplication de l'état entier — ici $|F\rangle$ — d'un système par un nombre complexe non nul ne modifie pas la situation physique.) Nous constatons ainsi que le photon *n'a pas* la possibilité d'être dans l'état $|G\rangle$; les deux rayons se combinent pour ne permettre que la seule possibilité $|F\rangle$. Ce résultat curieux survient parce que les *deux* rayons sont simultanément présents dans l'état physique du photon, entre ses rencontres avec le premier et le dernier miroir. On dit que les deux rayons *interfèrent* l'un avec l'autre. Ainsi, les deux « trajets » qui s'offrent au photon entre ces rencontres ne sont

pas réellement disjoints, mais peuvent s'influencer mutuellement à travers de tels phénomènes d'interférence.

Il faut garder à l'esprit que c'est là une propriété des photons *individuels*. Il faut donc considérer que chaque photon individuel ressent les deux voies qui s'offrent à lui, mais reste *un seul et même* photon ; il ne se divise pas en deux photons lors de la phase intermédiaire, mais sa localisation passe par cette forme d'étrange *coexistence* d'alternatives, pondérée par des nombres complexes et caractéristique de la théorie quantique.

5.8 La réduction **R** du vecteur d'état

Dans l'exemple précédent, le photon émerge du dispositif dans un état non mélangé. Imaginons des détecteurs (des cellules photoélectriques) placés aux points désignés par F et G sur la figure 5.12. Le photon émergeant dans un état (proportionnel à) $|F\rangle$, sans aucune contribution de $|G\rangle$, il s'ensuit que seul le détecteur situé en F enregistre le passage du photon, tandis que le détecteur situé en G n'enregistre rien.

Que se passe-t-il dans une situation plus générale où une superposition d'états telle que $w|F\rangle + z|G\rangle$ rencontre ces détecteurs ? Ils effectuent alors une *mesure* pour voir si le photon est dans un état $|F\rangle$ ou dans un état $|G\rangle$. Une mesure quantique amplifie les événements, les faisant passer du niveau quantique au niveau classique. Au niveau quantique, l'action permanente de l'évolution **U** ne modifie pas les superpositions linéaires. Toutefois, dès que les effets sont amplifiés pour atteindre le niveau classique, où ils sont perçus comme des événements réels, les superpositions à coefficients de pondération complexes disparaissent. Ainsi, dans notre exemple, on constate *soit* un enregistrement par le détecteur en F, *soit* un enregistrement par le détecteur en G, ces deux événements survenant selon certaines probabilités. L'état quantique semble avoir mystérieusement « sauté » de l'état de superposition $w|F\rangle + z|G\rangle$ à un autre état correspondant seulement à $|F\rangle$ ou à $|G\rangle$. Ce « saut » descriptif de l'état du système, qui fait passer de l'état quantique de superposition à un état associé à l'une ou l'autre des possibilités classiques, s'appelle la *réduction du vecteur d'état* ou encore l'*effondrement de la fonction d'onde*. Je désignerai cette opération par la lettre **R**. Cette opération **R** est-elle un processus physique réel, une forme d'illusion, voire une approximation ? Nous le verrons, cette question aura une importance primordiale dans la suite de notre discussion. Le fait que, du moins dans nos descriptions mathématiques, nous devions de temps en temps nous dispenser de **U** et introduire cette procédure **R** totalement différente constitue l'énigme-**X** fondamentale de la théorie quantique. Pour l'instant, il sera préférable de ne pas trop sonder en profondeur ce problème et de considérer (provisoirement) **R** comme un processus *intervenant* simplement (du moins dans les descriptions mathématiques

que nous utilisons) au titre de caractéristique de la procédure d'amplification d'un événement du niveau quantique au niveau classique.

Comment calcule-t-on les *probabilités* associées aux divers résultats possibles d'une mesure effectuée sur un état de superposition ? Il existe en fait une règle remarquable pour y parvenir. Cette règle dit que lorsqu'une mesure décide entre les états possibles $|F\rangle$ et $|G\rangle$ — par exemple, dans la situation précédente, en utilisant respectivement les détecteurs en F et en G — composant une superposition d'états,

$$w|F\rangle + z|G\rangle,$$

le rapport de la probabilité d'enregistrement par le détecteur en F à la probabilité d'enregistrement par le détecteur en G est donné par

$$|w|^2 / |z|^2,$$

où $|w|^2$ et $|z|^2$ sont respectivement les *carrés des modules* des nombres complexes w et z. Le carré du module d'un nombre complexe est par définition la somme des carrés de sa partie réelle et de sa partie imaginaire ; ainsi, pour le nombre complexe

$$z = x + iy,$$

où x et y sont des nombres réels, le carré du module est

$$\begin{aligned} |z|^2 &= x^2 + y^2 \\ &= (x+iy)\,(x-iy) \\ &= z\bar{z} \end{aligned}$$

où $\bar{z}$ $(= x - iy)$ est le *complexe conjugué* de z — et pareillement pour w. (Dans la discussion ci-dessus, j'ai implicitement supposé que les états $|F\rangle$, $|G\rangle$, etc., sont convenablement *normalisés*. J'expliquerai cela plus loin — *cf.* §5.12. Strictement parlant, cette normalisation est nécessaire à la validité de cette règle des probabilités.)

C'est ici, et ici seulement, que les *probabilités* de Cardan font leur entrée sur la scène quantique. On le voit, les probabilités relatives sont données non par les coefficients de pondération complexes du niveau quantique — parce que justement ils sont complexes —, mais par les *carrés des modules* — qui sont, eux, réels — de ces nombres complexes. En outre, c'est seulement maintenant, une fois les *mesures* effectuées, que l'indétermination et les probabilités entrent en jeu. Une mesure sur un état quantique se traduit en effet par une importante amplification d'un processus physique qui, répétons-le, le fait passer du niveau quantique au niveau classique. Dans le cas d'une cellule photoélectrique, l'enregistrement d'un seul événement quantique — la réception d'un photon — provoque une perturbation au niveau classique, par exemple sous forme d'un « clic » audible. On peut aussi utiliser une plaque photographique sensible pour enregistrer l'arrivée d'un seul photon ; celle-ci est alors amplifiée au niveau classique sous forme d'une marque visible sur la plaque. Dans chacun des deux cas, l'appareil de mesure est un dispositif délicatement étalonné

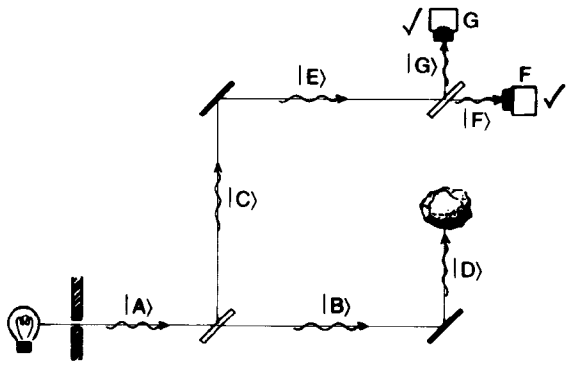


Figure 5.13. L'interposition d'un obstacle sur le chemin du faisceau |D⟩ permet au détecteur situé en G d'enregistrer l'arrivée du photon (à condition que l'obstacle *n'absorbe pas* le photon !).

qui utilise un infime effet quantique pour déclencher un effet bien plus important, observable au niveau classique. C'est dans ce passage du niveau quantique au niveau classique que les nombres complexes de Cardan voient leurs modules élevés au carré pour devenir des probabilités de Cardan !

Voyons comment appliquer cette règle dans une situation particulière. Reprenons le dispositif de la figure 5.12 et remplaçons le miroir de l'angle inférieur droit par une cellule photoélectrique. Celle-ci reçoit alors l'état

$$|B\rangle + i|C\rangle,$$

enregistre l'état |B⟩, mais ignore l'état |C⟩. Ainsi, le rapport des probabilités respectives est égal à $|1|^2/|i|^2 = 1/1$; autrement dit, ces deux possibilités ont des probabilités égales et le photon a autant de chances d'activer la cellule que de ne pas l'activer.

Considérons maintenant un arrangement légèrement plus compliqué. Replaçons le miroir initialement présent dans le coin inférieur droit et interposons à sa sortie — sur le chemin du photon dans l'état |D⟩ — un *obstacle* pouvant éventuellement absorber le photon (Fig. 5.13). L'interférence destructive qui existait auparavant au niveau du dernier miroir disparaît. Si le photon *n'est pas* absorbé par l'obstacle, il émerge du dispositif dans l'état |G⟩ — en plus de l'état |F⟩. Son état à l'approche du dernier miroir est en effet alors $-|E\rangle$ qui, à la traversée de ce miroir, se transforme en $-|F\rangle - i|G\rangle$, de sorte que l'état final se compose des deux possibilités |F⟩ et |G⟩. En revanche, *s'il est* absorbé, il n'émerge pas du dispositif, pas plus dans l'état |F⟩ que dans l'état |G⟩.

Lorsque le photon n'est pas absorbé, les coefficients de pondération complexes des deux possibilités |F⟩ et |G⟩ sont -1 et $-i$ (puisque l'état final est $-|F\rangle - i|G\rangle$). Le rapport des probabilités respectives est donc $|-1|^2/|-i|^2$; autrement dit, les deux états |F⟩ et |G⟩ sont équiprobables et le photon a autant de chances d'activer le détecteur placé en F que celui placé en G.

L'obstacle est en fait, lui aussi, un « appareil de mesure », car les deux termes de l'alternative « l'obstacle absorbe le photon » et « l'obstacle n'absorbe pas le photon » se situent au niveau classique et ne se voient donc pas attribuer de coefficients de pondération complexes. Certes, l'obstacle n'est pas suffisamment structuré pour que l'événement quantique qu'est son absorption d'un photon puisse être amplifié et devenir observable au niveau classique, mais on peut imaginer qu'il *pourrait être* suffisamment structuré pour permettre cette amplification. Concrètement, l'absorption du photon provoque une légère perturbation d'une part considérable du matériau constituant l'obstacle, et il est impossible de collecter toute l'information associée à cette perturbation de manière à reconstituer les effets d'interférences caractérisant les phénomènes quantiques. Ainsi, l'obstacle doit être considéré — du moins en pratique — comme un objet appartenant au niveau classique et jouant un rôle d'appareil de mesure, qu'il enregistre ou non l'absorption du photon d'une manière concrètement observable. (Je reviendrai sur ce point à la section 6.6.)

Dans ces conditions, nous pouvons calculer la probabilité d'absorption du photon par l'obstacle en utilisant la « règle du module au carré ». Lors de sa rencontre avec l'obstacle, le photon est dans l'état $i|D\rangle - |E\rangle$ et son absorption se produit si l'obstacle le rencontre alors qu'il est dans l'état $|D\rangle$ — et non $|E\rangle$. Le rapport absorption/non-absorption vaut $|i|^2/|-1|^2 = 1/1$, de sorte que ces deux possibilités sont équiprobables.

Nous pourrions également imaginer une situation peu différente dans laquelle, au lieu d'avoir un obstacle sur le trajet du photon dans l'état $|D\rangle$, nous attacherions un dispositif de mesure au miroir situé en bas à droite. Supposons que ce dispositif soit si sensible qu'il puisse détecter (*i.e.* amplifier au niveau classique) n'importe quelle impulsion communiquée au miroir par le photon, cette détection se manifestant, par exemple, par le déplacement d'une aiguille (Fig. 5.14). Lors de la rencontre avec le miroir, l'état $|B\rangle$ provoquerait le déplacement de l'aiguille, tandis que l'état $|C\rangle$ n'aurait aucune action. Face à l'état $|B\rangle + i|C\rangle$, ce dispositif déclencherait l'« effondrement de la fonction d'onde » et indiquerait que le photon est *soit* dans l'état $|B\rangle$ (l'aiguille se déplace), *soit* dans l'état $|C\rangle$ (l'aiguille reste immobile), avec des probabilités égales (de rapport $|1|^2/|i|^2$). C'est donc à *ce* stade que se produirait le processus **R**. Le comportement ultérieur du photon serait en gros identique à celui que nous avons décrit plus haut lors de la présence de l'obstacle. On trouverait que les détecteurs F et G ont des chances égales d'enregistrer le photon (que l'aiguille ait ou non bougé). Ce dispositif fonctionne uniquement si le miroir inférieur droit est légèrement « tremblant » afin de permettre l'activation de l'aiguille, et c'est cette non-rigidité du miroir qui perturbe la délicate organisation nécessaire pour assurer l'« interférence destructive » des deux voies suivies par le photon entre A et G, interférence qui avait initialement empêché le détecteur G d'enregistrer le photon.)

Peut-être le lecteur a-t-il le sentiment qu'il n'est pas totalement évident de savoir *quand* — ni même *pourquoi* — il faut changer les règles quantiques en passant du déterminisme quantique à coefficients de pondération complexes aux alternatives du niveau classique, non déterministes et à coefficients de

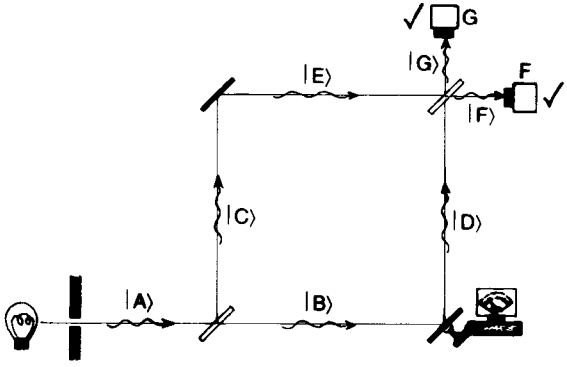


Figure 5.14. On peut obtenir une situation semblable à celle de la figure 5.13 en rendant légèrement « tremblant » le miroir situé en bas à droite et en utilisant son « tremblement » pour savoir, grâce à un détecteur, si le photon a ou non été réfléchi par ce miroir. Ici encore, il y a destruction des interférences et le détecteur situé en G peut éventuellement détecter le photon.

pondération réels, mathématiquement caractérisées par le recours aux carrés des modules des nombres complexes concernés. Qu'est-ce qui fait *réellement* que certains dispositifs matériels, tels les détecteurs de photons F et G ou le miroir inférieur droit — voire l'obstacle placé en D — sont des objets appartenant au niveau classique, tandis que le photon, lui, appartient au niveau quantique ? Est-ce simplement le fait que le photon est un système physique simple admettant un traitement complet en tant qu'objet quantique, tandis que les détecteurs et les obstacles sont des objets complexes qui exigent un traitement approximatif dans lequel les subtilités du comportement quantique, pour une raison ou une autre, disparaissent lors du passage à la moyenne ? Nombre de physiciens affirment qu'il en est ainsi, que nous devrions, en toute rigueur, traiter *tous* les objets physiques — y compris les grands systèmes ou les systèmes complexes — en termes quantiques et que c'est uniquement par commodité que nous les traitons en termes classiques, les règles de probabilité qui interviennent dans la procédure « **R** » découlant pour ainsi dire des approximations faites. Nous verrons aux sections 6.6 et 6.7 que ce point de vue ne nous libère pas vraiment des difficultés que soulèvent les énigmes-**X** de la théorie quantique. Il n'explique pas non plus la miraculeuse règle **R** qui fait surgir des probabilités s'exprimant par les carrés des modules des coefficients de pondération complexes. Pour le moment cependant, nous allons laisser ces problèmes de côté et poursuivre l'examen des conséquences de cette théorie, en nous concentrant sur ses énigmes-**Y**.

5.9 Solution du problème d'Elitzur-Vaidman

Nous disposons maintenant des outils nécessaires pour résoudre le problème proposé à la section 5.2. L'idée consiste à transformer le miroir très sensible fixé au nez de la bombe en un dispositif de mesure — à la manière de l'obstacle ou du système détecteur-miroir de la section précédente. Reprenons donc le dispositif de la figure 5.14, en remplaçant toutefois le miroir inférieur droit par celui de la bombe.

Si la bombe est une « ratée » — dans le sens défini lors de l'énoncé du problème —, son miroir est bloqué dans une position fixe et la situation se ramène à celle décrite à la figure 5.12. Le photon quitte la source de photons dans l'état $|A\rangle$ et — comme à la section 5.7 — émerge finalement dans l'état (proportionnel à) $|F\rangle$. Ainsi, le détecteur situé en F enregistre le passage du photon, tandis que celui situé en G n'enregistre rien.

En revanche, si la bombe *n'est pas* une ratée, son miroir est alors sensible au photon : elle explose si ce photon percute le miroir. La bombe est de fait un dispositif de mesure. Elle amplifie les deux possibilités quantiques « le photon percute le miroir » et « le photon ne percute pas le miroir » qui deviennent alors les deux éventualités classiques : « la bombe explose » et « la bombe n'explose pas ». Elle réagit à l'état $|B\rangle + i|C\rangle$ en explosant si elle trouve le photon dans l'état $|B\rangle$ et en n'explosant pas si elle trouve qu'il n'est pas dans cet état $|B\rangle$ — auquel cas, il est dans l'état $|C\rangle$. Le rapport des probabilités de ces deux événements est $|1|^2/|i|^2 = 1/1$. Si la bombe explose, elle a alors détecté la présence du photon et ce qui arrive ensuite ne nous concerne pas. Si, en revanche, elle n'explose pas, l'état du photon est réduit (par l'action de **R**) à l'état $i|C\rangle$ qui rencontre le miroir supérieur gauche puis émerge de ce miroir dans l'état $-|E\rangle$. Après sa rencontre avec le dernier miroir (semi-argenté), cet état devient $-|F\rangle - i|G\rangle$, de sorte que le rapport des probabilités des deux événements « le détecteur situé en F enregistre l'arrivée du photon » et « le détecteur situé en G enregistre l'arrivé du photon » est égale à $|-1|^2/|-i|^2 = 1/1$ — exactement comme dans les cas considérés à la section précédente, lorsque l'obstacle n'absorbe pas le photon ou lorsque l'aiguille reste immobile. Il existe alors une possibilité bien précise pour que le détecteur situé en G enregistre le photon.

Supposons ensuite que lors de l'envoi d'un photon, on constate que le détecteur situé en G enregistre parfois l'arrivée du photon alors que la bombe n'explose pas. En vertu de ce qui a été dit plus haut, cette situation survient uniquement si la bombe *n'est pas* une ratée ! Lorsqu'elle est une ratée, seul le détecteur situé en F peut enregistrer le photon. Ainsi, chaque fois que nous constatons que le détecteur situé en G enregistre le photon, nous avons la certitude que la bombe est opérationnelle, *i.e.* qu'elle n'est pas une ratée ! C'est la solution que nous recherchions*.

---

* ***L'interrupteur du sabbat.*** Elitzur et Vaidman enseignant tous deux dans des universités israéliennes, Artur Ekert et moi-même avons imaginé un dispositif permettant de venir en

Les valeurs des probabilités mentionnées à l'instant montrent que si on applique ce test à un stock de bombes suffisamment conséquent, la moitié des bombes opérationnelles explosent et sont donc perdues. En outre, le détecteur situé en G enregistre le photon pour la moitié seulement des cas où une bombe opérationnelle n'explose pas. Ainsi, une fois testées toutes les bombes, un quart seulement des bombes initialement opérationnelles se trouvent effectivement *garanties* opérationnelles. On procède alors à un deuxième test sur les bombes restantes en conservant uniquement celles pour lesquelles le détecteur situé en G enregistre le photon. En répétant indéfiniment la procédure, on obtient un tiers (car $\frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots = \frac{1}{3}$) seulement des bombes opérationnelles initialement présentes, mais elles sont *toutes* garanties. (J'ignore ce que l'on fera de ces bombes et je préfère ne pas le savoir !)

Si cette procédure peut sembler très peu rentable, elle a du moins le mérite d'exister. Il n'y a aucun moyen classique de résoudre ce problème. C'est seulement en théorie quantique que les contrafactuels peuvent réellement influencer un résultat physique. Cette procédure quantique permet de résoudre un problème apparemment insoluble — et qui *est* effectivement insoluble dans le cadre de la physique classique. Soulignons en outre que certaines améliorations permettent de réduire les pertes de deux tiers à la moitié (Elitzur-Vaidman 1993). Mieux encore, P. G. Kwiat, H. Weinfurter, A. Zeilinger et M. Kasevich ont récemment montré, en utilisant une procédure différente, que l'on peut réduire les pertes à zéro !

En ce qui concerne l'élément technique le plus délicat de la procédure d'Elitzur-Vaidman, à savoir la nécessité d'utiliser une source émettant un photon à la fois, signalons que l'on sait construire aujourd'hui de tels dispositifs (voir Grangier *et al.* 1986).

Pour terminer, je voudrais souligner qu'un dispositif de mesure n'a nullement besoin d'être aussi spectaculaire que la bombe de notre discussion. Il n'est en fait nullement nécessaire qu'un tel « dispositif » signale au monde extérieur s'il a ou non reçu le photon. Un simple miroir légèrement

---

aide à ceux qui observent strictement la religion juive et donc s'abstiennent rigoureusement de toucher tout interrupteur électrique durant le sabbat. Au lieu de breveter notre ingénieux dispositif, et d'assurer ainsi notre fortune, nous avons généreusement décidé de le rendre public pour le bien-être de l'ensemble de la communauté juive. Pour le construire, il suffit de disposer d'une source émettant des photons individuels, de deux miroirs semi-argentés, de deux miroirs pleinement argentés et d'une cellule photoélectrique liée à l'interrupteur en question. Ce dispositif est identique à celui de la figure 5.13 dans lequel la cellule photoélectrique serait placée en G. Pour fermer ou ouvrir l'interrupteur, on place un doigt sur le rayon en D — à l'instar de l'obstacle à la figure 5.13. Si le photon heurte le doigt, l'état de l'interrupteur n'est pas modifié — et aucun péché n'est commis. (Car les doigts sont en permanence soumis à un flux de photons, même durant le sabbat.) Mais si le photon ne heurte pas le doigt, il y a alors 50 pour cent de chances (si Dieu le veut) que l'état de l'interrupteur se trouve modifié. Ici non plus, on ne commet certainement aucun péché en *ne recevant pas* le photon qui active l'interrupteur! (On pourrait objecter qu'en pratique les sources qui émettent continuellement des photons individuels sont difficiles — et coûteuses — à fabriquer. Mais il n'est pas nécessaire d'avoir une telle source. Toute source de photons convient, car l'argument s'applique séparément à chacun des photons qu'elle émet.)

« tremblant » ferait aussi bien l'affaire, à condition qu'il soit assez léger pour se déplacer de manière appréciable lors de l'impact du photon, avant d'amortir son mouvement par frottement. Le seul fait que le miroir inférieur droit de la figure 5.14 soit « tremblant » permet au détecteur situé en G d'enregistrer le photon, même si en fait ce miroir *ne tremble pas* — ce qui indique alors que le photon a pris l'autre chemin. Il suffit qu'il soit *potentiellement* tremblant pour que le photon puisse atteindre G ! L'obstacle mentionné à la section précédente joue un rôle tout à fait similaire. Il sert à « mesurer » la présence du photon quelque part sur sa trajectoire, telle que la décrivent les états successifs $|B\rangle$ et $|D\rangle$. Le fait qu'il ne reçoive pas le photon, alors qu'il est capable de le recevoir, compte tout autant pour une « mesure » que s'il l'avait effectivement reçu.

Les mesures de ce type, négatives, opérant comme par défaut, s'appellent des mesures à *résultat nul* (ou sans interaction) — voir Dicke (1981) — et ont une importance théorique (voire, peut-être, pratique) considérable. Il existe des expériences qui testent directement les prédictions de la théorie quantique dans de telles situations. En particulier, Kwiat, Weinfurter et Zeilinger ont récemment effectué une expérience qui, dans son principe, correspond *exactement* à la procédure mise en jeu dans la résolution du problème d'Elitzur-Vaidman ! Cela ne nous surprend plus, les prédictions de la théorie quantique ont été entièrement confirmées. Les mesures à résultat nul font réellement partie des énigmes-**Y** de la théorie quantique.

5.10 Théorie quantique du spin ; la sphère de Riemann

Pour résoudre le second des deux mystères quantiques que j'ai proposés, nous allons devoir examiner un peu plus en détail la structure de la théorie quantique. Rappelons (*cf.* §5.3) que mon dodécaèdre et celui de mon collègue possèdent en leur centre un atome de spin $\frac{3}{2}$. Qu'est-ce que le spin ? Quelle est son importance pour la théorie quantique ?

Le spin est une propriété intrinsèque des particules. Il correspond essentiellement au concept physique de rotation sur lui-même (de *moment cinétique*) d'un objet classique, tel une balle de golf ou de cricket, ou la Terre entière, avec toutefois cette différence (mineure) que pour ces grands objets, la principale contribution au moment cinétique est, de loin, celle des mouvements orbitaux de toutes leurs particules autour d'une autre particule, tandis que pour une particule individuelle, le spin est une propriété intrinsèque à la particule elle-même. En fait, le spin d'une particule fondamentale a cette curieuse particularité que sa *valeur* reste *constante*, tandis que sa direction (l'axe de rotation) peut varier — bien que cet « axe » se comporte lui aussi d'une façon très curieuse et qui a peu de rapport, en général, avec un comportement classique. La valeur du spin s'exprime à l'aide de l'unité quantique fondamentale $\hbar$, h

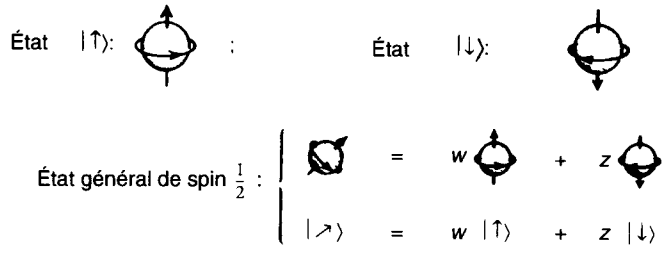


Figure 5.15. L'état de spin d'une particule de spin $\frac{1}{2}$ (par exemple, un électron, un proton ou un neutron) est une superposition complexe des deux états « spin *up* » et « spin *down* ».

étant la constante de Planck et $\hbar$ le symbole introduit par Dirac pour désigner $h/2\pi$. La mesure du spin d'une particule est toujours un multiple entier ou demi-entier (non négatif) de $\hbar$, à savoir 0, $\frac{1}{2}\hbar$, $\hbar$, $\frac{3}{2}\hbar$, $2\hbar$, etc. Les particules sont alors dites respectivement de spin 0, de spin $\frac{1}{2}$, de spin 1, de spin $\frac{3}{2}$, de spin 2, etc.

Considérons d'abord le cas le plus simple (hormis celui du spin 0, qui est *trop* simple ; il correspond à un seul état de spin, à symétrie sphérique), à savoir le cas du spin $\frac{1}{2}$, qui est, par exemple, celui de l'électron ou d'un nucléon (un proton ou un neutron). Pour cette valeur $\frac{1}{2}$, tous les états de spin sont des superpositions linéaires de *deux* états seulement, l'état « spin *up* », désigné par $|\uparrow\rangle$ et correspondant à une rotation vers la droite autour d'un axe vertical orienté *vers le haut*, et l'état « spin *down* », désigné par $|\downarrow\rangle$ et correspondant à une rotation vers la droite autour d'un axe vertical orienté *vers le bas* (Fig. 5.15). L'état de spin général s'écrit alors sous forme d'une combinaison linéaire à coefficients complexes $|\psi\rangle = w|\uparrow\rangle + z|\downarrow\rangle$. Il s'avère en fait que toute combinaison de ce type représente l'état de spin (de valeur $\frac{1}{2}\hbar$) autour d'une direction précise déterminée par le rapport des deux nombres complexes w et z. Le choix des deux états $|\uparrow\rangle$ et $|\downarrow\rangle$ est totalement arbitraire, mais toutes les combinaisons qu'ils engendrent donnent des états de spin parfaitement définis.

Nous allons maintenant expliciter ces relations en nous aidant de la géométrie. Cela nous permettra de voir que les coefficients de pondération complexes w et z ne sont pas des entités aussi abstraites qu'il a pu paraître jusqu'ici. En fait, ils entretiennent un lien précis avec la géométrie dans l'espace. (J'imagine que ce lien géométrique aurait séduit Cardan, voire aurait atténué ses « tortures mentales » — bien que la théorie quantique ait également introduit de nouvelles tortures mentales !)

Considérons la représentation — aujourd'hui classique — des nombres complexes par les points d'un plan. (Ce plan est indifféremment appelé plan d'Argand, plan de Gauss, plan de Wessel, ou tout simplement plan *complexe*.) L'idée consiste à représenter le nombre complexe $z = x + iy$, où x et y sont des nombres réels, par le point du plan dont les coordonnées cartésiennes,

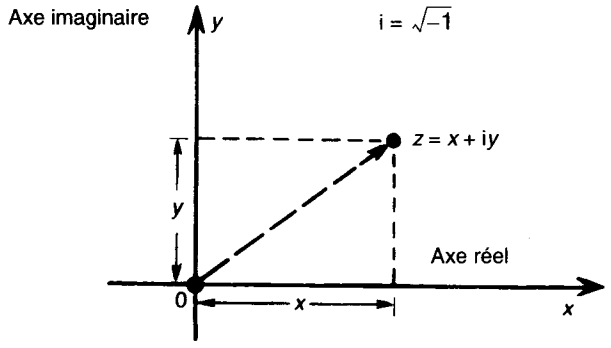


Figure 5.16. Représentation d'un nombre complexe dans le plan complexe (de Wessel-Argand-Gauss).

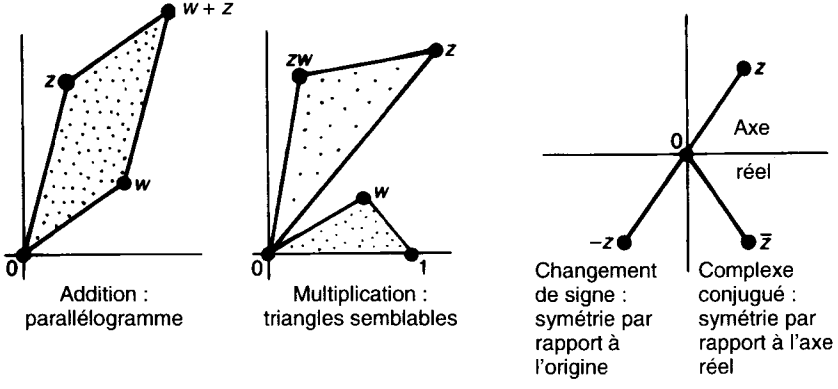


Figure 5.17. Description géométrique des opérations fondamentales sur les nombres complexes.

rapportées à un repère cartésien donné, sont (x, y) (Fig. 5.16). Par exemple, les quatre nombres complexes 1, 1 + i, i et 0 constituent les sommets d'un carré. L'addition et le produit de deux nombres complexes obéissent à des règles géométriques simples (Fig. 5.17). L'opposé $-z$ d'un nombre complexe z s'obtient par symétrie par rapport à l'origine ; le complexe conjugué $\bar{z}$ de z s'obtient par symétrie par rapport à l'axe réel.

Le module d'un nombre complexe est égal à la distance séparant son point représentatif de l'origine ; le carré du module est donc le carré de cette distance. Le *cercle unité* est le lieu des points situés à une distance unité de l'origine (Fig. 5.18) ; ces nombres complexes de *module un* sont de la forme

$$e^{i\theta} = \cos\theta + i \sin\theta,$$

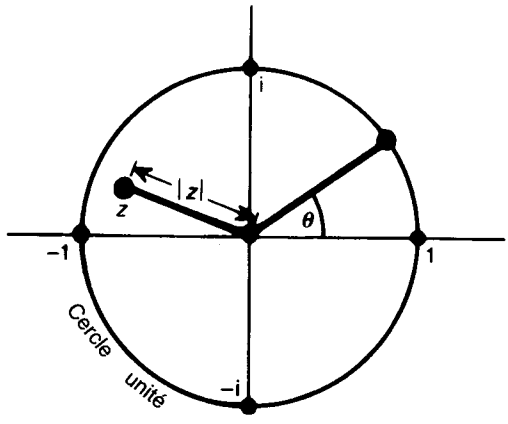


Figure 5.18. Le cercle unité est l'ensemble des nombres complexes de la forme $z = e^{i\theta}$, avec θ réel ; autrement dit c'est l'ensemble des nombres complexes z vérifiant $|z| = 1$.

où θ est un nombre réel mesurant l'angle compris entre l'axe réel et la droite reliant l'origine au point représentatif du nombre complexe*.

Voyons maintenant comment représenter un *rapport* de deux nombres complexes. Lors de la discussion précédente, j'ai indiqué qu'un état quantique n'est pas modifié lorsqu'on le multiplie par un nombre complexe non nul (rappelons, par exemple, que $-2|F\rangle$ et $|F\rangle$ représentent le même état physique). Ainsi, d'une manière générale, $|\psi\rangle$ est physiquement identique à $u|\psi\rangle$ pour tout nombre complexe u non nul. Appliquée à l'état

$$|\psi\rangle = w|\uparrow\rangle + z|\downarrow\rangle,$$

cette règle dit que si l'on multiplie simultanément w et z par le même nombre complexe u non nul, on ne modifie pas la situation physique représentée par cet état. Ce sont les différents *rapports* z/w des deux nombres complexes z et w qui donnent des états de spin physiquement distincts (uz/uw étant égal à z/w si $u \neq 0$).

Comment représenter géométriquement un rapport complexe ? La différence essentielle entre un rapport complexe et un nombre complexe est qu'outre toutes les valeurs complexes finies, ce rapport peut prendre la valeur *infinie* (désignée par le symbole « ∞ »). Ainsi, le rapport z/w présente une difficulté lorsque $w = 0$. Pour tenir compte de cette éventualité, on substitue le

* Le nombre réel e est la « base des logarithmes naturels » : e = 2, 718 281 828 5... ; l'expression e^z désigne « e élevé à la puissance z » et l'on a

$$e^z = 1 + z + \frac{z^2}{1\times 2} + \frac{z^3}{1\times 2\times 3} + \frac{z^4}{1\times 2\times 3\times 4} + \ldots$$

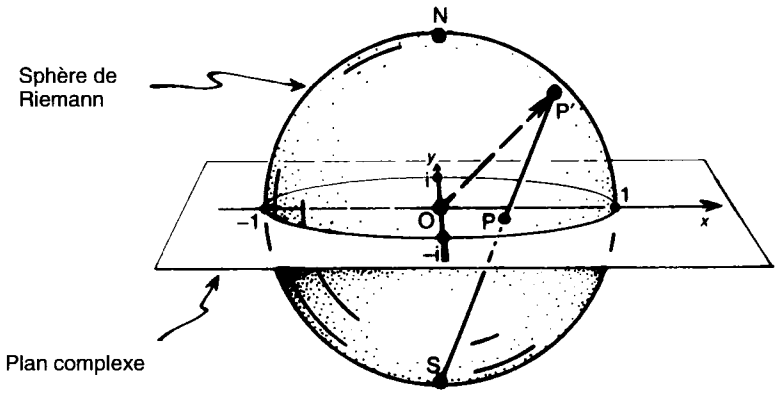


Figure 5.19. La sphère de Riemann. Le point P, qui représente $p = z/w$ dans le plan complexe, se projette à partir du pôle sud en un point P' de la sphère. La direction OP, définie à partir du centre O de la sphère, est la direction de spin pour l'état général de spin $\frac{1}{2}$ mentionné à la figure 5.15.

symbole ∞ à z/w lorsque $w = 0$. Ce cas survient lorsque l'on considère l'état « spin *down* », à savoir $|\psi\rangle = z|\downarrow\rangle = 0|\uparrow\rangle + z|\downarrow\rangle$. Rappelons qu'on ne peut avoir *simultanément* $w = 0$ *et* $z = 0$, mais que $w = 0$ lui-même est tout à fait permis. (Nous pourrions bien sûr utiliser w/z au lieu de z/w, mais nous aurions alors besoin de ∞ pour tenir compte du cas $z = 0$ correspondant à l'état purement « spin *up* ». Peu importe donc le choix adopté.)

Pour représenter l'ensemble de tous les rapports complexes possibles, on utilise une *sphère* appelée *sphère de Riemann.* Les points de cette sphère représentent non seulement les nombres complexes, mais aussi ∞. Cette sphère a un rayon unité, son plan équatorial est le plan complexe, et son centre est l'origine (le zéro) de ce plan. L'équateur de la sphère de Riemann s'identifie donc au cercle unité du plan complexe (Fig. 5.19). Pour représenter un rapport complexe donné, par exemple z/w, on marque sur le plan complexe le point P représentant le nombre complexe $p = z/w$ (nous supposons pour l'instant que $w \neq 0$), puis on projette ce point P sur la sphère à partir du *pôle sud* S. Autrement dit, on prolonge jusqu'à la sphère la droite reliant S à P. Soit P' le point ainsi obtenu. Cette application, qui associe les points du plan complexe aux points de la sphère, s'appelle une *projection stéréographique.* Dans cette projection stéréographique, le pôle S correspond à ∞ : si en effet le point P s'éloigne considérablement de l'origine du plan, le point P' qui lui correspond devient alors voisin de S et se confond avec S lorsque P va à l'infini.

La sphère de Riemann joue un rôle fondamental dans la représentation quantique des systèmes à deux états, même si ce rôle n'est pas toujours immédiatement perceptible. Elle donne une description abstraite et géométrique de l'espace des états physiquement discernables pouvant être construits par superposition linéaire quantique à partir de deux états quantiques distincts. Ces deux états peuvent être par exemple deux positions possibles $|B\rangle$ et $|C\rangle$

d'un photon. Toute superposition linéaire de ces deux états est alors de la forme $w|B\rangle + z|C\rangle$. À la section 5.7, nous avons uniquement utilisé le cas particulier $|B\rangle + i|C\rangle$ correspondant à la réflexion/transmission par un miroir semi-argenté, mais on aurait facilement pu obtenir les autres combinaisons. Il aurait suffi pour cela de modifier l'épaisseur de la couche d'argent sur le miroir et d'introduire un milieu réfringent sur le trajet de l'un des rayons émergents. On aurait ainsi pu construire une sphère de Riemann complète représentant les états associés à toutes les situations physiques de la forme $w|B\rangle + z|C\rangle$ pouvant s'obtenir à partir des deux possibilités $|B\rangle$ et $|C\rangle$.

Si dans un tel contexte le rôle géométrique de la sphère de Riemann n'est pas du tout apparent, il existe toutefois d'autres situations dans lesquelles ce rôle est manifeste. L'exemple le plus frappant est celui des états de spin d'une particule de spin $\frac{1}{2}$, telle un électron ou un proton. L'état de spin général est représenté par la combinaison

$$|\psi\rangle = w|\uparrow\rangle + z|\downarrow\rangle,$$

et il s'avère (en choisissant convenablement $|\uparrow\rangle$ et $|\downarrow\rangle$ à partir de la classe de proportionnalité des possibilités physiquement équivalentes) que ce $|\psi\rangle$ représente l'état de spin, de valeur $\frac{1}{2}\hbar$, correspondant à une rotation vers la droite autour du rayon de la sphère de Riemann aboutissant au point représentant le rapport z/w. Ainsi, chaque direction de l'espace est une direction de spin accessible à une particule de spin $\frac{1}{2}$. Même si la plupart des états sont représentés sous forme de mystérieuses « sommes pondérées à coefficients complexes d'états de base », on voit que ces combinaisons ne sont ni plus ni moins mystérieuses que les deux possibilités initiales $|\uparrow\rangle$ et $|\downarrow\rangle$ dont nous sommes partis. Chacune de ces combinaisons a autant de réalité physique que les autres.

Qu'en est-il des états de spin supérieur ? Là, les choses deviennent un peu plus complexes — et *plus* mystérieuses ! La description générale que je vais donner n'est pas très connue des physiciens d'aujourd'hui ; elle a été introduite en 1932 par Ettore Majorana, un brillant physicien italien qui disparut à l'âge de 31 ans lors d'un naufrage en baie de Naples, dans des circonstances restées mystérieuses.

Considérons tout d'abord ce que les physiciens connaissent le mieux. Supposons que nous ayons un atome (ou une particule) de spin $\frac{1}{2}n$ et concentrons-nous, pour commencer, sur la direction *up*. Demandons-nous quelle « proportion » du spin de cet atome est effectivement orientée dans cette direction (*i.e.* tourne vers la droite autour de cette direction). Il existe un dispositif classique, appelé appareil de Stern-Gerlach, qui permet de mesurer des spins atomiques en utilisant un champ magnétique inhomogène. Les mesures fournies par ce dispositif montrent que le spin peut prendre que $n+1$ orientations possibles, discernables par le fait que l'atome appartient à l'un des $n+1$ faisceaux émergents de l'appareil (Fig. 5.20). La proportion de spin correspondant à la direction choisie est déterminée par le faisceau particulier dans lequel se trouve l'atome. Si l'on prend comme unité $\frac{1}{2}\hbar$, la proportion de spin dans cette direction s'avère avoir l'une des $n+1$ valeurs n, $n-2$, $n-4$, ...,

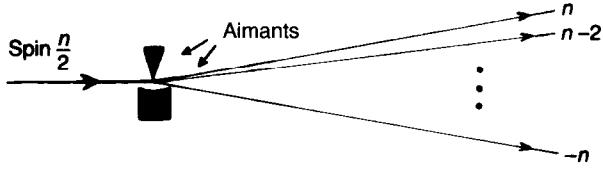


Figure 5.20. Expérience de Stern-Gerlach. La mesure d'une particule de spin $\frac{1}{2}n$ admet $n+1$ résultats possibles correspondant à la proportion de spin se trouvant dans la direction mesurée.

$2-n$, $-n$. Ainsi, les états de spin accessibles à un atome de spin $\frac{1}{2}n$ sont simplement les superpositions complexes de ces possibilités. Je désignerai les divers résultats possibles d'une mesure de Stern-Gerlach sur un spin $\frac{1}{2}n$ — le champ magnétique dans l'appareil étant vertical et orienté vers le haut — par

$$|\uparrow\uparrow\uparrow\ldots\uparrow\rangle, |\downarrow\uparrow\uparrow\ldots\uparrow\rangle, |\downarrow\downarrow\uparrow\ldots\uparrow\rangle, \ldots, |\downarrow\downarrow\downarrow\ldots\downarrow\rangle,$$

correspondant respectivement aux valeurs de spin n, $n-2$, $n-4$, ..., $2-n$, $-n$ sur l'axe du champ magnétique, chaque « ket » contenant exactement n flèches. On peut considérer que chaque flèche orientée vers le haut contribue pour $\frac{1}{2}\hbar$ au spin *up*, tandis que chaque flèche orientée vers le bas contribue pour $\frac{1}{2}\hbar$ au spin *down*. La somme de ces contributions donne la valeur totale du spin obtenue lors d'une mesure (de Stern-Gerlach) de spin effectuée dans la direction verticale.

La superposition générale de ces diverses contributions est donnée par la somme à coefficients complexes

$$z_0|\uparrow\uparrow\uparrow\ldots\uparrow\rangle + z_1|\downarrow\uparrow\uparrow\ldots\uparrow\rangle + z_2|\downarrow\downarrow\uparrow\ldots\uparrow\rangle + \ldots + z_n|\downarrow\downarrow\downarrow\ldots\downarrow\rangle,$$

où les nombres complexes z_0, z_1, z_2, ..., z_n ne sont pas tous nuls. Peut-on représenter un tel état en termes de directions qui ne soient pas simplement *up* ou *down* ? Majorana a démontré que c'était effectivement possible à condition que les diverses flèches pointent dans des directions totalement indépendantes, autrement dit, qu'elles ne soient pas obligatoirement alignées sur une paire de directions opposées, comme c'est le cas lors d'une mesure de type Stern-Gerlach. Dès lors, l'état général de spin $\frac{1}{2}n$ est représenté par un ensemble de n « flèches » indépendantes ; ces flèches sont associées à n points de la sphère de Riemann et partent du centre de la sphère pour aboutir chacune à l'un de ces points (Fig. 5.21). Il importe de noter que ces points (ces flèches) forment un ensemble *non ordonné*, et donc qu'aucun d'eux ne peut être qualifié de premier, de deuxième, de troisième, etc.

Tout cela donne une image très curieuse du spin dès que l'on tente de mettre ce concept en parallèle avec celui de rotation ordinaire propre au niveau classique. Lorsqu'il tourne sur lui-même, un objet classique — par exemple, une balle de golf — admet un axe de rotation bien défini, tandis qu'un objet quantique peut apparemment admettre, simultanément, une multitude d'axes

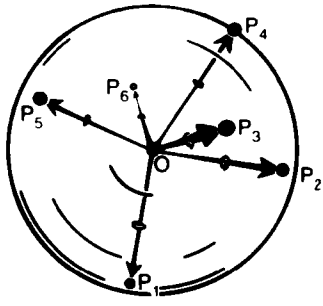


Figure 5.21. Dans la description de Majorana, l'état général de spin $\frac{1}{2}n$ est représenté par un ensemble non ordonné de n points P_1, P_2, ..., P_n de la sphère de Riemann. On peut considérer que chacun de ces points est un élément de spin $\frac{1}{2}$ associé au vecteur partant du centre de la sphère et aboutissant au point considéré.

de rotation orientés dans des directions différentes. Si l'on cherche à se représenter un objet classique comme un objet quantique qui serait en un certain sens simplement plus « gros », il semble que l'on soit confronté à un paradoxe. Plus le spin a une valeur élevée, plus nombreuses sont les directions qui lui sont accessibles. Pourquoi alors les objets classiques ne tournent-ils pas simultanément dans de nombreuses directions ? C'est là un exemple d'énigme-**X** quantique. Quelque chose intervient (à un niveau inconnu), et l'on constate que la plupart des états quantiques ne se manifestent pas (ou, du moins, ne se manifestent presque jamais) au niveau classique des phénomènes perceptibles par nos sens. Dans le cas du spin, on trouve que les seuls états qui subsistent de façon significative au niveau classique sont ceux qui correspondent aux axes proches d'une direction particulière, à savoir l'axe de rotation de l'objet classique sur lui-même.

En théorie quantique, un principe — appelé « principe de correspondance » — affirme que lorsque des grandeurs physiques (telles la valeur du spin) prennent des valeurs élevées, le système *peut* alors avoir un comportement qui est une très bonne approximation du comportement classique. Toutefois, ce principe ne nous dit pas comment ces états peuvent résulter de la seule action de l'équation de Schrödinger **U**. En fait, les « états classiques » n'apparaissent presque jamais de cette manière. Ils apparaissent sous l'action d'une autre procédure : la réduction **R** du vecteur d'état.

5.11 Position et quantité de mouvement d'une particule

Le concept quantique de *position* d'une particule offre un exemple encore plus net de ce type de situation. Nous avons vu que l'état d'une particule peut mettre en jeu des superpositions de deux ou plusieurs positions différentes.

(Rappelons-nous la discussion de la section 5.7 dans laquelle, après avoir rencontré un miroir semi-argenté, le photon est dans un état qui le place simultanément sur deux rayons différents.) De telles superpositions sont pareillement possibles pour d'autres types de particules — simples ou composites — telles un électron, un proton, un atome ou une molécule. En outre, rien, dans la description **U** du formalisme de la théorie quantique, n'interdit que les gros objets — par exemple, les balles de golf — puissent se trouver eux-mêmes en plusieurs endroits à la fois, par suite de la superposition d'états. Toutefois, nous n'observons jamais une balle de golf occupant simultanément plusieurs positions, pas plus que nous ne l'observons en train de tourner simultanément autour de plusieurs axes. Pourquoi certains objets semblent-ils trop gros ou trop massifs (ou trop autre chose) pour être des objets « quantiques » et n'occupent-ils jamais des états de superposition dans le monde réel ? En théorie quantique standard, seule **R** permet d'accomplir la transition des superpositions d'états quantiques vers un état classique, réel et unique. La seule action de **U** conduit presque invariablement à des superpositions classiques « d'apparence déraisonnable ». (Je reviendrai sur ce problème à la section 6.1.)

Au niveau quantique, en revanche, les états qui permettent à une particule de n'avoir aucune position bien définie jouent un rôle fondamental. Car si la particule a une *quantité de mouvement* bien déterminée (et se déplace de manière précise dans une certaine direction et non dans une superposition de plusieurs directions différentes), son état correspond alors nécessairement à une superposition de toutes les *positions* possibles. (C'est là une caractéristique propre à l'équation de Schrödinger ; elle est cependant trop technique pour pouvoir être développée ici ; voir EOLP, p. 261-269, ou, par exemple, Dirac (1947) et Davies (1984). Elle est également intimement liée au *principe d'indétermination* de Heisenberg qui impose des limites à la définition simultanée et précise de la position et de la quantité de mouvement.) En fait, les états dont la quantité de mouvement est bien définie présentent un comportement spatial oscillatoire dans la direction du mouvement — c'est ce comportement que nous avons ignoré lors de la discussion des états du photon à la section 5.7. Rigoureusement parlant, « oscillatoire » n'est pas le terme exact. Il s'avère que ces « oscillations » diffèrent des vibrations linéaires d'une corde, pour lesquelles on peut considérer que les coefficients de pondération complexes oscillent de part et d'autre de l'origine dans le plan complexe ; ces coefficients sont en fait des nombres complexes de module 1, dont seule la phase oscille, et (Fig. 5.18) qui décrivent un cercle autour de l'origine à vitesse constante — cette vitesse correspondant à une fréquence ν proportionnelle à l'énergie E de la particule en vertu de la célèbre relation de Planck $E = h\nu$ (Voir la figure 6.11 de EOLP pour une représentation imagée, en « tire-bouchon », des états de quantité de mouvement.) Bien qu'importants pour la théorie quantique, ces aspects ne seront pas particulièrement déterminants pour les discussions contenues dans ce livre, de sorte que le lecteur peut tranquillement les ignorer.

D'une manière générale, les coefficients de pondération complexes n'ont pas cette forme « oscillatoire », mais varient d'un point à un autre de manière

arbitraire. Ils sont une fonction complexe dépendant de la position et appelée *fonction d'onde* de la particule.

5.12 L'espace de Hilbert

Afin de décrire plus explicitement (et plus précisément) la manière dont, selon la théorie quantique standard, la procédure **R** est censée agir, il me faut au préalable présenter certaines définitions mathématiques (relativement élémentaires). L'ensemble de tous les états possibles d'un système quantique constitue ce que l'on appelle un *espace de Hilbert*. S'il ne nous sera pas nécessaire d'expliquer en détail toutes les caractéristiques mathématiques de cet espace, la connaissance de certaines de ses propriétés nous permettra cependant d'avoir une idée de l'image que les physiciens se font actuellement de l'univers quantique.

La première et la plus importante propriété d'un espace de Hilbert est qu'il s'agit d'un *espace vectoriel complexe*. Cela signifie que l'on peut y former les combinaisons à coefficients de pondération complexes que nous avons considérées pour les états quantiques. Autrement dit — en utilisant les « kets » de Dirac pour désigner les éléments de cet espace —, si $|\psi\rangle$ et $|\phi\rangle$ sont deux éléments d'un espace de Hilbert et si w et z sont deux nombres complexes, $w|\psi\rangle + z|\phi\rangle$ est aussi un élément de cet espace. Si $w = z = 0$, on obtient l'élément $\mathbf{0}$, unique élément de l'espace de Hilbert à ne représenter aucun état physique. Un espace de Hilbert vérifie les règles algébriques habituelles des espaces vectoriels, à savoir

$$\begin{aligned}
|\psi\rangle + |\phi\rangle &= |\phi\rangle + |\psi\rangle, \\
|\psi\rangle + (|\phi\rangle + |\chi\rangle) &= (|\psi\rangle + |\phi\rangle) + |\chi\rangle, \\
w(z|\psi\rangle) &= (wz)|\psi\rangle, \\
(w+z)|\psi\rangle &= w|\psi\rangle + z|\psi\rangle, \\
z(|\psi\rangle + |\phi\rangle) &= z|\psi\rangle + z|\phi\rangle, \\
0|\psi\rangle &= \mathbf{0}, \\
z\mathbf{0} &= \mathbf{0},
\end{aligned}$$

ce qui signifie plus ou moins que l'on peut utiliser les notations algébriques conformément à ce que dicte l'intuition.

Un espace de Hilbert peut avoir un nombre *fini* de dimensions, comme dans le cas des états de spin d'une particule. Pour un spin $\frac{1}{2}$, l'espace de Hilbert a simplement deux dimensions et ses éléments sont les combinaisons linéaires complexes des deux états $|\uparrow\rangle$ et $|\downarrow\rangle$. Pour un spin de valeur $\frac{1}{2}n$,

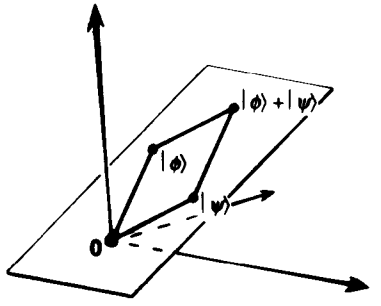


Figure 5.22. Si on assimile l'espace de Hilbert à un espace euclidien tridimensionnel, la somme de deux vecteurs $|\psi\rangle$ et $|\phi\rangle$ se détermine à l'aide de la règle habituelle du parallélogramme (appliquée dans le plan défini par **0**, $|\psi\rangle$ et $|\phi\rangle$).

l'espace de Hilbert a $n + 1$ dimensions. Un espace de Hilbert peut cependant avoir parfois un nombre *infini* de dimensions, comme c'est le cas pour l'espace des états de position d'une particule. Chaque position possible de la particule constitue alors une dimension de l'espace de Hilbert. L'état général décrivant la position quantique de la particule est une superposition complexe de *toutes* ces différentes positions possibles — c'est la fonction d'onde de la particule. En fait, le concept d'espace de Hilbert de dimension infinie engendre certaines complications mathématiques qui embrouilleraient inutilement notre discussion ; je me concentrerai donc principalement ici sur des espaces de Hilbert de dimension finie.

La visualisation des espaces de Hilbert se heurte à deux difficultés. D'une part, ces espaces ont généralement trop de dimensions pour parler directement à notre imagination, d'autre part ils sont non pas réels, mais *complexes.* Néanmoins, si l'on veut parvenir à une idée intuitive des mathématiques sous-jacentes à ces espaces, il est souvent commode d'ignorer provisoirement ces difficultés. Admettons donc pour l'instant que nous puissions représenter un espace de Hilbert par un espace ordinaire à deux ou à trois dimensions. La figure 5.22 illustre géométriquement l'opération de superposition linéaire dans le cas d'un espace réel tridimensionnel.

Rappelons qu'un vecteur d'état quantique $|\psi\rangle$ représente la même situation physique que tout multiple $u|\psi\rangle$ où u est un nombre complexe $\neq 0$. Géométriquement parlant, cela signifie qu'une situation physique donnée est représentée non par un point de l'espace de Hilbert, mais par une droite — on dit un *rayon* — passant par le point de l'espace de Hilbert associé à $|\psi\rangle$ et l'origine **0** (Fig. 5.23). Nous devons toutefois garder à l'esprit que les espaces de Hilbert sont non pas réels, mais complexes, et donc que si un rayon *ressemble* à une droite ordinaire à une dimension, il est en réalité à lui seul un plan complexe tout entier.

Ce que nous avons dit jusqu'ici concerne uniquement la structure vectorielle des espaces de Hilbert. Il existe cependant une autre propriété des espaces de Hilbert, qui est presque aussi cruciale que leur structure vectorielle et

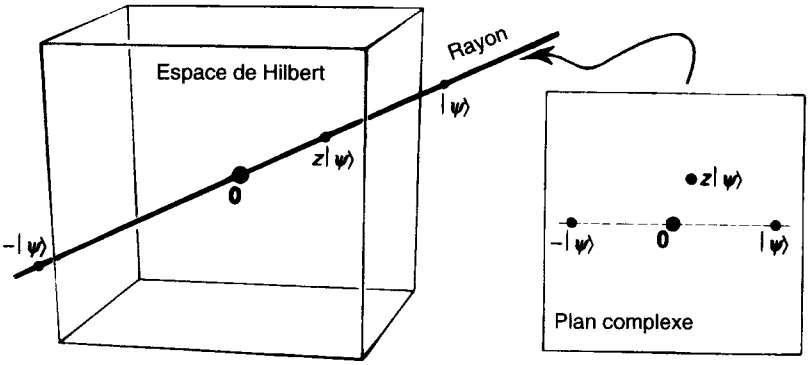


Figure 5.23. Un *rayon* de l'espace de Hilbert est l'ensemble de tous les multiples complexes d'un vecteur d'état $|\psi\rangle$. Si on peut représenter ce rayon par une droite passant par l'origine **0** de l'espace de Hilbert, il faut toutefois garder à l'esprit que cette droite est en réalité un plan complexe.

qui est essentielle à la description de la procédure de réduction **R**. Cette autre propriété est qu'ils possèdent un *produit scalaire hermitien* (ou produit *interne*). Ce produit scalaire, qui opère sur toute paire de vecteurs de l'espace de Hilbert pour donner un nombre complexe, permet d'exprimer deux choses importantes. La première est la notion de *longueur au carré* d'un vecteur de l'espace de Hilbert — c'est le produit scalaire du vecteur par *lui-même*. Un état *normalisé* (état qui, nous l'avons dit à la section 5.8, est nécessaire pour pouvoir appliquer rigoureusement la règle du module au carré) est un vecteur de l'espace de Hilbert dont la longueur au carré est égale à l'*unité*. La seconde est que le produit scalaire permet de définir la notion d'*orthogonalité* de deux vecteurs de l'espace de Hilbert — orthogonalité qui correspond à la nullité du produit scalaire des deux vecteurs concernés. On peut considérer que l'orthogonalité de deux vecteurs signifie, en un certain sens, que ces vecteurs « forment un angle droit ». En termes ordinaires, des états orthogonaux sont mutuellement *indépendants*. L'importance de ce concept en physique quantique tient au fait que les divers états résultant d'une mesure sont toujours mutuellement orthogonaux.

Comme exemple d'états orthogonaux, citons les états $|\uparrow\rangle$ et $|\downarrow\rangle$ d'une particule de spin $\frac{1}{2}$. (Remarquez que l'orthogonalité hilbertienne ne correspond pas généralement à la notion de perpendicularité dans l'espace ordinaire ; dans le cas du spin $\frac{1}{2}$, les états orthogonaux $|\uparrow\rangle$ et $|\downarrow\rangle$ représentent des configurations physiques non pas perpendiculaires, mais orientées en sens inverses.) Citons aussi les états $|\uparrow\uparrow\ldots\uparrow\rangle$, $|\downarrow\uparrow\ldots\uparrow\rangle$, ..., $|\downarrow\downarrow\ldots\downarrow\rangle$, que nous avons rencontrés lors de la discussion du spin $\frac{1}{2}n$, chacun de ces états étant orthogonal à tous les autres. Sont orthogonaux également *tous* les états de *position* possibles dans lesquels peut se trouver une particule. De même, les états $|B\rangle$ et $i|C\rangle$ (*cf.* §5.7) correspondant aux états du photon respectivement transmis et réfléchi par le miroir semi-argenté sont orthogonaux, tout comme les états $i|D\rangle$ et

$-|E\rangle$ survenant après réflexion de ces deux états par un miroir pleinement argenté.

Ce dernier fait illustre une importante propriété de l'évolution **U** donnée par l'équation de Schrödinger, à savoir que deux états quelconques initialement orthogonaux restent orthogonaux tant que chacun d'eux évolue, selon **U**, durant un même intervalle de temps. Autrement dit, **U** *préserve* l'orthogonalité. **U** préserve même la valeur du produit scalaire de deux états. En termes techniques, on dit que **U** est une représentation *unitaire*.

Nous l'avons dit, le rôle clé de l'orthogonalité est que chaque fois qu'une mesure est effectuée sur un système quantique, les divers états quantiques conduisant séparément — par amplification au niveau classique — à des résultats *discernables* sont nécessairement mutuellement orthogonaux. Cela vaut en particulier pour les mesures *à résultat nul*, comme lors du problème d'Elitzur-Vaidman (*cf.* §5.2 et §5.9). La *non*-détection d'un état quantique par un appareil capable de détecter cet état se traduit par le « saut » quantique sur un état *orthogonal* à l'état même que le détecteur est censé détecter.

Nous l'avons dit également, l'orthogonalité de deux états s'exprime mathématiquement par la *nullité* du produit scalaire de ces deux états. Ce produit scalaire est généralement un nombre complexe. Si l'on désigne par $|\psi\rangle$ et $|\phi\rangle$ les deux éléments (les deux états) de l'espace de Hilbert sur lesquels opère le produit scalaire, ce nombre complexe est alors noté $\langle\psi|\phi\rangle$. Le produit scalaire hilbertien vérifie un certain nombre de propriétés algébriques élémentaires qui s'expriment (de manière un peu rebutante) par les relations

$$\begin{aligned}
\overline{\langle\psi|\phi\rangle} &= \langle\phi|\psi\rangle, \\
\langle\psi|\,(|\phi\rangle + |\chi\rangle) &= \langle\psi|\phi\rangle + \langle\psi|\chi\rangle, \\
(z\langle\psi|)\,|\phi\rangle &= z\langle\psi|\phi\rangle \\
\langle\psi|\psi\rangle > 0 \text{ sauf si } |\psi\rangle &= \mathbf{0}.
\end{aligned}$$

En outre, on déduit de ces relations que $\langle\psi|\psi\rangle = 0$ si $|\psi\rangle = \mathbf{0}$. Je ne veux pas ennuyer le lecteur avec le détail de ces propriétés. (Ceux qui sont intéressés pourront consulter n'importe quel ouvrage standard sur la théorie quantique, par exemple Dirac 1947.)

Les choses essentielles que nous devrons retenir sur le produit scalaire sont les deux propriétés (définitions) suivantes évoquées plus haut :

$|\psi\rangle$ et $|\phi\rangle$ sont *orthogonaux* si et seulement si $\langle\psi|\phi\rangle = 0$,

$\langle\psi|\psi\rangle$ est la *longueur au carré* de $|\psi\rangle$.

Remarquons que l'orthogonalité est une relation symétrique entre $|\psi\rangle$ et $|\phi\rangle$ (car $\overline{\langle\psi|\phi\rangle} = \langle\phi|\psi\rangle$). De plus, $\langle\psi|\psi\rangle$ étant toujours un nombre non négatif, on peut donc considérer que sa racine carrée (non négative) représente la *longueur* de $|\psi\rangle$.

La multiplication d'un vecteur d'état par un nombre complexe non nul ne modifiant pas son interprétation physique, on peut toujours *normaliser* ce

vecteur d'état afin de lui donner une longueur unité, autrement dit de le transformer en un *vecteur unité* ou *état normalisé*. Reste cependant l'ambiguïté due au fait que ce vecteur d'état peut être multiplié par un nombre de la forme $e^{i\theta}$, avec θ réel (*cf.* §5.10).

5.13 Description de R dans l'espace de Hilbert

Comment représente-t-on l'action de **R** dans l'espace de Hilbert ? Considérons la plus simple des mesures, à savoir une mesure de type « oui/non » au cours de laquelle un appareil affiche **OUI** pour signaler catégoriquement que l'objet quantique mesuré possède telle propriété, et **NON** pour signaler tout aussi catégoriquement qu'il ne la possède pas. Cette situation inclut la possibilité qui nous concernera au premier chef ici, à savoir que l'éventualité **NON** peut être une mesure à *résultat nul*. Par exemple, certains détecteurs de photons de la section 5.8 effectuent une mesure de ce type. Ils affichent **OUI** s'ils reçoivent le photon et **NON** s'ils ne le reçoivent pas. Dans ce second cas, la mesure **NON** est une mesure à résultat nul. En dépit de son nom, une mesure à résultat nul n'en est pas moins une mesure : elle provoque un « saut » de l'état concerné dans quelque chose qui est orthogonal à ce que cet état serait devenu si la réponse obtenue avait été **OUI**. De même, les mesures de type Stern-Gerlach présentées à la section 5.10, mais effectuées sur un atome de spin $\frac{1}{2}$, seraient de pures mesures de type oui/non : on dirait que le résultat est **OUI** si le spin mesuré est $|\uparrow\rangle$, ce qui se produit si l'appareil trouve l'atome dans le faisceau correspondant à $|\uparrow\rangle$, et **NON** si l'appareil ne le trouve pas dans ce faisceau, auquel cas le résultat de la mesure est nécessairement orthogonal à $|\uparrow\rangle$, autrement dit est nécessairement $|\downarrow\rangle$.

On peut toujours décomposer des mesures plus complexes en une succession de mesures de type oui/non. Considérons par exemple un atome de spin $\frac{1}{2}n$. Pour obtenir les $n+1$ différentes possibilités associées à la mesure de ce spin dans la direction *up*, on regarde d'abord si l'état de spin est $|\uparrow\uparrow\ldots\uparrow\rangle$, *i.e.* on regarde si l'atome se trouve dans le faisceau correspondant à l'état de spin « complètement *up* ». Si l'on obtient **OUI**, on a terminé ; mais si l'on obtient **NON**, on a alors une mesure à résultat nul et on regarde ensuite si le spin est $|\downarrow\uparrow\ldots\uparrow\rangle$, puis $|\downarrow\downarrow\ldots\uparrow\rangle$ et ainsi de suite. Dans chaque cas, la réponse **NON** est une mesure à résultat nul indiquant simplement que la réponse **OUI** n'a pas été obtenue. De manière plus détaillée, supposons que l'état initial soit

$$z_0|\uparrow\uparrow\uparrow\ldots\uparrow\rangle + z_1|\downarrow\uparrow\uparrow\ldots\uparrow\rangle + z_2|\downarrow\downarrow\uparrow\ldots\uparrow\rangle + \ldots + z_n|\downarrow\downarrow\downarrow\ldots\downarrow\rangle,$$

et demandons-nous si le spin est entièrement *up*. Si la réponse est **OUI**, nous sommes alors certains que l'état de spin est $|\uparrow\uparrow\uparrow\ldots\uparrow\rangle$, ou plus exactement, nous pouvons considérer qu'il a « sauté » sur $|\uparrow\uparrow\uparrow\ldots\uparrow\rangle$ au moment de la

mesure. Si en revanche la réponse est **NON**, nous devons considérer que l'état a « sauté » dans l'état orthogonal

$$z_1|\downarrow\uparrow\uparrow...\uparrow\rangle + z_2|\downarrow\downarrow\uparrow...\uparrow\rangle + ... + z_n|\downarrow\downarrow\downarrow...\downarrow\rangle,$$

auquel cas nous regardons si l'état est $|\downarrow\uparrow\uparrow...\uparrow\rangle$. Si *maintenant* nous obtenons **OUI**, nous dirons que l'état est effectivement $|\downarrow\uparrow\uparrow...\uparrow\rangle$, ou qu'il a « sauté » sur $|\downarrow\uparrow\uparrow...\uparrow\rangle$. Mais si la réponse est **NON**, nous dirons qu'il a alors « sauté » sur

$$z_2|\downarrow\downarrow\uparrow...\uparrow\rangle + ... + z_n|\downarrow\downarrow\downarrow...\downarrow\rangle,$$

et ainsi de suite.

Ce « saut » auquel se livre le vecteur d'état — du moins auquel il *semble* se livrer — est l'élément le plus déconcertant de la théorie quantique. Et il faut bien avouer que la plupart des physiciens quantiques soit juge *très difficile* d'y voir une caractéristique de la réalité physique, soit refuse *catégoriquement* de croire que la réalité puisse se comporter d'une manière aussi absurde. Il n'en reste pas moins que ce « saut » est une caractéristique du formalisme quantique, quel que soit le degré de réalité que l'on accepte de lui accorder.

Les descriptions précédentes reposent sur ce que l'on nomme parfois le *postulat des projections*. Ce postulat définit la forme prise par ce « saut » (*e.g.* $z_0|\uparrow\uparrow\uparrow...\uparrow\rangle + z_1|\downarrow\uparrow\uparrow...\uparrow\rangle + z_2|\downarrow\downarrow\uparrow...\uparrow\rangle + ... + z_n|\downarrow\downarrow\downarrow...\downarrow\rangle$ « saute » sur $z_1|\downarrow\uparrow\uparrow...\uparrow\rangle + z_2|\downarrow\downarrow\uparrow...\uparrow\rangle + ... + z_n|\downarrow\downarrow\downarrow...\downarrow\rangle$). Nous verrons dans un instant la raison géométrique de cette terminologie. Certains physiciens affirment que ce postulat est une hypothèse superflue dans le cadre de la théorie quantique. Toutefois, ils se réfèrent alors généralement non pas à une mesure à résultat nul, mais à une mesure provoquant une *perturbation* de l'état quantique par une interaction physique. Une telle perturbation survient lorsque, dans les exemples précédents, on obtient la réponse **OUI** — lorsqu'un détecteur de photons absorbe un photon dans l'état recherché ou lorsqu'après avoir traversé un dispositif de Stern-Gerlach, un atome s'avère être réellement dans un faisceau particulier (*i.e.* **OUI**). Le postulat des projections est en fait essentiel aux mesures à résultat nul (réponse **NON**), car sans lui, on ne peut vérifier les prédictions (correctes) de la théorie quantique sur les mesures effectuées par la suite.

Je vais maintenant préciser ce que recouvre ce postulat en donnant la description hilbertienne du processus de mesure. Envisageons une mesure — que j'appellerai mesure *primitive* — de type oui/non, mais pour laquelle la réponse **OUI** garantit que l'état quantique est — ou vient de « sauter » dans — un état particulier $|\alpha\rangle$ (ou dans l'un de ses multiples non nuls $u|\alpha\rangle$). Ainsi, lors d'une mesure primitive, la réponse **OUI** établit que l'état physique est *une* chose précise, tandis que la réponse **NON** ouvre la voie à plusieurs éventualités. Les mesures de spin mentionnées à l'instant et qui tentent de vérifier si le spin est dans un état particulier, disons $|\downarrow\downarrow\uparrow...\uparrow\rangle$, sont des exemples de mesures primitives.

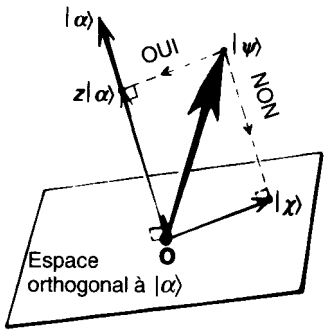


Figure 5.24. Une mesure primitive projette l'état $|\psi\rangle$ sur un multiple de l'état choisi $|\alpha\rangle$ (**OUI**) ou dans l'espace orthogonal à $|\alpha\rangle$ (**NON**).

Lors d'une mesure primitive, la réponse **NON** *projette* l'état sur quelque chose qui est orthogonal à $|\alpha\rangle$. La figure 5.24 illustre géométriquement cette opération de projection. $|\psi\rangle$ est l'état original, représenté par la flèche grasse ; lors de la mesure, soit cet état « saute » sur un multiple de $|\alpha\rangle$ — lorsque la réponse est **OUI** —, soit il se projette sur l'espace orthogonal à $|\alpha\rangle$ — lorsque la réponse est **NON**. Dans le cas **NON**, la théorie quantique standard est catégorique : le comportement de $|\psi\rangle$ est effectivement celui décrit ici. Dans le cas **OUI** en revanche, la situation se complique par le fait que le système quantique a maintenant interagi avec l'appareil de mesure, et que son état est devenu plus complexe que le simple état $|\alpha\rangle$. En fait, il se transforme généralement en ce que l'on appelle un *état quantique emmêlé*, résultant de l'interaction entre le système quantique original et l'appareil de mesure. (Les états quantiques emmêlés seront examinés à la section 5.17.) Toutefois, pour que son évolution après la mesure soit dépourvue d'ambiguïté, l'état quantique doit s'être comporté *comme s'il* avait effectivement sauté sur un multiple de $|\alpha\rangle$.

On peut exprimer ce saut de la manière algébrique suivante. Le vecteur d'état $|\psi\rangle$ peut toujours s'écrire (de façon unique si $|\alpha\rangle$ est donné) sous la forme

$$|\psi\rangle = z|\alpha\rangle + |\chi\rangle,$$

où $|\chi\rangle$ est orthogonal à $|\alpha\rangle$. Le vecteur $z|\alpha\rangle$ est la projection orthogonale de $|\psi\rangle$ sur le rayon défini par $|\alpha\rangle$, et $|\chi\rangle$ est la projection orthogonale de $|\psi\rangle$ sur l'*espace orthogonal* à $|\alpha\rangle$ (*i.e.* sur l'espace de tous les vecteurs orthogonaux à $|\alpha\rangle$). Si le résultat de la mesure est **OUI**, on considère alors que le vecteur d'état a sauté sur $z|\alpha\rangle$ (on dira plus simplement « sur $|\alpha\rangle$ ») et ce $z|\alpha\rangle$ est le point de départ de son évolution ultérieure ; si la réponse est **NON**, on considère que le vecteur d'état a sauté sur $|\chi\rangle$.

Quelles sont les probabilités de voir se réaliser chacune de ces deux éventualités ? Afin de pouvoir utiliser la « règle du module au carré », nous

devons supposer d'une part que $|\alpha\rangle$ est un *vecteur unité*, d'autre part que $|\chi\rangle = w|\phi\rangle$, où $|\phi\rangle$ est le vecteur unité dans la direction de $|\chi\rangle$. Nous avons

$$|\psi\rangle = z|\alpha\rangle + w|\phi\rangle,$$

avec $z = \langle\alpha|\psi\rangle$ et $w = \langle\phi|\psi\rangle$. Le rapport $|z|^2/|w|^2$ représente alors le rapport des probabilités des résultats **OUI** et **NON**. Si $|\psi\rangle$ est lui-même un vecteur unité, $|z|^2$ et $|w|^2$ sont respectivement les probabilités *absolues* de **OUI** et **NON**.

On peut formuler cela d'une manière un peu plus simple dans le présent contexte (et je laisse au lecteur intéressé le soin de vérifier que cette formulation est effectivement équivalente). Pour obtenir la probabilité réelle de chacun des résultats **OUI** ou **NON**, on examine la longueur au carré du vecteur $|\psi\rangle$ (qui n'est ici pas nécessairement un vecteur unité) et on regarde dans quelles proportions cette longueur au carré se trouve réduite sur chacune de ses projections respectives. Les deux coefficients de réduction que l'on obtient ainsi sont les probabilités recherchées.

Signalons pour terminer que la discussion d'une mesure *générale* (*i.e.* non nécessairement primitive) de type oui/non, dans laquelle les états **OUI** n'appartiennent pas obligatoirement à un seul rayon vecteur, est tout à fait similaire. On a alors un sous-espace **OUI** — appelons-le **O** — et un sous-espace **NON** — appelons-le **N**. Chacun de ces deux sous-espaces est le complément orthogonal de l'autre — dans le sens où chaque vecteur de l'un est orthogonal à chaque vecteur de l'autre et où ces sous-espaces engendrent à eux deux l'espace de Hilbert original. Le postulat des projections affirme alors que la mesure projette orthogonalement le vecteur original $|\psi\rangle$ sur **O** lorsqu'on obtient la réponse **OUI** et sur **N** si la réponse est **NON**. Les probabilités respectives sont encore données par les coefficients de réduction de la longueur au carré du vecteur d'état $|\psi\rangle$ (voir EOLP, p. 284, figure 6.23). Toutefois, le statut du postulat des projections est ici un peu moins clair que dans le cas des mesures à résultat nul, car lors d'une mesure positive, l'état résultant est un état quantique emmêlé, résultant de l'interaction avec l'appareil de mesure. C'est la raison pour laquelle, dans les discussions qui vont suivre, je me bornerai aux mesures *primitives*, pour lesquelles l'espace **OUI** se compose d'un seul rayon (les multiples de $|\alpha\rangle$). Cela suffira à nos besoins.

5.14 Mesures commutatives

En physique quantique, l'ordre dans lequel on effectue une succession de mesures sur un système est en général important. Les mesures pour lesquelles les vecteurs d'état résultants dépendent de l'ordre dans lequel elles sont effectuées sont dites *non commutatives*. Si cet ordre ne joue aucun rôle (n'entraînant même pas de modification des phases des coefficients), les mesures sont

dites *commutatives*. En termes d'espace de Hilbert, la non-commutativité des mesures se traduit par le fait que le vecteur d'état résultant des projections orthogonales successives d'un vecteur d'état initial $|\psi\rangle$ dépend de l'ordre dans lequel ces projections sont effectuées. Pour les mesures commutatives, cet ordre ne joue aucun rôle.

Que se passe-t-il au niveau des mesures *primitives* ? Il n'est pas difficile de voir que la condition pour que deux mesures primitives distinctes commutent est que le rayon **OUI** de l'une des mesures soit *orthogonal* au rayon **NON** de l'autre.

Par exemple, lors des mesures primitives de spin mentionnées à la section 5.10 et effectuées sur un atome de spin $\frac{1}{2}n$, l'ordre ne joue aucun rôle. La raison en est que les divers états mis en jeu, à savoir $|\uparrow\uparrow\ldots\uparrow\rangle$, $|\downarrow\uparrow\ldots\uparrow\rangle$, ..., $|\downarrow\downarrow\ldots\downarrow\rangle$ sont tous mutuellement orthogonaux. Ainsi, l'ordre dans lequel on choisit d'effectuer ces mesures primitives n'a aucune influence sur le résultat final et ces mesures commutent toutes entre elles. Il en va toutefois différemment lorsque les diverses mesures de spins sont effectuées dans des directions différentes : ces mesures, en général, *ne commutent pas*.

5.15 Le « et » quantique

Il existe en théorie quantique une procédure qui permet de traiter les systèmes comprenant plusieurs parties indépendantes. Cette procédure nous sera notamment nécessaire pour la discussion quantique (donnée à la section 5.18) du système composé des deux particules de spin $\frac{3}{2}$ que *Trucs Quintessentiels* a placées au centre des deux dodécaèdres magiques de la section 5.3. Elle intervient également, par exemple, dans la description quantique d'un détecteur lorsque celui-ci commence à interagir avec l'état quantique de la particule qu'il est en train de détecter.

Considérons tout d'abord un système composé uniquement de *deux* parties indépendantes (non interactives). Supposons que ces deux parties, isolées l'une de l'autre, soient respectivement décrites par les vecteurs d'état $|\alpha\rangle$ et $|\beta\rangle$. Comment décrire le système résultant de la mise en présence de ces *deux* parties ? La procédure habituelle consiste à former ce que l'on appelle le *produit tensoriel* (ou produit *extérieur*) des deux vecteurs d'état $|\alpha\rangle$ et $|\beta\rangle$, à savoir

$$|\alpha\rangle|\beta\rangle .$$

Ce produit tensoriel est l'équivalent quantique du concept ordinaire « et », dans le sens où les deux systèmes quantiques indépendants respectivement associés à $|\alpha\rangle$ et $|\beta\rangle$ sont maintenant *tous deux* simultanément présents. (Par exemple, $|\alpha\rangle$ peut représenter un électron occupant la position A et $|\beta\rangle$ un atome d'hydrogène situé à un point B distant de A. L'état dans lequel l'électron est en A *et* l'atome d'hydrogène est en B est alors représenté par $|\alpha\rangle|\beta\rangle$.)

La quantité $|\alpha\rangle|\beta\rangle$ forme toutefois un *seul* vecteur d'état quantique. Appelons-le $|\chi\rangle$. On a donc

$$|\chi\rangle = |\alpha\rangle|\beta\rangle.$$

Soulignons que ce concept « et » diffère radicalement de celui de superposition linéaire quantique. Celle-ci correspondrait en l'occurrence à $|\alpha\rangle + |\beta\rangle$ ou, d'une manière plus générale, à $z|\alpha\rangle + w|\beta\rangle$, où z et w sont des coefficients de pondération complexes. Par exemple, si $|\alpha\rangle$ et $|\beta\rangle$ sont les états possibles d'un photon correspondant respectivement à sa localisation en deux endroits tout à fait différents A et B, $|\alpha\rangle + |\beta\rangle$ est alors un état possible pour un *seul* photon dont la position se répartit entre A et B selon les étranges prescriptions de la théorie quantique — et non un état associé à *deux* photons. Une *paire* de photons situés l'un en A et l'autre en B serait représentée par l'état $|\alpha\rangle|\beta\rangle$.

Le produit tensoriel vérifie les règles algébriques que l'on attend habituellement d'un « produit », à savoir

$$\begin{aligned}
(z|\alpha\rangle)\,|\beta\rangle &= z\,(|\alpha\rangle|\beta\rangle) = |\alpha\rangle z|\beta\rangle,\\
(|\alpha\rangle + |\gamma\rangle)\,|\beta\rangle &= |\alpha\rangle|\beta\rangle + |\gamma\rangle|\beta\rangle,\\
|\alpha\rangle\,(|\beta\rangle + |\gamma\rangle) &= |\alpha\rangle|\beta\rangle + |\alpha\rangle|\gamma\rangle,\\
(|\alpha\rangle|\beta\rangle)\,|\gamma\rangle &= |\alpha\rangle\,(|\beta\rangle|\gamma\rangle),
\end{aligned}$$

sauf qu'il n'est pas strictement correct d'écrire « $|\alpha\rangle|\beta\rangle = |\beta\rangle|\alpha\rangle$ ». Toutefois, il ne serait pas raisonnable d'attribuer au « et » quantique un sens qui impliquerait que le système combiné « $|\alpha\rangle$ et $|\beta\rangle$ » *est* physiquement différent du système combiné « $|\beta\rangle$ et $|\alpha\rangle$ ». Nous contournerons ce problème en analysant plus en profondeur le comportement de la nature au niveau quantique. Au lieu d'interpréter l'état $|\alpha\rangle|\beta\rangle$ comme un « produit tensoriel » au sens mathématique, je supposerai à partir de maintenant que la notation « $|\alpha\rangle|\beta\rangle$ » représente ce que les physiciens théoriciens désignent aujourd'hui sous le nom de *produit de Grassmann*. Nous avons alors la règle supplémentaire

$$|\beta\rangle|\alpha\rangle = \pm|\alpha\rangle|\beta\rangle,$$

où le signe « moins » apparaît lorsque les *deux* états $|\alpha\rangle$ et $|\beta\rangle$ possèdent un nombre *impair* de particules de spin non entier. (Le spin de telles particules a alors l'une des valeurs $\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$, $\frac{7}{2}$, ..., et ces particules sont appelées *fermions*. Les particules de spin 0, 1, 2, 3, ..., sont appelées *bosons* et ne contribuent pas au signe de cette expression.) Que le lecteur ne s'inquiète pas de ces aspects techniques : en tant qu'états physiques, « $|\alpha\rangle$ et $|\beta\rangle$ » et « $|\beta\rangle$ et $|\alpha\rangle$ » sont, dans cette description, identiques.

Pour décrire les états comportant plus de deux composantes indépendantes, il suffit de répéter cette procédure. Ainsi, l'état associé à trois composantes d'états respectifs $|\alpha\rangle$, $|\beta\rangle$ et $|\gamma\rangle$ est

$$|\alpha\rangle|\beta\rangle|\gamma\rangle,$$

cette notation désignant pareillement (au sens de produits de Grassmann) $(|\alpha\rangle|\beta\rangle)\,|\gamma\rangle$ ou $|\alpha\rangle\,(|\beta\rangle|\gamma\rangle)$. Le cas de quatre — et plus — composantes indépendantes se traite de manière analogue.

Une importante propriété de la représentation de Schrödinger **U** est que, pour des systèmes $|\alpha\rangle$ et $|\beta\rangle$ sans interactions mutuelles, la représentation de leur combinaison $|\alpha\rangle|\beta\rangle$ est simplement la combinaison des représentations des systèmes individuels. Ainsi, après un certain temps t, si le système $|\alpha\rangle$ s'est transformé (suivant son évolution propre) en $|\alpha'\rangle$ et si le système $|\beta\rangle$ s'est transformé (suivant son évolution propre) en $|\beta'\rangle$, le système combiné $|\alpha\rangle|\beta\rangle$ s'est transformé, après le même temps t, en $|\alpha'\rangle|\beta'\rangle$. De même, si les trois composantes $|\alpha\rangle$, $|\beta\rangle$ et $|\gamma\rangle$ d'un système $|\alpha\rangle|\beta\rangle|\gamma\rangle$ se transforment respectivement en $|\alpha'\rangle$, $|\beta'\rangle$ et $|\gamma'\rangle$, le système combiné se transforme en $|\alpha'\rangle|\beta'\rangle|\gamma'\rangle$. Le même raisonnement vaut pour quatre — ou plus — composantes.

Remarquez que tout cela est très similaire à la propriété de *linéarité* de **U** mentionnée à la section 5.7 et selon laquelle l'évolution d'une superposition d'états est donnée par la superposition des évolutions des états individuels. Par exemple, $|\alpha\rangle + |\beta\rangle$ deviendrait $|\alpha'\rangle + |\beta'\rangle$. Il importe toutefois de remarquer que dans le cas d'un système combiné, la situation est tout à fait *différente*. Il n'y a rien de surprenant dans le fait qu'un système constitué de composantes indépendantes non interactives évolue globalement comme si chaque composante ignorait la présence des autres. Pour cela, il est effectivement essentiel que ces composantes n'interagissent pas entre elles, sinon la propriété serait fausse. Ce qui est surprenant, en revanche, c'est la linéarité. Ici, selon **U**, les composantes du système évoluent dans l'ignorance totale les unes des autres, absolument *indépendamment* de la présence ou de l'absence d'une interaction. Ce seul fait peut nous conduire à mettre en question la validité absolue de la propriété de linéarité. Pourtant, elle est très bien confirmée pour les phénomènes quantiques. C'est uniquement l'opération **R** qui semble la violer. Nous aurons à revenir sur ce sujet.

5.16 Orthogonalité des états produits

La notion d'orthogonalité soulève quelques difficultés lorsqu'on veut l'appliquer aux états produits définis à la section précédente. Si l'on a deux états $|\alpha\rangle$ et $|\beta\rangle$ *orthogonaux*, on pourrait penser que les états $|\psi\rangle|\alpha\rangle$ et $|\psi\rangle|\beta\rangle$ sont eux aussi orthogonaux quel que soit $|\psi\rangle$. Par exemple, $|\alpha\rangle$ et $|\beta\rangle$ pourraient être deux états accessibles à un photon, l'état $|\alpha\rangle$ étant celui détecté par une cellule photoélectrique, tandis que l'état orthogonal $|\beta\rangle$ serait l'état *déduit* pour le photon lorsque la cellule photoélectrique ne détecte rien (mesure à résultat nul). Supposons maintenant que le photon fasse partie d'un système *combiné*, comprenant par exemple un autre photon situé sur la Lune et dont

l'état est représenté par un vecteur $|\psi\rangle$. Les deux possibilités pour le système total sont alors $|\psi\rangle|\alpha\rangle$ et $|\psi\rangle|\beta\rangle$, et il semble que la simple inclusion de l'état $|\psi\rangle$ dans cette description ne doive pas modifier l'orthogonalité des deux états $|\alpha\rangle$ et $|\beta\rangle$. Il en serait effectivement ainsi avec la définition du « produit tensoriel » ordinaire (au contraire d'une forme de produit de Grassmann utilisée ici), et l'orthogonalité de $|\alpha\rangle$ et $|\beta\rangle$ entraînerait l'orthogonalité de $|\psi\rangle|\alpha\rangle$ et $|\psi\rangle|\beta\rangle$.

Toutefois, la nature ne semble pas se comporter aussi simplement que le voudrait la théorie quantique. Si l'état $|\psi\rangle$ pouvait être considéré comme totalement indépendant de $|\alpha\rangle$ et $|\beta\rangle$, sa présence ne jouerait effectivement aucun rôle. Mais techniquement parlant, l'état d'un photon, fût-il sur la Lune, n'est jamais totalement dissocié de celui mis en jeu lors de la détection par la cellule photoélectrique*. (Cela est lié au fait que le produit « $|\psi\rangle|\alpha\rangle$ » est de Grassmann. En termes plus familiers, cela est lié à la « statistique de Bose-Einstein » associée aux états photoniques et, plus généralement, aux états bosoniques, ou à la « statistique de Fermi-Dirac » associée aux états d'électrons, de protons et autres fermions ; *cf.* EOLP, p. 301-303, 504 (note 23) ; *cf.* aussi, par exemple, Dirac 1947.) Si l'on voulait respecter scrupuleusement les règles de la théorie quantique, on devrait, lors de la discussion de l'état d'un seul photon, prendre en compte tous les photons de l'Univers. Fort heureusement, cela n'est pas nécessaire, même à une très haute précision de mesure. Nous admettrons que si un état $|\psi\rangle$ n'a pratiquement rien à voir avec un système physique décrit par deux états orthogonaux $|\alpha\rangle$ et $|\beta\rangle$, les états $|\psi\rangle|\alpha\rangle$ et $|\psi\rangle|\beta\rangle$ sont eux-mêmes orthogonaux (même si ces produits sont des produits de Grassmann).

5.17 L'emmêlement quantique

Nous voulons comprendre la physique quantique gouvernant les *effets EPR* — les énigmes-**Y** illustrées par les dodécaèdres magiques de la section 5.3 (*cf.* §5.4). Nous souhaitons également éclaircir l'énigme-**X** fondamentale de la théorie quantique, à savoir la relation paradoxale existant entre les deux processus **U** et **R** sous-jacents au *problème de la mesure* — problème qui fera l'objet du prochain chapitre. Pour parvenir à ces deux objectifs, il me faut introduire une autre notion importante : l'*emmêlement quantique.*

Examinons tout d'abord les éléments intervenant lors d'un processus de mesure élémentaire. Considérons un photon dans une superposition d'états

* Curieusement, ce phénomène peut avoir un lien profond avec les observations réelles. La méthode de Hanbury Brown et Twiss (1954, 1956), qui a permis de mesurer le diamètre de certaines étoiles proches, dépend du caractère « bosonique » des photons atteignant la Terre depuis des points opposés de l'étoile !

$|\alpha\rangle + |\beta\rangle$ telle que l'état $|\alpha\rangle$ active un détecteur, tandis que l'état $|\beta\rangle$, orthogonal à $|\alpha\rangle$, n'interagit pas avec ce détecteur. (Nous avons rencontré un exemple de cette situation à la section 5.8 : lorsque le détecteur G recevait l'état $|F\rangle - i|G\rangle$, il réagissait à $|G\rangle$ et restait insensible à $|F\rangle$.) Je vais supposer que l'on puisse également attribuer un état quantique, disons $|\Psi\rangle$, à ce détecteur. Bien que cette pratique soit tout à fait courante en théorie quantique et ne soit habituellement pas mise en question dans les discussions de ce genre, je dois admettre que je ne vois pas clairement ce que peut recouvrir une description quantique d'un objet classique. On peut toujours supposer que les éléments du détecteur *initialement* rencontrés par le photon peuvent effectivement être traités selon les règles standard de la théorie quantique. Ceux qui doutent que l'on puisse procéder ainsi peuvent considérer que le vecteur $|\Psi\rangle$ correspond à l'état des éléments quantiques (particules, atomes, molécules) initialement rencontrés par le photon.

Juste avant que le photon n'atteigne le détecteur (ou, plus exactement, juste avant que la composante $|\alpha\rangle$ de la fonction d'onde du photon n'atteigne le détecteur), la situation physique se compose de l'état du détecteur *et* de l'état du photon, autrement dit de $|\Psi\rangle\,(|\alpha\rangle + |\beta\rangle)$. On a alors :

$$|\Psi\rangle\,(|\alpha\rangle + |\beta\rangle) \;=\; |\Psi\rangle|\alpha\rangle + |\Psi\rangle|\beta\rangle\,.$$

C'est une superposition de l'état $|\Psi\rangle|\alpha\rangle$, qui décrit le détecteur (plus précisément, ses éléments) et le photon en phase d'approche, et de l'état $|\Psi\rangle|\beta\rangle$, qui décrit le détecteur (plus précisément, ses éléments) et le photon ailleurs. Supposons ensuite qu'en vertu de l'évolution **U** donnée par l'équation de Schrödinger, l'état $|\Psi\rangle|\alpha\rangle$ (détecteur plus photon en phase d'approche) se transforme en un nouvel état $|\Psi_O\rangle$ (indiquant que le détecteur a affiché la réponse **OUI**) à la suite des interactions survenues entre le photon et les éléments du détecteur après la rencontre du premier avec les seconds. Supposons également que si le photon *ne rencontre pas* le détecteur, l'action de **U** place l'état du détecteur $|\Psi\rangle$ dans un état $|\Psi_N\rangle$ (le détecteur affiche **NON**) et que $|\beta\rangle$ se transforme en $|\beta'\rangle$. Alors, en vertu des propriétés de l'évolution schrödingerienne mentionnées à la section précédente, l'état total devient

$$|\Psi_O\rangle + |\Psi_N\rangle|\beta'\rangle\,.$$

C'est là un exemple particulier d'état quantique *emmêlé*, où l'« emmêlement » quantique traduit le fait que l'état total ne peut s'écrire comme le *produit* de l'état associé à l'un des sous-systèmes (le photon) par l'état associé à l'autre sous-système (le détecteur). En fait, l'état $|\Psi_O\rangle$ est probablement lui-même un état quantique emmêlé avec son propre environnement, mais cela dépend du détail d'autres interactions — qui ne nous concernent pas ici.

Notons que si $|\Psi\rangle$ est totalement indépendant de $|\alpha\rangle$ et de $|\beta\rangle$, les états $|\Psi\rangle|\alpha\rangle$ et $|\Psi\rangle|\beta\rangle$, dont la superposition représente l'état du système combiné juste avant l'interaction, sont (essentiellement) *orthogonaux* — car $|\alpha\rangle$ et $|\beta\rangle$ sont eux-mêmes orthogonaux. Ainsi, les états $|\Psi_O\rangle$ et $|\Psi_N\rangle|\beta'\rangle$ que deviennent respectivement $|\Psi\rangle|\alpha\rangle$ et $|\Psi\rangle|\beta\rangle$ sous l'action de **U** sont eux-mêmes

orthogonaux. (**U** conserve toujours l'orthogonalité.) L'état $|\Psi_O\rangle$ peut alors évoluer pour devenir quelque chose de macroscopiquement observable, par exemple un « clic » audible indiquant que le photon a effectivement été détecté, tandis que l'absence de clic signifie que l'état est devenu — a « sauté » sur — l'état orthogonal $|\Psi_N\rangle|\beta'\rangle$. En l'absence de clic, la simple *éventualité* « contrafactuelle » de son occurrence provoque un « saut » de l'état sur $|\Psi_N\rangle|\beta'\rangle$ qui n'est plus maintenant un état quantique emmêlé. La mesure à résultat nul a mis fin à l'interaction de l'état avec l'appareil de mesure.

L'un des traits caractéristiques des états quantiques emmêlés est que le « saut » associé à l'opération **R** peut avoir une action apparemment non locale (voire apparemment rétroactive) qui est encore plus déconcertante que celle d'une simple mesure à résultat nul. Cette non-localité accompagne notamment les effets baptisés « effets EPR » — pour Einstein-Podolsky-Rosen. Ces effets sont d'authentiques mystères quantiques qui comptent parmi les plus embarrassantes des énigmes-**Y** de la théorie. Ils furent envisagés pour la première fois par Einstein, alors qu'il tentait de montrer que le formalisme de la théorie quantique ne fournissait probablement pas une description complète de la nature. On a depuis proposé quantité de versions du phénomène EPR (telle celle des dodécaèdres magiques présentée à la section 5.3) ; nombre d'entre elles ont été directement confirmées par l'expérience, montrant ainsi que ce phénomène correspond aux mécanismes *réels* du monde dans lequel nous vivons (*cf.* §5.4).

Les effets EPR surviennent dans les situations suivantes. Considérons un système physique dont l'état initial $|\Omega\rangle$, connu, se transforme, sous l'action de l'évolution **U**, en une superposition de deux états orthogonaux dont chacun est lui-même le produit de deux états indépendants décrivant deux systèmes physiques séparés spatialement, superposition que nous écrirons sous la forme

$$|\psi\rangle|\alpha\rangle + |\phi\rangle|\beta\rangle .$$

Nous supposerons que $|\psi\rangle$ et $|\phi\rangle$ d'une part, $|\alpha\rangle$ et $|\beta\rangle$ d'autre part, sont respectivement les états orthogonaux décrivant le premier et le second de ces deux systèmes. Une mesure qui montre que le premier système est dans l'état $|\psi\rangle$ ou $|\phi\rangle$ détermine instantanément que le second est dans l'état correspondant $|\alpha\rangle$ ou $|\beta\rangle$.

Jusqu'ici, il n'y a rien de mystérieux. La situation est tout à fait semblable à celle des chaussettes du bon Dr Bertlmann (§5.4). Si l'on sait que ses deux chaussettes sont toujours de couleurs différentes et si, par exemple, on sait qu'aujourd'hui il a choisi de porter une chaussette rose et une verte, il en résulte que, selon que l'on constate que sa chaussette gauche est verte (état $|\psi\rangle$) ou rose (état $|\phi\rangle$), on en déduit immédiatement que sa chaussette droite est, respectivement, rose (état $|\alpha\rangle$) ou verte (état $|\beta\rangle$). Les effets de l'emmêlement quantique peuvent toutefois être profondément différents et leurs manifestations observationnelles n'admettre aucune explication de type « chaussettes de Bertlmann ». Les problèmes apparaissent lorsqu'on a le choix d'effectuer des mesures correspondant à différentes *alternatives* sur les deux parties du système.

À titre d'exemple, supposons que l'état initial $|\Omega_0\rangle$ décrive l'état de spin d'une particule de spin 0 et que cette particule se désintègre en deux nouvelles particules, chacune de spin $\frac{1}{2}$, s'éloignant ensuite considérablement l'une de l'autre — disons vers la droite et vers la gauche. En vertu des propriétés de conservation du moment cinétique, les orientations des spins des deux particules sont mutuellement *inverses*, et l'état de spin nul $|\Omega_0\rangle$ se transforme en

$$|\Omega\rangle = |G\uparrow\rangle|D\downarrow\rangle - |G\downarrow\rangle|D\uparrow\rangle,$$

où « G » se réfère à la particule de gauche, « D » à la particule de droite (et où le signe « moins » exprime les conventions standard). Ainsi, si l'on choisit de mesurer le spin de la particule de gauche dans l'orientation *up*, la réponse **OUI** (*i.e.* l'affichage de $|G\uparrow\rangle$) place automatiquement la particule de droite dans l'état d'orientation *down* $|D\downarrow\rangle$. La réponse **NON** ($|G\downarrow\rangle$ placerait automatiquement la particule de droite dans l'état de spin *up* ($|D\uparrow\rangle$). Il semble qu'une mesure sur l'une des deux particules affecte instantanément l'état d'une particule tout à fait différente située en un lieu tout à fait différent — mais jusqu'ici, cela n'est pas plus mystérieux que les chaussettes de Bertlmann !

On peut cependant représenter notre état quantique emmêlé sous une autre forme, correspondant à un choix de mesure différent. Par exemple, on peut décider de mesurer le spin de la particule de gauche dans une direction, cette fois, *horizontale*, de sorte que **OUI** correspond, disons, à $|G\leftarrow\rangle$ et **NON** à $|G\rightarrow\rangle$. Un calcul élémentaire (*cf.* EOLP, p. 288) montre que le *même* état $|\Omega\rangle$ peut s'écrire :

$$|\Omega\rangle = |G\leftarrow\rangle|D\rightarrow\rangle - |G\rightarrow\rangle|D\leftarrow\rangle.$$

On trouve ainsi que la réponse **OUI** sur la particule de gauche place automatiquement la particule de droite dans l'état $|D\rightarrow\rangle$, et que la réponse **NON** à gauche place la particule de droite dans l'état $|D\leftarrow\rangle$. Un phénomène analogue s'observerait *quelle que soit* la direction choisie pour mesurer le spin de la particule de gauche.

Cette situation a ceci de remarquable que la simple *connaissance* de la direction de mesure du spin sur la particule de gauche semble *fixer* la direction du spin de la particule de droite. En fait, tant que l'on *n'effectue pas* de mesure sur la particule de gauche, aucune information n'est transmise à la particule de droite. La connaissance de la direction de l'axe du spin ne suscite, à elle seule, aucun événement observable. Bien que ce soit là un fait admis, on rencontre néanmoins de temps en temps des gens qui affirment que le processus **R** « réduisant » simultanément l'état quantique des deux particules EPR — quelle que soit la distance qui les sépare —, on pourrait utiliser les effets EPR pour transmettre *instantanément* des signaux d'un lieu à un autre. Or il n'y a en fait aucun moyen d'envoyer, par cette procédure, un signal de la particule de gauche à la particule de droite (*cf.* Ghirardi *et al.* 1980).

Selon le formalisme standard de la théorie quantique, dès que l'on fait une mesure sur l'une des deux particules, disons celle de gauche, l'état entier se trouve instantanément réduit et passe de l'état quantique emmêlé initial —

dans lequel aucune des deux particules ne possède *en elle-même* un état de spin défini — à une situation dans laquelle les états de spin des particules de gauche et de droite se trouvent « démêlés » et prennent une valeur bien définie. Dans la description *mathématique* du vecteur d'état, la mesure sur la particule de gauche a un effet instantané sur la particule de droite. Mais comme je l'ai indiqué, cet « effet instantané » ne permet pas de transmettre un signal physique.

Selon les principes de la relativité, les signaux physiques — capables de véhiculer une information — se déplacent au mieux à la vitesse de la lumière. Ces principes, toutefois, ne s'appliquent pas aux effets EPR. Il ne serait pas cohérent avec les prédictions de la théorie quantique de traiter ces effets comme des signaux se propageant à vitesse finie, au plus égale à la vitesse de la lumière. (L'exemple des dodécaèdres magiques est une illustration de ce fait : les emmêlements existant entre mon dodécaèdre et celui de mon collègue ont des actions immédiates et n'attendent pas les quatre années que mettrait un signal pour passer entre nous ; *cf.* §5.3, §5.4, et la note 4 du présent chapitre.) Ainsi, les effets EPR ne sont pas des signaux au sens ordinaire du terme.

On peut dès lors se demander pourquoi ils ont des conséquences observables, ainsi qu'il s'ensuit du fameux théorème de John Bell (*cf.* §5.4). Aucun modèle classique d'objets sans interaction ne peut rendre compte des probabilités combinées prédites par la théorie quantique pour diverses mesures pouvant être effectuées sur nos deux particules de spin $\frac{1}{2}$ — avec des choix indépendants pour la direction du spin, tant pour la particule de droite que pour celle de gauche. (Voir EOLP, p. 310 et p. 504 (note 24).) Les exemples analogues à celui des dodécaèdres magiques de la section 5.3 sont encore plus impressionnants, car on a affaire ici non à de simples probabilités, mais à des contraintes précises s'exprimant par des alternatives oui/non. Ainsi, bien que les particules de droite et de gauche ne soient pas en communication dans le sens où elles ne peuvent s'envoyer instantanément des messages, elles sont cependant encore *emmêlées* l'une avec l'autre dans le sens où l'on ne peut les considérer comme des objets distincts et indépendants — jusqu'à ce qu'elles se trouvent « démêlées » par une mesure. L'emmêlement quantique est un phénomène mystérieux qui se situe quelque part entre la communication directe et la séparation totale — et n'admet aucun équivalent classique. En outre, il ne s'atténue pas avec la distance (contrairement, par exemple, à la loi « en inverse du carré » de la gravitation ou de l'attraction électrique). Einstein jugeait profondément troublante la perspective d'un tel effet, le qualifiant d'« effrayante action à distance » (voir Mermin 1985).

En fait, l'emmêlement quantique semble ignorer totalement non seulement la séparation spatiale, mais aussi la séparation temporelle. Selon la description quantique standard, si l'on effectue une mesure sur une composante d'une paire EPR *avant* de l'effectuer sur l'autre, la première mesure détruit l'emmêlement entre ces deux composantes, de sorte que la seconde mesure opère sur une composante qui est de fait non emmêlée. Toutefois, on observerait exactement les mêmes conséquences si l'on considérait que c'est la *seconde* mesure — et non la première — qui, de quelque façon, détruit rétroactivement l'emmêlement. On peut exprimer autrement cette indépen-

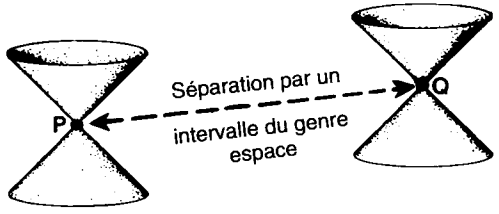


Figure 5.25. Deux événements spatio-temporels sont dits séparés par un intervalle du genre espace si chacun d'eux se trouve hors du cône de lumière de l'autre (voir aussi la figure 4.1). Aucun de ces deux événements ne peut alors influencer l'autre, et les mesures réalisées sur chacun d'eux commutent.

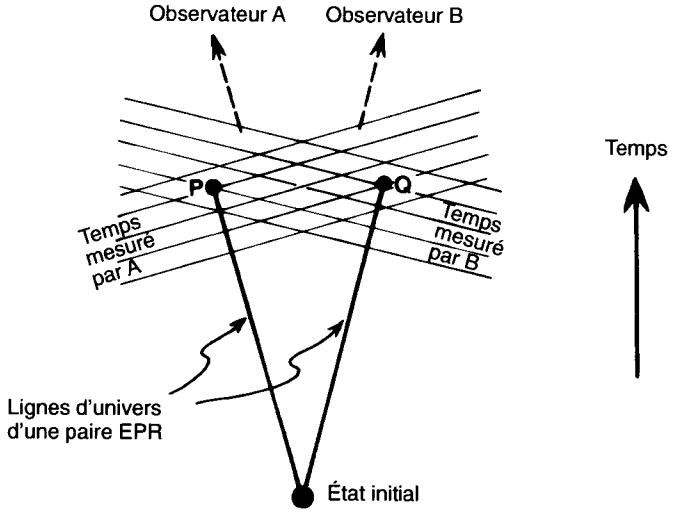


Figure 5.26. Selon la relativité restreinte, les observateurs A et B, en mouvement relatif l'un par rapport à l'autre, perçoivent différemment la chronologie de deux événements P et Q séparés par un intervalle du genre espace (A pense que Q s'est produit en premier, tandis que B pense que P a précédé Q).

dance par rapport à l'ordre temporel en disant que ces deux mesures *commutent* (*cf.* §5.14).

Sans cette symétrie, les mesures EPR contrediraient les conséquences observables de la relativité restreinte. Les mesures effectuées sur des événements séparés par des intervalles du genre espace (*i.e.* sur des événements dont chacun est situé à l'extérieur du cône de lumière de l'autre ; voir la figure 5.25 et la discussion de la section 4.4) commutent *nécessairement* et peu importe alors quelle est celle qui a agi « en premier » — conformément aux principes fondamentaux de la relativité restreinte. Pour voir qu'il en est effectivement ainsi, on peut considérer (Fig. 5.26) la situation physique tout entière telle qu'elle

est décrite dans des systèmes de référence associés à deux observateurs différents (voir aussi EOLP, p. 312). (Ces deux « observateurs » n'ont aucun lien avec ceux qui effectuent les mesures.) Les deux observateurs ont des avis différents sur celle des deux mesures qui a été réalisée en « premier ». Au niveau des mesures de type EPR, le phénomène d'emmêlement quantique — ou, en l'occurrence, de « *dé*mêlement »* — est indifférent à la séparation spatiale et à l'ordre temporel !

5.18 Les dodécaèdres magiques : solution

Pour la paire de particules de spin $\frac{1}{2}$ considérée dans l'expérience de pensée EPR (*cf.* §5.4), la non-localité spatiale ou temporelle intervient uniquement au niveau des *probabilités*. L'emmêlement quantique est toutefois un phénomène bien plus précis et concret qu'une simple « influence sur des probabilités ». Il existe des exemples — tel celui des dodécaèdres magiques (et certaines configurations proposées antérieurement[10]) — qui montrent que l'étrange non-localité de l'emmêlement quantique *n'est pas* une simple affaire de probabilités, mais fournit également des effets précis de type oui/non inexplicables classiquement de façon locale.

Examinons maintenant la mécanique quantique sous-jacente aux dodécaèdres magiques de la section 5.3. Rappelons que, sur Bételgeuse, la société *Trucs Quintessentiels* a délicatement placé au centre de chaque dodécaèdre un atome de spin $\frac{3}{2}$ provenant de la scission en deux d'un atome de spin 0 (c'est l'état initial $|\Omega\rangle$). Ces dodécaèdres nous ont ensuite été soigneusement expédiés, à mon collègue sur α du Centaure et à moi-même sur Terre, de sorte que les états de spin des deux atomes n'ont subi aucune modification durant le trajet — jusqu'à ce que mon collègue ou moi procédions à une mesure de spin sur nos atomes respectifs en pressant l'un des boutons situés sur les sommets de nos dodécaèdres. Chaque pression déclenche une mesure de type Stern-Gerlach — par exemple, à l'aide d'un champ magnétique inhomogène comme indiqué à la section 5.10 — sur l'atome suspendu au centre du dodécaèdre. Rappelons que pour un spin $\frac{3}{2}$, cette mesure admet quatre résultats possibles, correspondant (dans le cas où le dispositif de mesure est orienté vers le haut) aux quatre vecteurs mutuellement orthogonaux $|\uparrow\uparrow\uparrow\rangle$, $|\downarrow\uparrow\uparrow\rangle$, $|\downarrow\downarrow\uparrow\rangle$ et $|\downarrow\downarrow\downarrow\rangle$ respectivement associés aux quatre faisceaux possibles dans lesquels l'atome peut se trouver après la traversée de l'appareil de mesure. *Trucs Quintessentiels* s'est arrangé pour que la pression sur un bouton quelconque oriente cet appareil dans la direction (définie à partir du centre du dodécaèdre) de ce bouton. La sonnette retentit (**OUI**) si l'atome est dans le

* Il existe des exemples (Zeilinger *et al.* 1992) dans lesquels l'emmêlement d'une paire de particules est lui-même une propriété emmêlée !

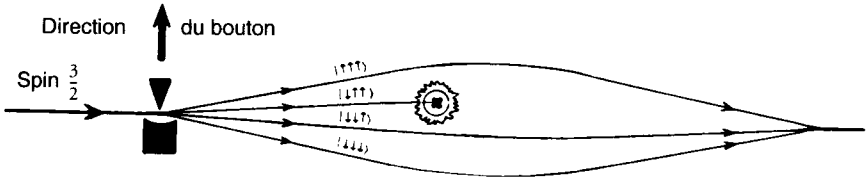


Figure 5.27. La société *Trucs Quintessentiels* s'est arrangée pour que la pression d'un bouton du dodécaèdre déclenche une mesure de spin sur l'atome de spin $\frac{3}{2}$ dans la direction du bouton (considérée ici comme « *up* »). Si l'atome est dans l'état $|\downarrow\uparrow\uparrow\rangle$, la sonnette retentit (**OUI**). Si la réponse est **NON**, les faisceaux se recombinent et le dodécaèdre est prêt pour une mesure dans une autre direction.

deuxième des quatre faisceaux possibles (Fig. 5.27). Autrement dit (en utilisant la notation associée à l'orientation vers le haut), l'état $|\downarrow\uparrow\uparrow\rangle$ suscite la réponse **OUI** — la sonnette retentit et un magnifique feu d'artifice s'ensuit —, tandis que les trois autres états ne suscitent *aucune* réponse (*i.e.* **NON**). Si la réponse est **NON**, les trois autres faisceaux dans lesquels l'atome peut se trouver sont alors réunis (par exemple, en inversant la direction du champ magnétique inhomogène) sans que les différences qui les caractérisent aient eu extérieurement le moindre effet perturbateur. L'atome est alors prêt pour la sélection d'une autre direction, associée à la pression d'un autre bouton. Notons que la pression d'un bouton correspond à l'exécution d'une mesure *primitive* (*cf.* §5.13).

L'état total des deux atomes de spin $\frac{3}{2}$ issus de l'état de spin nul $|\Omega\rangle$ peut s'exprimer sous la forme

$$|\Omega\rangle = |G\uparrow\uparrow\uparrow\rangle|D\downarrow\downarrow\downarrow\rangle - |G\uparrow\uparrow\downarrow\rangle|D\downarrow\downarrow\uparrow\rangle + |G\uparrow\downarrow\downarrow\rangle|D\downarrow\uparrow\uparrow\rangle - |G\downarrow\downarrow\downarrow\rangle|D\uparrow\uparrow\uparrow\rangle.$$

Supposons que mon atome soit celui de droite et que je constate que son état est effectivement $|D\downarrow\uparrow\uparrow\rangle$ parce que la sonnette retentit lorsque je presse en premier le bouton le plus en haut. Il s'ensuit que si mon collègue choisit de presser en premier le bouton *opposé* au mien — son état $|G\uparrow\downarrow\downarrow\rangle$ —, sa sonnette retentit nécessairement. En outre, si ma sonnette *reste muette* lorsque je presse ce premier bouton, la sonnette de mon collègue doit également rester silencieuse s'il presse le bouton opposé.

Vérifions maintenant que les propriétés (a) et (b) (*cf.* §5.3) garanties par *Trucs Quintessentiels* sont effectivement satisfaites par ces mesures primitives. Pour cela, nous allons utiliser quelques propriétés mathématiques, présentées dans l'appendice C, de la description des états de spin — en particulier, de spin $\frac{3}{2}$ — donnée par Majorana. Pour simplifier notre démonstration, nous identifierons la sphère de Riemann à la sphère passant par tous les sommets du dodécaèdre — *i.e.*, à la sphère *circonscrite* au dodécaèdre. Notons ensuite que la description de Majorana de l'état **OUI** pour un bouton situé en un sommet P du dodécaèdre est simplement deux fois P associé au point P* qui

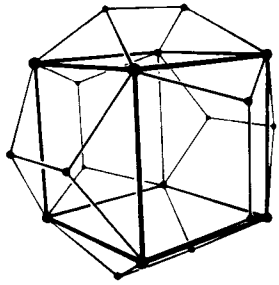


Figure 5.28. On peut inscrire un cube à l'intérieur d'un dodécaèdre régulier. Les sommets de ce cube sont 8 sommets du dodécaèdre. Remarquez que les sommets adjacents de ce cube sont des « deuxièmes voisins » du dodécaèdre.

lui est antipodique — ce qui est de fait l'état $|\downarrow\uparrow\uparrow\rangle$ pour P pris au pôle nord. On peut étiqueter cet état **OUI** par $|P^*PP\rangle$.

Une propriété clé du spin $\frac{3}{2}$ est que les états **OUI** pour les mesures primitives correspondant aux pressions de boutons situés sur deux sommets qui sont « deuxièmes voisins » sont *orthogonaux* ; autrement dit, les états de Majorana $|A^*AA\rangle$ et $|C^*CC\rangle$ sont orthogonaux si A et C sont « deuxièmes voisins » sur le dodécaèdre. Or on voit sur la figure 5.28 que si A et C jouissent de cette propriété, ils sont des sommets *adjacents* d'un *cube* inscrit dans le dodécaèdre, cube dont le centre et les huit sommets sont respectivement le centre et huit sommets du dodécaèdre. En vertu du dernier paragraphe de l'appendice C, il en résulte que $|A^*AA\rangle$ et $|C^*CC\rangle$ sont effectivement orthogonaux.

Qu'est-ce que cela signifie ? Cela signifie notamment que les pressions exercées sur les trois sommets adjacents à un sommet SÉLECTIONNÉ correspondent à des mesures *commutatives* (§5.14), ces sommets étant tous « deuxièmes voisins » les uns des autres. Ainsi, l'ordre dans lequel ils sont pressés n'influe en rien sur le résultat. Cet ordre est également sans influence pour mon collègue sur α du Centaure. Si par hasard son sommet SÉLECTIONNÉ est *opposé* au mien, les trois boutons qu'il peut presser sont alors opposés aux trois que je peux moi-même presser. En vertu de ce qui a été dit plus haut, soit ma sonnette et sa sonnette retentissent lorsqu'elles correspondent à des sommets opposés — indépendamment de l'ordre dans lequel chacun de nous exerçons nos pressions —, soit aucune de nos deux sonnettes ne retentit. Cela établit (a).

Qu'en est-il de (b) ? Notons que l'espace de Hilbert pour le spin $\frac{3}{2}$ est quadridimensionnel, de sorte que les trois éventualités mutuellement orthogonales dans lesquelles ma sonnette peut retentir, disons $|A^*AA\rangle$, $|C^*CC\rangle$ et $|G^*GG\rangle$ — dans le cas où mon sommet SÉLECTIONNÉ est B (Fig. 5.29) —, n'épuisent pas toutes les possibilités. Il se peut également que ma sonnette reste muette lorsque je presse l'un de ces trois boutons et que j'obtienne alors une mesure à résultat nul (la sonnette ne retentit sur aucun

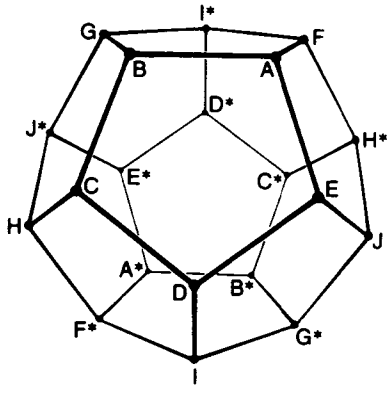


Figure 5.29. Les sommets du dodécaèdre tels qu'ils sont étiquetés dans la discussion de la section 5.18 et à l'appendice B.

de ces trois boutons) signifiant que l'état est l'état (unique) mutuellement orthogonal à |A*AA⟩, |C*CC⟩ et |G*GG⟩. Désignons cet état par |RST⟩, les trois points R, S et T de la sphère de Riemann correspondant à la description de Majorana de cet état. Ces trois points ne sont en fait pas faciles à localiser. (Ils ont été précisément localisés par Jason Zimba, 1993). Mais peu importe, pour notre raisonnement, leur localisation exacte. Il nous suffit de savoir qu'ils occupent des positions dépendant de la géométrie du dodécaèdre et du sommet SÉLECTIONNÉ B. Ainsi notamment (par symétrie), si, au lieu de B, je choisissais de faire de B* — antipodique à B — mon sommet SÉLECTIONNÉ, l'état |R*S*T*⟩ — où R*, S* et T* sont antipodiques à R, S et T — serait l'état de spin si les sonnettes restaient muettes sur les trois sommets A*, C* et G* adjacents à B*.

Supposons maintenant que mon collègue SÉLECTIONNE le sommet B de son dodécaèdre, sommet qui est exactement homologue au sommet B que j'ai moi-même SÉLECTIONNÉ. Si la sonnette reste *muette* sur chacun des trois sommets A, C et G adjacents à *son* sommet B, ses mesures (commutatives) obligent successivement *mon* atome à se placer dans un état orthogonal aux trois états correspondant aux pressions sur les boutons des trois sommets *opposés* A*, C* et G* sur mon dodécaèdre, *i.e.* mon atome est contraint de se placer dans l'état |R*S*T*⟩. En revanche, si ma sonnette reste également *muette* lorsque je presse n'importe lequel de *mes* trois boutons A, C et G, cela oblige mon état à être |RST⟩. Mais en vertu de la propriété **C.1** de l'appendice C, |RST⟩ est *orthogonal* à |R*S*T*⟩, de sorte qu'il est impossible que nos sonnettes restent muettes sur ces six boutons. Cela établit (b).

Cela explique comment *Trucs Quintessentiels* a pu utiliser le phénomène d'emmêlement quantique pour garantir les propriétés (a) et (b). À la section 5.3, nous avons observé que *si* les deux dodécaèdres se comportent comme des objets *indépendants*, il s'ensuit d'une part que l'on a les règles de coloration (c), (d) et (e), mais d'autre part que l'application de ces règles nous

confronte à un problème insoluble (comme le montre explicitement l'appendice B). Ainsi, *Trucs Quintessentiels* a réussi, grâce au phénomène d'emmêlement quantique, à construire quelque chose qui est *impossible* si nos deux dodécaèdres sont des objets pouvant être traités comme indépendants l'un de l'autre après avoir quitté la fabrique. Ce phénomène d'emmêlement quantique n'est pas un simple obstacle épistémologique nous avertissant que nous devons parfois tenir compte des effets probabilistes de l'environnement extérieur lors de l'étude d'une situation physique, car lorsqu'on peut isoler convenablement ses effets, ce phénomène s'avère être mathématiquement très précis et possède souvent une structure géométrique très bien définie.

Ces prédictions du formalisme quantique n'admettent aucune explication en termes d'entités supposées mutuellement indépendantes. Le phénomène d'emmêlement quantique n'admet en général aucune explication du type « chaussettes de Bertlmann ». Les règles de l'évolution quantique standard — notre procédure **U** — nous amènent à conclure que les objets *restent* « emmêlés » de cette manière étrange quelle que soit la distance qui les sépare. C'est seulement avec **R** que ces emmêlements disparaissent. Mais **R** est-il un processus réel ? Si l'on pense qu'il n'en est rien, ces emmêlements doivent alors se maintenir éternellement, même s'ils restent dissimulés aux regards en raison de la complexité excessive du monde réel.

Cela signifie-t-il qu'il faille considérer tous les éléments de l'Univers comme mutuellement emmêlés ? Nous l'avons déjà remarqué (§5.17), le phénomène d'emmêlement quantique n'a pas d'équivalent en physique classique, physique dans laquelle les effets s'atténuent généralement avec la distance, de sorte que nous n'avons pas besoin de savoir ce qui se passe dans la galaxie d'Andromède pour expliquer le comportement des objets lors d'une expérience dans un laboratoire terrestre. L'emmêlement quantique, en revanche, semble effectivement correspondre à cette « effrayante action à distance » si choquante pour Einstein. Mais il s'agit d'une « action » extrêmement subtile, que l'on ne peut utiliser pour transmettre de l'information.

En dépit de cette impossibilité, on ne peut ignorer les (« effrayants ») effets à distance de l'emmêlement quantique. Aussi longtemps que ces effets persistent, aucun objet dans l'Univers ne peut, rigoureusement parlant, être considéré comme véritablement indépendant. Selon moi, ce fait place la théorie physique dans une situation qui est loin d'être satisfaisante. Rien dans la théorie quantique standard n'explique réellement pourquoi, en pratique, on *peut* ignorer ces emmêlements. Pourquoi peut-on se dispenser de considérer que l'Univers n'est qu'un désordre incroyablement complexe d'emmêlements quantiques sans aucun lien avec le monde classique que nous observons réellement ? En pratique, c'est l'utilisation répétée de la procédure **R** qui met fin aux emmêlements — comme lorsque mon collègue et moi-même effectuons nos mesures sur les atomes emmêlés situés aux centres de nos dodécaèdres. La question qui se pose est alors la suivante : **R** est-elle un processus physique *réel* qui, en un certain sens, détruit *réellement* les emmêlements quantiques ? Ou peut-on seulement dire que tout cela n'est qu'une forme d'illusion ?

Ces questions troublantes seront analysées dans le prochain chapitre. Elles jouent selon moi un rôle capital pour la recherche d'une action physique non calculable.

Appendice B : la non-colorabilité du dodécaèdre

Rappelons que le problème posé à la section 5.3 consiste à montrer qu'il n'existe aucun moyen de colorer en NOIR et BLANC tous les sommets d'un dodécaèdre de sorte que deux « deuxièmes voisins » (sommets adjacents à un même sommet) ne soient jamais simultanément BLANCS et que les six sommets adjacents à une paire de sommets opposés ne soient pas tous NOIRS. La symétrie du dodécaèdre va nous être extrêmement précieuse pour éliminer les diverses possibilités.

Désignons les sommets comme l'indique la figure 5.29. A, B, C, D, E sont les sommets successifs d'une face pentagonale et F, G, H, I, J les sommets qui leur sont respectivement adjacents. Comme à la section 5.18, A*, ..., J* sont les sommets antipodiques respectifs de A, ..., J. Notons tout d'abord qu'en vertu de la seconde propriété, le dodécaèdre doit posséder au moins un sommet BLANC. Supposons que ce soit A.

Supposons alors que ce sommet BLANC A ait, parmi ses voisins immédiats, *un autre* sommet BLANC — par exemple, B (Fig. 5.29). Les dix sommets entourant A et B, à savoir C, D, E, J, H*, F, I*, G, J* et H sont tous nécessairement NOIRS, car chacun d'eux est adjacent à un sommet adjacent à A ou à B. Considérons maintenant les six sommets adjacents à chacun des deux sommets de la paire antipodique H, H*. L'un des ces six sommets est nécessairement BLANC, de sorte que l'un au moins des deux sommets F* et C* est nécessairement BLANC. Le même raisonnement appliqué à la paire antipodique J, J* permet de conclure que l'un au moins des deux sommets G* et E* est nécessairement BLANC. Or cela est *impossible*, car G* et E* sont tous deux « deuxièmes voisins » à la fois de F* et de C*. Cela exclut donc que le sommet BLANC A puisse avoir un voisin immédiat BLANC — et par symétrie, cela exclut également la possibilité de deux voisins immédiats BLANCS.

Ainsi, les sommets B, C, D, E, J, H*, F, I*et G qui entourent le sommet BLANC A sont nécessairement NOIRS, car chacun d'eux est soit adjacent à A, soit adjacent à un sommet adjacent à A. Si l'on examine maintenant les six sommets adjacents à l'un des sommets composant la paire antipodique A, A*, on conclut que l'un des sommets B*, E* et F* est nécessairement BLANC. Par symétrie, peu importe celui des trois qui est effectivement BLANC — supposons que ce soit F*. Remarquons que E* et G* étant adjacents à des sommets adjacents à F*, ils sont tous deux NOIRS. Il en est donc de même pour H — car il est adjacent à F*, et le raisonnement précédent nous a permis d'exclure les sommets adjacents BLANCS. Toutefois, cette coloration est impossible,

car les sommets antipodiques J et J* n'ont maintenant que des sommets NOIRS comme voisins immédiats. La démonstration est ainsi achevée : les dodécaèdres magiques sont une impossibilité *classique.*

Appendice C : l'orthogonalité des états de spin généralisés

La description des états de spin généralisés donnée par Majorana n'est pas très connue des physiciens ; pourtant, elle offre une image du spin qui est à la fois très utile et géométriquement éclairante. Je vais donner ici un bref exposé des formules de base et de certaines de leurs conséquences géométriques. Cet exposé nous fournira en particulier les relations d'orthogonalité sous-jacentes à la géométrie des dodécaèdres magiques de la section 5.18. Ma présentation différera notablement de celle donnée par Majorana (1932) et suivra d'assez près celles de Penrose (1994*a*) et Zimba et Penrose (1993).

L'idée est de considérer l'ensemble non ordonné des n points de la sphère de Riemann comme les n racines d'un polynôme complexe de degré n et (essentiellement) de faire des coefficients de ce polynôme les coordonnées de l'espace de Hilbert — à $n+1$ dimensions — des états de spin d'une particule (massive) de spin $\frac{1}{2}n$. Prenant (*cf.* §5.10) comme états de base les divers résultats possibles d'une mesure de spin dans la direction verticale, nous allons leur associer les monômes suivants (affectés de coefficients de normalisation pour les transformer en vecteurs unité) :

$$|\uparrow\uparrow\uparrow\uparrow\ldots\uparrow\uparrow\rangle \text{ correspond à } x^n$$

$$|\downarrow\uparrow\uparrow\uparrow\ldots\uparrow\uparrow\rangle \text{ correspond à } n^{1/2}x^{n-1}$$

$$|\downarrow\downarrow\uparrow\uparrow\ldots\uparrow\uparrow\rangle \text{ correspond à } \{n(n-1)/2!\}^{1/2}x^{n-2}$$

$$|\downarrow\downarrow\downarrow\uparrow\ldots\uparrow\uparrow\rangle \text{ correspond à } \{n(n-1)(n-2)/3!\}^{1/2}x^{n-3}$$

$$\ldots$$

$$|\downarrow\downarrow\downarrow\downarrow\ldots\downarrow\uparrow\rangle \text{ correspond à } n^{1/2}x$$

$$|\downarrow\downarrow\downarrow\downarrow\ldots\downarrow\downarrow\rangle \text{ correspond à } 1.$$

(Les expressions figurant entre les accolades sont les coefficients du binôme.) Ainsi, l'état de spin généralisé $\frac{1}{2}n$

$$z_n|\uparrow\uparrow\uparrow\ldots\uparrow\rangle + z_{n-1}|\downarrow\uparrow\uparrow\ldots\uparrow\rangle + z_{n-2}|\downarrow\downarrow\uparrow\ldots\uparrow\rangle$$
$$+ z_{n-3}|\downarrow\downarrow\downarrow\ldots\uparrow\rangle + \ldots + z_0|\downarrow\downarrow\downarrow\ldots\downarrow\rangle$$

correspond au polynôme

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0$$

où

$$a_n = z_n, \; a_{n-1} = n^{1/2} z_{n-1},$$

$$a_{n-2} = \{n(n-1)/2!\}^{1/2} z_{n-2}, \ldots, a_0 = z_0$$

Les racines $x = \alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n$ de $p(x) = 0$ fournissent les n points de la sphère de Riemann (avec leurs multiplicités) qui correspondent à la description de Majorana. Celle-ci inclut l'éventualité d'un point de Majorana en $x = \infty$ (le pôle sud), qui se réalise lorsque le degré du polynôme $p(x)$ est égal à n moins la valeur de multiplicité de ce point.

Une rotation de la sphère correspond à la substitution

$$x \mapsto (\lambda x - \mu)\,(\bar{\mu} x + \bar{\lambda})^{-1}$$

(avec $\lambda\bar{\lambda} + \mu\bar{\mu} = 1$) suivie de l'élimination des dénominateurs obtenue en multipliant toute l'expression par $(\bar{\mu} x + \bar{\lambda})^n$. Les polynômes correspondant aux résultats de mesures de spin (par exemple, de type Stern-Gerlach) dans une direction arbitraire ont donc pour terme général

$$c(\lambda x - \mu)^p (\bar{\mu} x + \bar{\lambda})^{n-p}.$$

Les points μ/λ et $-\bar{\lambda}/\bar{\mu}$ sont antipodiques sur la sphère de Riemann et correspondent respectivement à la direction de mesure du spin et à sa direction opposée. (Cela suppose un choix de phases approprié pour les états $|\uparrow\uparrow\uparrow\ldots\uparrow\rangle$, $|\downarrow\uparrow\uparrow\ldots\uparrow\rangle$, $|\downarrow\downarrow\uparrow\ldots\uparrow\rangle$, …, $|\downarrow\downarrow\downarrow\ldots\downarrow\rangle$. Les propriétés mentionnées plus haut et la vérification de ces calculs s'apprécient mieux dans le formalisme des spineurs de rang 2. Je renvoie le lecteur à Penrose et Rindler (1984), en particulier p. 162 et §4.15. L'état de spin généralisé $\frac{1}{2}n$ y est décrit à l'aide d'un spineur symétrique à n composantes, et la description de Majorana découle de sa décomposition canonique en un produit symétrisé de vecteurs de spin.)

Un point α de la sphère de Riemann a pour antipode $-1/\bar{\alpha}$. Ainsi, la réflexion, par rapport au centre de cette sphère, de tous les points de Majorana qui sont racines du polynôme

$$a(x) \equiv a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_{n-1} x^{n-1} + a_n x^n$$

donne les racines du polynôme

$$a^*(x) \equiv \bar{a}_n + \bar{a}_{n-1} x + \bar{a}_{n-2} x^2 - \ldots - (-1)^n \bar{a}_1 x^{n-1} + (-1)^n \bar{a}_0 x^n.$$

Si l'on a deux états $|\alpha\rangle$ et $|\beta\rangle$ donnés respectivement par les polynômes $a(x)$ et $b(x)$, où

$$b(x) \equiv b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \ldots + b_{n-1} x^{n-1} + b_n x^n,$$

leur produit scalaire est alors

$$\langle \beta | \alpha \rangle = \bar{b}_0 a_0 + \frac{1}{n} \bar{b}_1 a_1 + \frac{2!}{n(n-1)} \bar{b}_2 a_2 + \frac{3!}{n(n-1)(n-2)} \bar{b}_3 a_3 + \ldots + \bar{b}_n a_n$$

Cette expression est invariante sous l'action des rotations de la sphère, ainsi qu'on peut le vérifier directement en utilisant les formules données plus haut.

Appliquons cette expression du produit scalaire au cas particulier où $b(x) = a^*(x)$, ce qui, dans la description de Majorana, correspond à la situation de deux états mutuellement antipodiques. Le produit scalaire de ces deux états est (au signe près)

$$a_0 a_n - \frac{1}{n} a_1 a_{n-1} + \frac{2!}{n(n-1)} a_2 a_{n-2} - \ldots - (-1)^n \frac{1}{n} a_{n-1} a_1 + (-1)^n a_n a_0.$$

Cette expression s'annule si n est *impair*. On en déduit alors le théorème suivant. (L'état donné par P, Q, ..., S dans la description de Majorana est désigné par $|PQ... S\rangle$. Le point antipodique de X est désigné par X*.)

C.1 Si n est impair, l'état $|PQR... T\rangle$ est orthogonal à $|P^*Q^*R^*... T^*\rangle$.

L'expression générale du produit scalaire permet de déduire les deux propriétés suivantes :

C.2 L'état $|PP...P\rangle$ est orthogonal à chaque état $|P^*AB... D\rangle$.

C.3 Si la projection stéréographique de Q* à partir de P* est l'isobarycentre des projections stéréographiques de A, B, C, ..., D à partir de P*, l'état $|QPP... P\rangle$ est orthogonal à $|ABC... E\rangle$.

(L'isobarycentre d'un ensemble de points est le centre de masse de la configuration formée par des masses ponctuelles égales situées en ces points. La projection stéréographique a été décrite à la section 5.10, figure 5.19.) Pour démontrer **C.3**, on tourne la sphère de Riemann de manière à amener P* au pôle sud. L'état $|QPP... P\rangle$ est alors représenté par le polynôme $x^{n-1}(x-\chi)$, où χ définit le point Q sur la sphère. Le produit scalaire avec l'état représenté par le polynôme $(x-\alpha_1)(x-\alpha_2)(x-\alpha_3)\ldots(x-\alpha_n)$ — dont la description de Majorana est donnée par $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n$ — s'annule pour

$$1 + n^{-1} \bar{\chi} (\alpha_1 + \alpha_2 + \alpha_3 + \ldots + \alpha_n) = 0,$$

i.e. lorsque $-1/\bar{\chi}$ est égal à $(\alpha_1 + \alpha_2 + \alpha_3 + \ldots + \alpha_n)/n$, ce qui est l'isobarycentre, dans le plan complexe, des points donnés par $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n$. Cela établit **C.3**. Pour démontrer **C.2**, on place cette fois P — au lieu de P* — au pôle sud. L'état $|PPP... P\rangle$ est représenté par la constante 1, considé-

rée comme un polynôme de degré 0. Le produit scalaire correspondant s'annule alors lorsque

$$\alpha_1\alpha_2\alpha_3\ldots\alpha_n = 0,$$

i.e. lorsque l'un au moins des α_1, α_2, α_3, ..., α_n s'annule — le point 0 du plan complexe représentant le pôle nord P*. Cela établit **C.2**.

Le résultat **C.2** permet de donner une interprétation physique des points de Majorana. Il entraîne en effet que ces points définissent les directions dans lesquelles une mesure de spin (de type Stern-Gerlach) donne une probabilité nulle de trouver ce spin entièrement orienté dans la direction opposée à celle mesurée (*cf.* EOLP p. 297). Il implique en outre, comme cas particulier, que les états de spin orthogonaux de particules de spin $\frac{1}{2}$ ($n = 1$) sont exactement ceux dont les points de Majorana sont antipodiques. Le résultat **C.3** permet de déduire l'interprétation géométrique générale de l'orthogonalité dans le cas de spin 1 ($n = 2$). Un cas particulier remarquable survient lorsque les deux états sont représentés par deux paires de points antipodiques correspondant à des rayons de la sphère perpendiculaires entre eux. Dans le cas du spin $\frac{3}{2}$ ($n = 3$), **C.3** et **C.1** constituent tout ce dont nous avons besoin pour l'analyse de la section 5.18. (Une interprétation géométrique de l'orthogonalité dans le cas général sera donnée ailleurs.)

Le cas particulier de **C.3** nécessaire à l'analyse de la section 5.18 correspond à la situation où P et Q sont deux sommets adjacents sur un cube inscrit dans la sphère de Riemann, de sorte que dans ce cube, les côtés PQ et P*Q* sont opposés. Les longueurs de PQ*et QP* sont égales à $\sqrt{2}$ fois celles de PQ et P*Q*. Des considérations géométriques élémentaires permettent de déduire de **C.3** que les états $|P^*PP\rangle$ et $|Q^*QQ\rangle$ sont orthogonaux.

6
Théorie quantique et réalité

6.1 R est-elle un processus réel ?

Le chapitre précédent a analysé les énigmes-**Y** de la théorie quantique. Si certains de ces phénomènes — comme l'emmêlement quantique sur des distances de plusieurs années-lumière[1] — n'ont pas été concrètement vérifiés, nombre d'indices expérimentaux suffisamment solides incitent à penser que ces énigmes-**Y** sont des aspects réels du comportement des constituants de l'Univers dans lequel nous vivons.

Toutefois, le comportement du monde physique au niveau quantique heurte profondément notre intuition et diffère énormément, à de nombreux égards, du comportement « classique » semblant prévaloir au niveau plus familier de notre expérience sensible. Ce comportement quantique inclut manifestement des effets d'emmêlement agissant sur des distances considérables, du moins tant qu'elles n'impliquent que des objets quantiques — des électrons, des photons, des atomes ou des molécules. Le contraste entre l'étrange comportement *quantique* des « petits » objets, même sur de grandes distances, et le comportement *classique* plus familier des objets plus gros, est à l'origine des énigmes-**X** de la théorie quantique. Y aurait-il ainsi deux types de lois physiques, opérant chacun à un niveau phénoménologique différent ?

Une telle idée contredit fortement ce que nous avons appris à attendre de la physique. L'un des plus grands succès de la dynamique galiléo-newtonienne du XVII[e] siècle fut en effet de démontrer que les mouvements des corps célestes obéissent exactement aux mêmes lois que celles agissant sur Terre. Depuis la

Grèce antique, voire même avant, les hommes croyaient en l'existence de deux corps de lois totalement distincts, l'un pour le ciel, l'autre pour la Terre. Galilée et Newton nous ont révélé que ces lois étaient en fait identiques à toutes les échelles — et cette révélation s'avéra cruciale pour le progrès de la science. Pourtant (ainsi que l'a souligné le professeur Ian Percival, de l'Université de Londres), la théorie quantique semble nous mettre en présence d'un schéma analogue à celui des anciens Grecs, avec un corps de lois pour le niveau classique et un autre, très différent, pour le niveau quantique. Mon opinion — qui est partagée par une minorité non négligeable de physiciens — est que la situation dans laquelle se trouve actuellement notre compréhension de la physique est toute provisoire et qu'il est fort probable que la découverte de lois quantiques/classiques adéquates, opérant uniformément à *toutes* les échelles, ouvrira la voie à un progrès scientifique comparable à celui initié par Galilée et Newton.

Le lecteur peut toutefois légitimement se demander si la théorie quantique, telle qu'on la comprend ordinairement, rend compte également des phénomènes classiques. Selon moi, il n'en est rien. Nombre de personnes contestent ce point de vue et affirment que les systèmes physiques qui sont en un certain sens grands ou complexes et entièrement régis par des lois quantiques se comportent exactement comme des objets classiques, du moins en très bonne approximation. Dans un premier temps, nous allons examiner la crédibilité de cette affirmation — *i.e.* examiner si le comportement apparemment « classique » des objets macroscopiques découle du comportement quantique de leurs infimes parties constitutives. S'il s'avère que tel n'est pas le cas, nous devrons alors nous orienter dans une direction nouvelle afin de dégager un schéma cohérent à *tous* les niveaux. Je préviens toutefois le lecteur : le problème de la réalité de **R** est très controversé et admet de nombreuses approches. Il serait téméraire de ma part de les présenter toutes, voire de démonter en détail celles d'entre elles que je trouve peu plausibles ou indéfendables. On me pardonnera d'exposer ces approches en me plaçant de mon propre point de vue, au risque de manquer d'impartialité envers celles qui se démarquent trop de ma conception. Je présente par avance mes excuses pour les injustices que je ne pourrai éviter de commettre.

Lorsqu'on tente de définir clairement le point où le comportement *quantique*, caractérisé par la persistance des superpositions quantiques des divers états, cède réellement la place — sous l'action de **R** — au comportement *classique* apparemment dépourvu, lui, de telles superpositions, on se heurte à une difficulté fondamentale : la procédure **R** est en effet si « sournoise » que, d'un point de vue observationnel, on ne peut localiser précisement le niveau auquel elle entre en jeu — c'est d'ailleurs là une des raisons pour lesquelles nombre de physiciens refusent de la considérer comme un processus réel. Il semble que le point où l'on pense que **R** intervient n'ait aucune incidence expérimentale, pourvu qu'on le situe au-dessus du niveau où s'observent les interférences quantiques, mais pas au-dessus de celui où nous percevons directement la présence d'alternatives classiques au lieu de superpositions linéaires complexes (bien que, nous le verrons dans un instant, certains physiciens soutiennent que les superpositions persistent même à ce dernier niveau).

Comment déterminer le niveau où **R** intervient *vraiment* — si tant est qu'elle intervienne bien physiquement ? Il est difficile de répondre à une telle question à l'aide d'une expérience physique. Si la procédure **R** *est* un processus physique réel, elle peut intervenir à une multitude de niveaux, situés entre ceux, microscopiques, où l'on *observe* les interférences quantiques et celui, macroscopique, où l'on perçoit le comportement classique. En outre, les effets de l'emmêlement quantique pouvant se faire sentir sur des distances considérables (*cf.* §5.4), ces « différences de niveaux » ne semblent pas liées à la taille physique. Nous le verrons, les *différences d'énergie* s'avèrent plus appropriées pour mesurer cette échelle de niveaux. Quoi qu'il en soit, au niveau macroscopique, l'endroit où s'opère le « passage du témoin » est défini par nos *perceptions conscientes.* Cela pose un problème délicat du point de vue théorique, car nous ne connaissons pas vraiment les processus physiques cérébraux associés à la perception. Cependant, la nature physique de ces processus semble imposer une limite macroscopique à toute théorie fondée sur l'hypothèse de la *réalité* de **R**. Cela autorise cependant encore quantité de possibilités entre les deux extrêmes et donc laisse une place considérable aux multiples conceptions que l'on peut échafauder sur les mécanismes *réellement* mis en jeu lors de l'intervention de **R.**

L'un des grands problèmes concerne la « réalité » du formalisme quantique — voire du monde quantique lui-même. Je ne peux m'empêcher de citer à ce propos une remarque que m'a faite lors d'un dîner, il y a quelques années, le professeur Bob Wald, de l'université de Chicago :

> Si on croit réellement à la mécanique quantique, on ne peut la prendre au sérieux.

Cette phrase me semble exprimer une vérité profonde sur la théorie quantique et l'attitude des gens à son égard. Ceux qui proclament avec le plus d'ardeur que cette théorie ne nécessite aucune modification ne considèrent généralement *pas* qu'elle représente le vrai comportement d'un monde quantique « réel ». Niels Bohr, qui fut un artisan éminent du développement et de l'interprétation de la théorie quantique, fut sur ce plan l'un des plus extrémistes. Il semble qu'il n'ait vu dans le vecteur d'état qu'un outil commode servant uniquement à calculer les probabilités des résultats de « mesures » pouvant être effectuées sur un système. Selon Bohr, le vecteur d'état *ne constitue pas* une description objective de la *réalité* quantique, quelle qu'elle soit. Il ne représente que « notre connaissance » du système. Bohr doutait d'ailleurs que le concept même de « réalité » eût un sens au niveau quantique. Si assurément il « croyait réellement en la mécanique quantique », il semble cependant avoir estimé qu'on ne pouvait « prendre au sérieux » l'idée que le vecteur d'état fournît une description de la réalité physique quantique.

Le point de vue diamétralement opposé affirme que le vecteur d'état donne une description mathématique précise d'un monde quantique *réel,* un monde évoluant avec une précision extraordinaire — pouvant toutefois ne pas être absolue — selon les règles mathématiques fournies par les équations de la théorie. Les partisans de ce point de vue me semblent alors avoir le choix entre

deux possibilités. Soit ils considèrent que l'évolution de l'état quantique est entièrement régie par la procédure **U**, la procédure **R** devenant ainsi une illusion, un outil commode ou une approximation, sans participation *aucune* à l'évolution *réelle* de la réalité décrite par l'état quantique. Ce choix débouche apparemment sur une interprétation (ou sur des interprétations) de type *mondes multiples* d'Everett[2]. Je donnerai dans un instant quelques détails sur cette interprétation. Soit, en revanche, ils « prennent au sérieux » le formalisme quantique dans sa totalité et considèrent que **U** *et* **R** représentent (avec une précision considérable) le comportement physique *réel* d'un monde quantique/classique *physiquement réel* et décrit par un vecteur d'état. Mais accepter ainsi le formalisme quantique dans son intégralité équivaut à accepter que la théorie ne soit pas totalement précise à tous les niveaux. Car telle qu'elle se présente, **R** ne vérifie pas nombre des propriétés de **U**, en particulier sa *linéarité*. En ce sens, cela revient à ne pas « croire réellement en la mécanique quantique ». Les sections qui suivent vont examiner plus à fond ces problèmes.

6.2 Le point de vue de type mondes multiples

Regardons tout d'abord jusqu'où peut mener la première voie « réaliste », celle qui débouche sur l'idée de « mondes multiples ». Selon cette conception, le vecteur d'état, dont l'évolution est entièrement régie par l'action de **U**, représente effectivement la réalité. Il faut dès lors admettre que les objets du niveau classique — les balles de golf, les êtres humains, etc. — sont eux aussi soumis aux lois de superpositions linéaires quantiques. On pourrait certes concevoir que cela ne pose aucune véritable difficulté, puisque les superpositions d'états se manifestent très rarement au niveau classique. Malheureusement le problème est la *linéarité* de **U**. Quelle que soit la quantité de matière impliquée, **U** *ne modifie pas* les coefficients de pondération entrant en jeu dans les superpositions d'états. D'elle-même, **U** ne permet pas la « désuperposition » des superpositions dès qu'un système s'agrandit ou se complexifie. Ces superpositions ne tendant nullement à « disparaître » chez les objets classiques, ceux-ci devraient donc apparaître fréquemment dans des états manifestement superposés. La question qui se pose est alors la suivante : pourquoi ces superpositions de possibilités macroscopiques n'ont-elles aucun impact sur notre perception du monde classique ?

Voyons comment les tenants du point de vue « mondes multiples » expliquent cela. Considérons une situation, semblable à celle examinée à la section 5.17, dans laquelle un détecteur de photons décrit par un état $|\Psi\rangle$ rencontre un photon se trouvant dans un état superposé $|\alpha\rangle + |\beta\rangle$, où $|\alpha\rangle$ active le détecteur tandis que $|\beta\rangle$ n'interagit pas avec lui. (Par exemple, une source émet un photon qui heurte un miroir semi-argenté, et $|\alpha\rangle$ et $|\beta\rangle$ représentent respectivement les parties transmise et réfléchie de l'état du photon.) Soulignons que

l'on accepte ici la pertinence du concept de vecteur d'état pour un objet classique tel qu'un détecteur, car selon ce point de vue, les vecteurs d'état sont des représentations fidèles de tous les niveaux de réalité. Ainsi, $|\Psi\rangle$ décrit le détecteur tout entier, et non, comme à la section 5.17, les quelques éléments de sa structure qui sont initialement en contact avec le photon. Rappelons que, comme à la section 5.17, l'état $|\Psi\rangle\,(|\alpha\rangle + |\beta\rangle)$ de l'ensemble détecteur-photon avant la rencontre devient, après la rencontre, l'état quantique emmêlé

$$|\Psi_O\rangle + |\Psi_N\rangle|\beta'\rangle\,.$$

Cet état quantique emmêlé *tout entier* est maintenant censé représenter la *réalité* de la situation. On ne dit pas que *soit* le détecteur a reçu et absorbé le photon (état $|\Psi_O\rangle$), *soit* le détecteur n'a pas reçu le photon et celui-ci reste libre (état $|\Psi_N\rangle|\beta'\rangle$), mais on affirme que ces *deux* éventualités coexistent dans un état de superposition au sein d'une réalité complète dans laquelle toutes ces superpositions sont préservées. On peut prolonger ce raisonnement et imaginer qu'un expérimentateur humain examine le détecteur pour voir s'il a ou non enregistré le photon. Avant d'examiner le détecteur, cet être humain est lui aussi dans un état quantique, disons $|\Sigma\rangle$, de sorte que le système être humain-détecteur-photon est alors dans l'état quantique « produit »

$$|\Sigma\rangle\,(|\Psi_O\rangle + |\Psi_N\rangle|\beta'\rangle)\,.$$

Puis, après avoir examiné l'état, l'observateur perçoit soit que le détecteur a reçu et absorbé le photon (état $|\Sigma_O\rangle$), soit qu'il ne l'a pas reçu (état orthogonal $|\Sigma_N\rangle$). Si l'on fait l'hypothèse que l'observateur n'interagit pas avec le détecteur une fois qu'il l'a examiné, la situation finale est alors décrite par le vecteur d'état

$$|\Sigma_O\rangle|\Psi'_O\rangle + |\Sigma_N\rangle|\Psi'_N\rangle|\beta''\rangle\,.$$

On est maintenant en présence de deux états (orthogonaux) pour l'observateur, tous deux partie intégrante de l'état complet du système. Dans l'état $|\Sigma_O\rangle$, l'observateur a constaté que le détecteur a affiché la réception du photon ; le détecteur est alors dans un état exprimant qu'il a effectivement enregistré le photon. Dans l'état $|\Sigma_N\rangle$, l'observateur a constaté que le détecteur n'a pas affiché la réception du photon ; le détecteur est alors dans un état exprimant qu'il n'a pas enregistré le photon et le photon dans un état exprimant qu'il se déplace librement. Le point de vue « mondes multiples » affirme donc la coexistence de divers exemplaires du « moi » de l'observateur au sein de l'état total, correspondant à des appréciations différentes du monde extérieur. L'état réel du monde associé à chaque exemplaire de l'observateur est cohérent avec les perceptions de cet exemplaire.

On peut généraliser cela aux situations physiques plus « réalistes » dans lesquelles il y aurait, non pas les deux seules possibilités de cet exemple, mais un nombre fantastique d'états quantiques différents surgissant en permanence à mesure que progresse l'histoire de l'Univers. Ainsi, selon le point de vue des mondes multiples, l'état total de l'Univers comprendrait de nombreux

« mondes » différents, associés aux exemplaires tout aussi nombreux de tout observateur humain. Chaque exemplaire percevrait un monde cohérent avec ses propres perceptions, et cela suffirait pour construire une théorie satisfaisante. Selon ce point de vue, la procédure **R** serait alors une *illusion*, résultant apparemment des mécanismes de perception d'un observateur macroscopique dans un monde soumis aux effets de l'emmêlement quantique.

Je dois dire que ce point de vue m'apparaît très peu satisfaisant, non pas tant parce qu'il témoigne d'un extraordinaire manque d'économie — bien que cela soit très embarrassant, c'est le moins que l'on puisse dire —, mais parce qu'il n'offre pas *réellement* de solution au « problème de la mesure » qu'il est censé résoudre.

Ce *problème de la mesure quantique* consiste à comprendre comment la procédure **R** peut survenir — ou survient effectivement — en tant que propriété du comportement macroscopique de systèmes quantiques soumis à l'évolution **U**. Ce problème ne se résout pas en indiquant simplement un mécanisme pouvant rendre compte d'un comportement de type **R**. Il exige une théorie offrant une certaine compréhension des *circonstances* dans lesquelles (l'illusion de ?) **R** intervient. En outre, cette théorie se doit d'expliquer la remarquable *précision* associée à **R**. Les gens pensent souvent que la précision de la théorie quantique réside dans ses équations dynamiques, à savoir dans **U**. Mais **R** fournit elle aussi des prédictions de probabilités très précises, et tant que l'on ne comprendra pas l'origine de ces probabilités, on n'aura pas une théorie satisfaisante.

En l'absence d'autres éléments, le point de vue de type mondes multiples ne donne pas d'explication acceptable de ces deux aspects. Sans une théorie expliquant comment un « être percevant » diviserait l'Univers en alternatives orthogonales, nous n'avons aucune raison d'escompter qu'un tel être puisse ignorer les superpositions linéaires de balles de golf ou d'éléphants situés en des endroits totalement distincts. (Remarquons que la simple *orthogonalité* des « états du percepteur », comme $|\Sigma_O\rangle$ et $|\Sigma_N\rangle$ à l'instant, ne permet en rien de distinguer entre ces états. Comparez avec le cas de $|G\leftarrow\rangle$ et de $|G\rightarrow\rangle$ par opposition à $|G\uparrow\rangle$et $|G\downarrow\rangle$, dans la discussion du paradoxe EPR à la section 5.17. Dans chacun des cas, les deux états sont orthogonaux, comme le sont $|\Sigma_O\rangle$ et $|\Sigma_N\rangle$, mais rien ne permet de choisir une paire au détriment de l'autre.) En outre, le point de vue « mondes multiples » n'explique absolument pas la règle merveilleuse et extrêmement précise qui transforme miraculeusement en probabilités les carrés des modules des coefficients de pondération complexes[3]. (Comparez également avec les analyses présentées aux sections 6.6 et 6.7.)

6.3 Faut-il prendre $|\psi\rangle$ au sérieux ?

Il existe de nombreuses versions du point de vue selon lequel le vecteur d'état $|\psi\rangle$ *ne constitue pas* une représentation de la réalité physique au niveau quantique, mais est simplement un outil de calcul commode pour déterminer les probabilités ou exprimer la « connaissance » qu'un expérimentateur a d'un système physique. On considère parfois aussi que $|\psi\rangle$ représente non pas l'état d'un système physique individuel, mais un *ensemble* de systèmes physiques similaires possibles. On affirme souvent qu'un vecteur d'état $|\psi\rangle$ emmêlé de manière complexe a, « en pratique » (ou EP pour reprendre la formule de John Bell[4])*, le même comportement qu'un tel ensemble de systèmes physiques — et que cet ensemble constitue tout ce que les physiciens ont besoin de connaître sur le problème de la mesure. On affirme même parfois que $|\psi\rangle$ *ne peut* décrire la réalité quantique car il n'y a aucun sens, à ce niveau, à attribuer une « réalité » à notre monde, la réalité étant uniquement le résultat de « mesures ».

Pour certaines personnes dont moi-même (mais aussi Einstein et Schrödinger — je suis donc en bonne compagnie), cela n'a aucun sens de réserver le mot « réalité » aux seuls objets que nous pouvons percevoir, tels les dispositifs de mesure (ou plutôt certains types de dispositifs de mesure), et de refuser de l'appliquer à un niveau plus profond. Certes le monde quantique a un comportement étrange, insolite, mais il n'est pas « irréel ». Comment d'ailleurs pourrait-on construire des objets réels à partir de constituants irréels ? En outre, les lois mathématiques qui gouvernent le monde quantique sont remarquablement précises — aussi précises que les équations plus familières qui régissent le comportement des objets macroscopiques —, en dépit des images confuses qu'évoquent des expressions comme « fluctuations quantiques » ou « principe d'indétermination ».

Pourtant, si l'on accepte l'existence d'une forme de réalité au niveau quantique, on peut douter que cette réalité soit correctement décrite par un vecteur d'état $|\psi\rangle$. Diverses objections s'opposent en effet à la « réalité » de $|\psi\rangle$. En premier lieu, $|\psi\rangle$ semble devoir subir, de temps en temps, ce mystérieux « saut » discontinu et non local que j'ai désigné par la lettre **R**. Cela ne semble pas être une description physiquement acceptable du monde, en particulier parce que nous avons déjà l'équation de Schrödinger **U**, continue et merveilleusement précise, qui est censée régir (la plupart du temps) l'évolution de $|\psi\rangle$. Pourtant, nous avons vu que, par elle-même, **U** conduit aux difficultés et aux énigmes associées au point de vue de type mondes multiples, et qu'on ne peut obtenir d'image reproduisant fidèlement l'Univers que nous pensons percevoir sans faire intervenir une procédure de type **R**.

On oppose également parfois à la réalité de $|\psi\rangle$ le fait que les alternances du type **U**, **R**, **U**, **R**, **U**, **R**, ... utilisées en théorie quantique ne constituent pas

* En anglais, FAPP = « for all practical purposes ». (N.d.T.)

une description symétrique par rapport au temps (parce que c'est **R** qui détermine le *début*, et non la fin, de chaque action de **U**) et qu'il existe une autre description possible, complètement équivalente, correspondant à une inversion de l'évolution temporelle de **U** (*cf.* EOLP, p. 385-388, Fig. 8.1 et Fig. 8.2). Pourquoi considérer que l'une seulement de ces deux descriptions, et non l'autre, traduit la « réalité » ? Il existe même des points de vue selon lesquels *les deux* vecteurs d'état, évoluant respectivement vers le futur et vers le passé, coexistent dans la description de la réalité physique (Costa de Beauregard 1989 ; Werbos 1989 ; Aharonov et Vaidman 1990). Ces points de vue renferment probablement des idées profondes. Je les examinerai — en même temps que d'autres problèmes annexes — à la section 7.12.

L'une des objections le plus fréquemment opposées à ceux qui considèrent sérieusement $|\psi\rangle$ comme une description de la réalité est que ce vecteur d'état n'est pas directement « mesurable » — dans le sens où face à un état totalement inconnu, on n'a aucun moyen expérimental de déterminer ce qu'est réellement ce vecteur d'état (à un facteur de proportionnalité près). Considérez par exemple le cas d'un atome de spin $\frac{1}{2}$. Rappelons (§5.10, Fig. 5.19) que chaque état de spin possible se caractérise par une direction particulière dans l'espace ordinaire. Mais si l'on n'a aucune idée de cette direction, on n'a alors aucun moyen de la déterminer. Tout ce que l'on peut faire, c'est se fixer une direction et se demander : le spin est-il dans cette direction (**OUI**) ou dans la direction opposée (**NON**) ? Quel que soit l'état de spin dans lequel on suppose initialement que se trouve l'atome, la direction qui lui est associée dans l'espace de Hilbert se projette soit dans l'espace **OUI**, soit dans l'espace **NON**, avec une certaine probabilité. Et une fois cette projection opérée, nous avons perdu l'essentiel de l'information sur ce qu'était « réellement » cet état de spin. Toute mesure de la direction du spin d'un atome de spin $\frac{1}{2}$ ne donne qu'une information *fragmentaire* (*i.e.* la réponse à une question de type oui/non), tandis que les directions associées aux états de spin possibles forment un continuum dont la description précise exigerait un nombre infini d'éléments d'information.

Tout cela est vrai, mais il est difficile d'adopter la position opposée, qui affirme que le vecteur d'état $|\psi\rangle$ est d'une manière ou d'une autre physiquement « irréel » et ne représente au mieux que la totalité de « notre connaissance » du système physique considéré. Je trouve cela très difficile à accepter, notamment parce que le statut d'une telle « connaissance » semble particulièrement empreint de subjectivité. *Qui* détient cette connaissance ? Certainement pas moi. Ma connaissance des vecteurs d'état individuels associés au comportement détaillé de tous les objets qui m'entourent est très limitée. Pourtant, ces objets évoluent selon un comportement précis et sont totalement indifférents à la « connaissance » que l'on pourrait avoir de leurs vecteurs d'état ou à la personne qui pourrait détenir cette connaissance. Des expérimentateurs différents, possédant des connaissances différentes sur un même système physique, utilisent-ils des vecteurs d'état différents pour décrire ce système ? Pas vraiment ; il peut exister des différences concernant certaines caractéristiques de l'expérience, mais cela n'a aucune influence sur son résultat.

L'une des raisons les plus sérieuses de rejeter ce point de vue subjectif sur la réalité[5] de $|\psi\rangle$ est que, quoi que puisse être $|\psi\rangle$, il y a toujours — du moins en principe — une *mesure primitive* (*cf.* §5.13) dont l'espace **OUI** se compose du rayon de l'espace de Hilbert défini par $|\psi\rangle$. L'état physique $|\psi\rangle$ (le rayon des multiples complexes de $|\psi\rangle$) est en effet *entièrement* déterminé par le fait que, pour cet état, le résultat **OUI** est *certain.* Aucun autre état physique ne possède cette propriété. À tout autre état est associée non pas la certitude, mais simplement une probabilité d'obtenir le résultat **OUI** — et donc éventuellement le résultat **NON**. Ainsi, bien qu'il n'y ait aucune mesure qui nous dise ce qu'*est* réellement $|\psi\rangle$, cet état physique est déterminé de manière unique par ce que serait le résultat d'une mesure qui *serait* effectuée sur lui. Nous sommes de nouveau en présence de contrafactuels (§5.2, §5.3), dont nous avons déjà noté l'importance au niveau des prédictions de la théorie quantique.

Pour rendre cette explication plus convaincante, imaginez qu'un système quantique ait été placé dans un état connu $|\phi\rangle$ et que le calcul montre qu'au bout d'un certain temps t, cet état se transforme, sous l'action de l'évolution **U**, en un autre état $|\psi\rangle$. Par exemple, $|\phi\rangle$ pourrait être l'état « spin *up* » ($|\phi\rangle = |\uparrow\rangle$) d'un atome de spin $\frac{1}{2}$ qui aurait été obtenu par l'action d'une mesure antérieure sur cet atome. Supposons que notre atome possède un moment magnétique aligné sur son spin (*i.e.* que cet atome soit un petit aimant pointant dans la direction du spin). Lorsqu'on place l'atome dans un champ magnétique, la direction du spin subit un mouvement de précession bien précis pour donner, au bout du temps t, un nouvel état, disons $|\psi\rangle = |\rightarrow\rangle$, que l'on peut calculer exactement à partir de l'évolution **U**. Doit-on considérer sérieusement que ce nouvel état, obtenu par calcul, fait partie de la réalité physique ? On voit difficilement comment il pourrait en être autrement. Car $|\psi\rangle$ doit être prêt à l'éventualité d'être mesuré à l'aide de la mesure primitive mentionnée plus haut, à savoir celle dont l'espace **OUI** se compose précisément des multiples de $|\psi\rangle$. Ici, cette mesure est la mesure du spin dans la direction $\rightarrow$. Le système doit connaître son état pour pouvoir donner avec *certitude* la réponse **OUI** pour cette mesure, tandis qu'*aucun* état de spin de l'atome *autre* que $|\psi\rangle = |\rightarrow\rangle$ ne peut donner cette réponse avec cette certitude.

Il existe certes, en pratique, de nombreuses situations physiques qui, contrairement à celles associées aux déterminations de spin, rendent une telle mesure primitive totalement irréalisable, mais les règles standard de la théorie quantique n'interdisent pas la mise en œuvre, en principe, de telles mesures. Nier la possibilité de ce type de mesures pour certains types de $|\psi\rangle$ « trop compliqués » reviendrait à changer la structure de la théorie quantique. Peut-être devrait-on effectivement modifier cette structure (et à la section 6.12, j'émettrai quelques suggestions dans ce sens). Il faut en tout cas être bien conscient que si l'on nie les distinctions *objectives* entre différents états quantiques, *i.e.* si l'on *ne considère pas* que $|\psi\rangle$ est, en un sens physique clairement défini, *objectivement* réel (du moins, à un facteur de proportionnalité près), on doit alors modifier un tant soit peu la théorie.

En ce qui concerne la théorie de la mesure, la modification « minimale » souvent suggérée consiste à introduire ce que l'on appelle des *règles de*

superselection[6] qui interdisent efficacement la réalisation de certains types de mesures primitives sur un système. Je ne vais pas discuter ces règles en détail, car elles n'ont, selon moi, pas encore abouti à une vision générale et cohérente du problème de la mesure. Le seul point sur lequel je voudrais insister ici est que même une modification minime de cette nature n'en est pas moins une modification — et cela fait ressortir la nécessité de certaines modifications.

Il me faut enfin mentionner qu'il existe diverses autres approches de la théorie quantique qui, bien que compatibles avec les prédictions habituelles, donnent des « images de la réalité » différant sur plusieurs plans de l'approche qui considère le vecteur d'état $|\psi\rangle$ comme une représentation fidèle, à lui seul, de cette réalité. Parmi celles-ci figure la théorie de l'*onde pilote*, développée par le prince Louis de Broglie (1956) et David Bohm (1952) — une théorie non locale où intervient quelque chose d'équivalent à une fonction d'onde $|\psi\rangle$ *et* à un système de particules de type classique, ces *deux* objets étant censés être « réels ». (Voir aussi Bohm et Hiley 1994.) Il y a également des points de vue (inspirés de l'approche de la théorie quantique due à Richard Feynman 1948) reposant sur les « histoires » entières des comportements possibles et selon lesquels le concept de « réalité physique » diffère sensiblement de celui fourni par un vecteur d'état $|\psi\rangle$ ordinaire. Plus récemment, Griffiths (1984), Omnès (1992), Gell-Mann et Hartle (1993) ont proposé un schéma assez voisin, mais qui (selon une idée due à Aharonov *et al.* 1964) tient compte en outre de la possibilité d'envisager des procédures qui sont, de fait, des mesures partielles répétées. Une analyse de ces diverses idées sortirait du cadre de ce livre (bien que pour certaines d'entre elles — ainsi que pour l'approche algébrique de Haag (1992) —, le formalisme de la matrice densité présenté à la prochaine section joue un rôle important). Disons cependant que si nombre d'aspects de ces procédures présentent un intérêt considérable et une certaine originalité stimulante, je doute fort que le problème de la mesure puisse réellement trouver sa solution dans le seul cadre de ces diverses présentations. Bien entendu, il se peut que l'avenir me donne tort.

6.4 La matrice densité

Nombre de physiciens, se targuant de pragmatisme, ne s'intéressent pas au problème de la « réalité » de $|\psi\rangle$. Tout ce que l'on attend de $|\psi\rangle$, disent-ils, c'est qu'il permette de calculer correctement les probabilités concernant le comportement futur d'un système physique. Souvent, un état représentant une situation physique évolue pour se transformer en un état extrêmement compliqué. Il s'« emmêle » si intimement avec l'environnement qu'il devient impossible, en pratique, d'observer les interférences quantiques distinguant cet état de nombreux autres états similaires. Ces physiciens « pragmatiques » affirment bien sûr qu'il serait insensé de considérer le vecteur d'état particulier résultant de cette évolution comme plus « réel » que d'autres dont il est, en pratique, indiscerna-

ble. En fait, affirment-ils, on peut décrire la « réalité » en utilisant aussi bien un *mélange de probabilités* de vecteurs d'état qu'un vecteur d'état *particulier* : si l'application de **U** sur un vecteur d'état représentant l'état initial d'un système donne quelque chose qui, *en pratique* (le EP de Bell), est indiscernable d'un tel mélange de probabilités, ce mélange donne une description du monde valant celle que l'on obtiendrait à partir du vecteur d'état résultant de l'évolution **U**.

On affirme souvent que — du moins EP — la procédure **R** peut s'interpréter en ces termes. C'est sur cet important problème que je me concentrerai à la section 6.6. Je me demanderai si l'on peut effectivement résoudre le paradoxe (apparent) **U/R** à l'aide de ce seul moyen. Pour cela, il me faut d'abord formuler plus explicitement les procédures adoptées par l'approche standard EP pour expliquer le processus (apparent ?) **R**.

L'élément clé de ces procédures est un objet mathématique baptisé *matrice densité*, un concept important en théorie quantique. C'est cette matrice densité — plus que le vecteur d'état — qui sous-tend généralement la plupart des descriptions mathématiques standard du processus de mesure. Elle jouera également un rôle central dans mon approche moins conventionnelle, notamment au niveau de son lien avec les procédures EP standard. Pour cette raison, il sera malheureusement nécessaire de faire une autre incursion dans le formalisme mathématique de la théorie quantique. J'espère qu'elle ne rebutera pas le lecteur non initié. Même s'il n'en retire pas une compréhension pleine et entière, je pense qu'elle lui facilitera la lecture des raisonnements mathématiques qu'il rencontrera par la suite, voire même qu'elle l'aidera à cerner les idées qui leur sont sous-jacentes. Elle lui sera en outre extrêmement précieuse pour comprendre certains des arguments que j'exposerai plus tard, ainsi que certaines subtilités expliquant pourquoi nous avons effectivement besoin d'une théorie quantique plus sophistiquée !

Une matrice densité représente un mélange statistique d'un certain nombre de vecteurs d'état orthogonaux — et non plus simplement un seul vecteur d'état. Le recours à ce mélange statistique signifie qu'il y a une certaine incertitude sur l'état réel du système et qu'à chaque vecteur d'état possible est attachée une probabilité exprimée, au sens ordinaire du terme, à l'aide d'un nombre réel. Une matrice densité introduit cependant une confusion (délibérée) entre les probabilités *classiques* associées à ce mélange et les probabilités *quantiques* résultant de la procédure **R**. L'idée est que puisqu'aucune expérience ne permet de distinguer entre ces deux types de probabilités, la description mathématique appropriée — la matrice densité — est celle qui, expérimentalement, *ne fait pas de distinction* entre eux.

Je ne donnerai ici que les concepts de base de cette description. La matrice densité est en fait une idée très élégante*. Tout d'abord, à la place de chaque état individuel $|\psi\rangle$, on utilise un objet noté

$$|\psi\rangle\langle\psi|\,.$$

* Elle fut introduite en 1932 par le brillant mathématicien américain d'origine hongroise John von Neumann. Ce dernier fut aussi le principal architecte de la théorie qui, se fondant sur les

La définition mathématique précise de cet objet importe peu ici. Disons seulement que cette expression représente le « produit » (analogue au produit tensoriel mentionné à la section 5.15) du vecteur d'état $|\psi\rangle$ avec son « complexe conjugué », noté $\langle\psi|$. $|\psi\rangle$ est un vecteur d'état *normalisé* ($\langle\psi|\psi\rangle = 1$), et $|\psi\rangle\langle\psi|$ est déterminé de manière unique par l'état physique représenté par le vecteur $|\psi\rangle$. (La modification du facteur de phase $e^{i\theta}$ faisant passer de $|\psi\rangle$ à $e^{i\theta}|\psi\rangle$ ne change rien à l'état physique du système ; *cf.* §5.10.) Dans la terminologie de Dirac, le vecteur $|\psi\rangle$ est un « ket » et $\langle\psi|$ est le « bra » qui lui correspond. On peut également combiner un « bra » $\langle\psi|$ et un « ket » $|\phi\rangle$ pour former leur *produit scalaire* (le *bracket**)

$$\langle\psi|\phi\rangle,$$

conformément à la notation de la section 5.12. Ce produit scalaire n'est qu'un nombre complexe, tandis que le produit tensoriel $|\psi\rangle\langle\phi|$ qui apparaît dans une matrice densité est un objet mathématique plus compliqué — un élément d'un certain espace vectoriel.

En effectuant l'opération mathématique qui consiste à « prendre la trace » de cet objet, on obtient un nombre complexe ordinaire. Pour une expression telle que $|\psi\rangle\langle\phi|$, « prendre la trace » équivaut simplement à inverser l'ordre des termes pour former le produit scalaire $\langle\phi|\psi\rangle$:

$$\text{trace}\ (|\psi\rangle\langle\phi|) = \langle\phi|\psi\rangle.$$

« Prendre la trace » est une opération linéaire ; par exemple,

$$\text{trace}\ (z\,|\psi\rangle\langle\phi| + w\,|\alpha\rangle\langle\beta|) = z\,\langle\phi|\psi\rangle + w\,\langle\beta|\alpha\rangle.$$

Sans entrer dans le détail de toutes les propriétés mathématiques d'objets tels que $\langle\psi|$ et $|\psi\rangle\langle\phi|$, précisons cependant que le produit $|\psi\rangle\langle\phi|$ vérifie les mêmes relations que celles énoncées p. 276 pour le produit $|\psi\rangle|\phi\rangle$ (à l'exception de la dernière qui n'interviendra pas ici) :

$$(z\,|\psi\rangle)\langle\phi| = z\,(|\psi\rangle\langle\phi|) = |\psi\rangle\,(z\,\langle\phi|),$$
$$(|\psi\rangle + |\chi\rangle)\langle\phi| = |\psi\rangle\langle\phi| + |\chi\rangle\langle\phi|,$$
$$|\psi\rangle\,(\langle\phi| + \langle\chi|) = |\psi\rangle\langle\phi| + |\phi\rangle\langle\chi|.$$

Notons également que le bra $\bar{z}\,\langle\psi|$ est le complexe conjugué du ket $\bar{z}\,|\psi\rangle$ ($\bar{z}$ étant le complexe conjugué du nombre complexe z ; *cf.* p. 252), et que $\langle\psi| + \langle\chi|$ est le complexe conjugué de $|\psi\rangle + |\chi\rangle$.

Supposons que nous ayons un mélange de deux états normalisés $|\alpha\rangle$ et $|\beta\rangle$, de probabilités respectives a et b. La matrice densité représentant ce mélange est alors

$$\boldsymbol{D} = a|\alpha\rangle\langle\alpha| + b|\beta\rangle\langle\beta|.$$

---

idées d'Alan Turing, permit l'avènement des ordinateurs. La théorie des jeux, mentionnée à la note 9 du chapitre 3, fut aussi l'œuvre de von Neumann. Mais ce que nous retiendrons surtout ici, c'est que von Neumann fut le premier à distinguer clairement entre les deux procédures quantiques que j'ai désignées par les lettres « **U** » et « **R** ».

* *Parenthèse* en anglais. Voir la note † p. 246. (N.d.T.)

Pour trois états normalisés $|\alpha\rangle$, $|\beta\rangle$, $|\gamma\rangle$ de probabilités respectives a, b et c, on a

$$\boldsymbol{D} = a|\alpha\rangle\langle\alpha| + b|\beta\rangle\langle\beta| + c|\gamma\rangle\langle\gamma|,$$

et ainsi de suite. La somme des probabilités de toutes les éventualités étant égale à 1, il en résulte l'importante propriété suivante, valable pour toute matrice densité $\boldsymbol{D}$:

$$\text{trace } \boldsymbol{D} = 1.$$

Comment utilise-t-on une matrice densité pour calculer les probabilités associées à une mesure ? Considérons d'abord le cas d'une mesure primitive et demandons-nous si le système est dans l'état physique $|\psi\rangle$ (**OUI**) ou dans un état orthogonal à $|\psi\rangle$ (**NON**). Cette mesure primitive est représentée par un objet mathématique (appelé *projecteur*) très semblable à une matrice densité et défini par :

$$\boldsymbol{E} = |\psi\rangle\langle\psi|.$$

La probabilité p d'obtenir **OUI** est alors

$$p = \text{trace } (\boldsymbol{DE}),$$

où le produit $\boldsymbol{DE}$ est lui-même un objet de type matrice densité qui s'obtient de manière presque ordinaire par les règles de l'algèbre — en faisant toutefois attention à l'ordre dans lequel sont effectuées les « multiplications ». Par exemple, pour la somme à deux termes $\boldsymbol{D} = a|\alpha\rangle\langle\alpha| + b|\beta\rangle\langle\beta|$, mentionnée plus haut, on a :

$$\begin{aligned}\boldsymbol{DE} &= (a|\alpha\rangle\langle\alpha| + b|\beta\rangle\langle\beta|)\,|\psi\rangle\langle\psi| \\ &= a|\alpha\rangle\langle\alpha|\psi\rangle\langle\psi| + b|\beta\rangle\langle\beta|\psi\rangle\langle\psi| \\ &= (a\langle\alpha|\psi\rangle)\,|\alpha\rangle\langle\psi| + (b\langle\beta|\psi\rangle)\,|\beta\rangle\langle\psi|.\end{aligned}$$

Les termes $\langle\alpha|\psi\rangle$ et $\langle\beta|\psi\rangle$ étant de simples nombres, on peut les intervertir avec d'autres expressions ; il faut en revanche être plus prudent sur l'ordre de « choses » telles que $\langle\alpha|$ et $\langle\beta|$. On en déduit (en se rappelant que $z\bar{z} = |z|^2$, *cf.* p. 252) que

$$\begin{aligned}\text{trace } (\boldsymbol{DE}) &= (a\langle\alpha|\psi\rangle)\langle\psi|\alpha\rangle + (b\langle\beta|\psi\rangle)\langle\psi|\beta\rangle \\ &= a|\langle\alpha|\psi\rangle|^2 + b|\langle\beta|\psi\rangle|^2.\end{aligned}$$

Rappelons (*cf.* §5.13) que $|\langle\alpha|\psi\rangle|^2$ et $|\langle\beta|\psi\rangle|^2$ sont respectivement les probabilités *quantiques* des résultats $|\alpha\rangle$ et $|\beta\rangle$, tandis que a et b sont les contributions *classiques* à la probabilité totale. Ainsi, l'expression finale est un mélange de probabilités quantiques et classiques.

Dans le cas plus général d'une mesure de type oui/non, la discussion est similaire, mais à la place du « $\boldsymbol{E}$ » défini plus haut, on utilise un projecteur plus général tel que :

$$\boldsymbol{E} = |\psi\rangle\langle\psi| + |\phi\rangle\langle\phi| + \ldots + |\chi\rangle\langle\chi| .$$

où $|\psi\rangle$, $|\phi\rangle$, ..., $|\chi\rangle$ sont des états normalisés orthogonaux engendrant l'espace des états **OUI** dans l'espace de Hilbert. La propriété

$$\boldsymbol{E}^2 = \boldsymbol{E}$$

est caractéristique des projecteurs. La probabilité pour que la mesure définie par le projecteur $\boldsymbol{E}$ et effectuée sur le système doté de la matrice densité $\boldsymbol{D}$ soit **OUI** est égale à trace $(\boldsymbol{DE})$, exactement comme plus haut.

Il importe de remarquer que pour calculer la probabilité recherchée, il suffit de connaître la matrice densité et le projecteur décrivant la mesure. Il n'est pas nécessaire de connaître la manière particulière dont la matrice densité a été constituée à partir des états. La probabilité totale apparaît automatiquement sous forme d'une combinaison de probabilités quantiques et classiques, sans que l'on ait à se soucier des contributions de chacune de ces probabilités à la probabilité résultante.

Examinons de plus près cet étrange enchevêtrement de probabilités quantiques et classiques au sein de la matrice densité. Supposons par exemple que nous ayons une particule de spin $\frac{1}{2}$ et que nous ignorions totalement si l'état de spin (normalisé) est $|\uparrow\rangle$ ou $|\downarrow\rangle$. Les probabilités associées à ces deux états étant toutes deux $\frac{1}{2}$, la matrice densité est alors

$$\boldsymbol{D} = \tfrac{1}{2}|\uparrow\rangle\langle\uparrow| + \tfrac{1}{2}|\downarrow\rangle\langle\downarrow| .$$

Un calcul élémentaire montre toutefois que l'on obtient exactement la *même* matrice densité $\boldsymbol{D}$ en partant de n'importe quelles autres possibilités orthogonales de probabilités $\frac{1}{2}$ et $\frac{1}{2}$, par exemple des états (normalisés) $|\rightarrow\rangle$ et $|\leftarrow\rangle$ (où $|\rightarrow\rangle = (|\uparrow\rangle + |\downarrow\rangle)/\sqrt{2}$ et $|\leftarrow\rangle = (|\uparrow\rangle - |\downarrow\rangle)/\sqrt{2}$) :

$$\boldsymbol{D} = \tfrac{1}{2}|\rightarrow\rangle\langle\rightarrow| + \tfrac{1}{2}|\leftarrow\rangle\langle\leftarrow| .$$

Supposons que nous choisissions de mesurer le spin de la particule dans la direction *up*. Le projecteur correspondant est alors

$$\boldsymbol{E} = |\uparrow\rangle\langle\uparrow| .$$

La probabilité de **OUI** dans la première description est égale à

$$\begin{aligned} \text{trace } (\boldsymbol{DE}) &= \tfrac{1}{2}|\langle\uparrow|\uparrow\rangle|^2 + \tfrac{1}{2}|\langle\downarrow|\uparrow\rangle|^2 \\ &= \tfrac{1}{2}\times 1^2 + \tfrac{1}{2}\times 0^2 \\ &= \tfrac{1}{2}, \end{aligned}$$

où l'on a utilisé les relations $\langle\uparrow|\uparrow\rangle = 1$ et $\langle\downarrow|\uparrow\rangle = 0$ (ces états sont normalisés et orthogonaux). Dans la seconde description, on a

$$\begin{aligned}\text{trace } (\boldsymbol{DE}) &= \tfrac{1}{2}|\langle\rightarrow|\uparrow\rangle|^2 + \tfrac{1}{2}|\langle\leftarrow|\uparrow\rangle|^2 \\ &= \tfrac{1}{2}\times (1/\sqrt{2})^2 + \tfrac{1}{2}\times (1/\sqrt{2})^2 \\ &= \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2},\end{aligned}$$

où les états droite/gauche $|\rightarrow\rangle$ et $|\leftarrow\rangle$ ne sont ni orthogonaux ni parallèles à l'état mesuré $|\uparrow\rangle$ (on a en fait $|\langle\rightarrow|\uparrow\rangle| = |\langle\leftarrow|\uparrow\rangle| = 1/\sqrt{2}$).

Bien que les probabilités résultantes soient identiques (comme il se doit, car la matrice densité est la même), les interprétations de ces deux descriptions sont tout à fait différentes. Certes la « réalité » physique de toute situation se décrit par *un* vecteur d'état bien précis, mais il existe, en termes classiques, une incertitude sur le vecteur d'état réellement associé au système. Dans la première des deux descriptions ci-dessus, on ignore l'état du système, mais on sait qu'il est soit $|\uparrow\rangle$, soit $|\downarrow\rangle$. Dans la seconde, on ignore toujours l'état du système, mais on sait qu'il est soit $|\rightarrow\rangle$, soit $|\leftarrow\rangle$. Dans la première description, lorsqu'on effectue une mesure pour savoir si l'état est $|\uparrow\rangle$, le problème est une simple affaire de probabilités classiques : il y a effectivement une probabilité $\frac{1}{2}$ pour que cet état soit $|\uparrow\rangle$. Dans la seconde description, la mesure de $|\uparrow\rangle$ s'effectue en intervenant sur un mélange statistique de $|\rightarrow\rangle$ et $|\leftarrow\rangle$, et la contribution de chacun de ces deux états est égale à une contribution classique $\frac{1}{2}$ fois une contribution quantique $\frac{1}{2}$, ce qui donne au total $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. On voit donc que la matrice densité s'arrange intelligemment pour donner la probabilité correcte, indépendamment de la décomposition de cette probabilité en parties classique et quantique.

L'exemple ci-dessus est un peu particulier dans la mesure où la matrice densité y possède ce que l'on appelle des « valeurs propres dégénérées » (*i.e.* les deux probabilités classiques sont égales à $\frac{1}{2}$), ce qui permet d'avoir plus d'une description en termes de mélange statistique d'états orthogonaux. Cette particularité n'est toutefois pas essentielle pour notre discussion. (Je la mentionne uniquement pour rassurer les spécialistes.) Rien n'oblige un mélange statistique d'états à être limité à un ensemble d'états orthogonaux. Par exemple, dans la situation ci-dessus, nous pourrions avoir des mélanges statistiques compliqués composés de multiples directions de spin. Il s'avère que *toute* matrice densité — et non seulement celles possédant des valeurs propres dégénérées — admet une multitude de représentations sous forme d'un mélange statistique d'états orthogonaux.

6.5 Matrices densité pour paires EPR

Examinons maintenant une situation dans laquelle la description en termes de matrice densité est particulièrement appropriée — bien qu'elle fasse

ressortir un aspect presque paradoxal de son interprétation. Cette situation est liée aux effets EPR et d'emmêlement quantique. Considérons la situation physique décrite à la section 5.17, dans laquelle une particule de spin 0 (dans l'état $|\Omega\rangle$) se scinde en deux particules de spin $\frac{1}{2}$ qui s'éloignent ensuite considérablement l'une de l'autre, disons vers la droite et vers la gauche. Leur état de spin (emmêlé) a pour expression :

$$|\Omega\rangle = |G\uparrow\rangle|D\downarrow\rangle - |G\downarrow\rangle|D\uparrow\rangle .$$

Supposons qu'un observateur (ou une observatrice !) entreprenne d'examiner le spin de la particule de droite à l'aide d'un appareil de mesure, mais que la particule de gauche soit si éloignée qu'elle lui soit inaccessible. Comment cet observateur pourrait-il décrire le spin de la particule de droite ?

Il serait bien inspiré d'utiliser la matrice densité

$$\boldsymbol{D} = \tfrac{1}{2}|D\uparrow\rangle\langle D\uparrow| + \tfrac{1}{2}|D\downarrow\rangle\langle D\downarrow| .$$

Il pourrait en effet imaginer qu'un autre observateur — un collègue situé très loin de lui — ait choisi de mesurer le spin de la particule de gauche dans une direction *up/down*. Il n'a aucun moyen de savoir quel résultat son collègue imaginaire pourrait obtenir lors de cette mesure de spin, mais il sait que si son collègue a obtenu le résultat $|G\uparrow\rangle$, l'état de sa propre particule est alors nécessairement $|D\downarrow\rangle$, tandis que si son collègue a obtenu $|G\downarrow\rangle$, l'état de sa propre particule est alors nécessairement $|D\uparrow\rangle$. Il sait également (d'après les règles standard de la théorie quantique sur les probabilités associées à ce genre de situation) que son collègue a autant de chances d'obtenir $|G\uparrow\rangle$ que $|G\downarrow\rangle$. Il en conclut que l'état de sa propre particule est un mélange statistique des deux éventualités équiprobables (*i.e.* de probabilités respectives $\frac{1}{2}$, $\frac{1}{2}$) $|D\uparrow\rangle$ et $|D\downarrow\rangle$, de sorte que sa matrice densité est effectivement le $\boldsymbol{D}$ tel qu'il est exprimé ci-dessus.

Il pourrait cependant imaginer que son collègue vient de mesurer le spin de la particule de gauche, non pas dans la direction *up/down*, mais dans la direction gauche/droite. Un raisonnement identique (fondé maintenant sur l'état de spin décrit par $|\Omega\rangle = |G\leftarrow\rangle\langle D\rightarrow| - |G\rightarrow\rangle\langle D\leftarrow|$, *cf.* p. **281**) l'amènerait à conclure que l'état de spin de sa propre particule est un mélange équiprobable de droite et de gauche, ce qui lui donnerait la matrice densité

$$\boldsymbol{D} = \tfrac{1}{2}|D\rightarrow\rangle\langle D\rightarrow| + \tfrac{1}{2}|D\leftarrow\rangle\langle D\leftarrow| .$$

Nous l'avons vu, c'est exactement la même matrice densité qu'il avait lors du premier raisonnement, mais son *interprétation* en tant que mélange statistique d'états orthogonaux est tout à fait différente ! Peu importe l'interprétation adoptée par l'observateur. Sa matrice densité lui donne toute l'information dont il peut disposer pour calculer les probabilités associées aux résultats des mesures de spin effectuées sur la seule particule de droite. En outre, puisque son collègue est simplement *imaginé*, notre observateur n'a nullement besoin de considérer que la particule de gauche a été soumise à la moindre mesure de spin. Cette même matrice densité $\boldsymbol{D}$ lui dit tout ce qu'il peut connaître sur l'état de spin de la particule de droite avant qu'il effectue une mesure sur cette

particule. De fait, nous pourrions considérer que l'« état réel » de la particule de droite est décrit plus correctement par la matrice densité $\boldsymbol{D}$ que par n'importe quel vecteur d'état.

Ce type de considérations conduit parfois les gens à penser que, dans certaines circonstances, les matrices densité donnent une représentation plus fidèle de la « réalité » quantique que ne le font les vecteurs d'état. Cela *ne permet pas*, cependant, d'avoir une vision globale d'une situation comme celle que nous venons d'envisager. Car rien en principe n'interdit au collègue imaginaire de notre observateur de devenir réel, et que tous deux se communiquent leurs résultats. Les emmêlements existant entre les mesures de ces deux observateurs ne peuvent s'expliquer en termes de matrices densité distinctes, une pour la particule de droite et une pour la particule de gauche. Pour expliquer ces emmêlements, il faut recourir à l'état quantique emmêlé complet fourni par le vecteur d'état réel $|\Omega\rangle$ tel qu'il a été donné plus haut.

Par exemple, si les deux observateurs décident de mesurer les spins de leurs particules dans la direction *up/down*, ils doivent alors nécessairement obtenir des réponses opposées. Des matrices densité individuelles pour ces deux particules ne fournissent pas une telle information. Plus grave même, le théorème de Bell (*cf.* §5.4) montre qu'il n'existe *aucune* description locale classique (de type « chaussettes de Bertlmann ») de l'état quantique emmêlé de ces deux particules avant la mesure. (Voir EOLP, chapitre 6, p.310 et note 24, p. 504, pour une démonstration élémentaire de ce fait — due essentiellement à Stapp (1979) ; *cf.* aussi Stapp 1993 — dans le cas où l'un des observateurs décide de mesurer le spin de sa particule soit dans la direction *up/down*, soit dans la direction droite/gauche, tandis que l'autre observateur choisit l'une des deux directions inclinées à 45° par rapport aux directions de son collègue. Cette démonstration est encore plus convaincante — grâce aux dodécaèdres magiques de la section 5.3 — si l'on remplace les deux particules de spin $\frac{1}{2}$ par deux particules de spin $\frac{3}{2}$, car aucune probabilité n'intervient alors.)

Cela montre que la matrice densité ne décrit de façon satisfaisante la « réalité » de cette situation que s'il existe une raison *de principe* pour laquelle on ne peut effectuer et comparer les mesures sur les deux parties du système. Rien, dans les situations normales, ne semble l'interdire. Certaines situations anormales — telle celle envisagée par Stephen Hawking (1982) et dans laquelle une particule d'une paire EPR pourrait être capturée par un trou noir — autoriseraient, à un niveau théorique fondamental, des descriptions à l'aide de matrices densité (ainsi que l'affirme Hawking). Mais cela constituerait en soi une modification de la structure de la théorie quantique. Sans une telle modification, le statut de la matrice densité reste essentiellement EP, et non pas fondamental — tout en étant bien sûr très important.

6.6 Une explication EP de **R** ?

Regardons maintenant le rôle des matrices densité dans l'approche standard — EP — visant à expliquer l'intervention « apparente » du processus **R**. L'idée est que l'ensemble formé par un système quantique, un appareil de mesure et l'environnement dans lequel ils se situent — tout cela étant censé évoluer selon **U** — se comporte, chaque fois que les effets d'une mesure s'emmêlent intimement avec cet environnement, *comme s'il* y avait intervention de **R**.

Le système quantique est supposé initialement isolé de son environnement. Dès qu'il est « mesuré », il déclenche dans l'appareil de mesure des effets macroscopiques engendrant rapidement des emmêlements avec des parties considérables et sans cesse croissantes de cet environnement. À ce stade, la situation ressemble à bien des égards à la situation EPR analysée à la section précédente, où le système quantique et l'appareil de mesure qu'il vient de déclencher joueraient le rôle de la particule de droite, tandis que l'environnement perturbé jouerait celui de la particule de gauche. Un physicien entreprenant d'examiner l'appareil de mesure jouerait un rôle analogue à celui de l'observateur qui, dans la discussion précédente, examinait la particule de droite. De même que cet observateur n'avait aucun accès aux mesures effectuées sur la particule de gauche, de même notre physicien ignorerait la perturbation provoquée par l'appareil de mesure sur l'environnement. Cet environnement se composant d'un nombre énorme de particules animées de mouvements aléatoires, on peut légitimement supposer que ce physicien n'aurait strictement aucun moyen de connaître concrètement le détail des perturbations auxquelles seraient soumises ces particules. Cette situation est analogue au fait que toute information sur le spin de la particule de gauche, dans l'exemple précédent, est inaccessible à l'observateur de la particule de droite. À l'instar de ce qui se passait pour la particule de droite, la description adéquate de l'état de l'appareil de mesure est celle donnée non par un état quantique pur, mais par une matrice densité, *i.e.* par un mélange statistique d'états. L'idée est donc que ce mélange statistique fournit les résultats que donnerait — du moins EP – la procédure **R**.

Supposons par exemple qu'une source de photons émette un photon en direction d'un détecteur. Entre la source et le détecteur est interposé un miroir partiellement argenté, de sorte qu'après sa rencontre avec ce miroir, le photon se trouve dans l'état de superposition

$$w|\alpha\rangle + z|\beta\rangle\,,$$

où l'état transmis $|\alpha\rangle$ active le détecteur (**OUI**) et l'état réfléchi $|\beta\rangle$ ne l'active pas (**NON**). Je suppose ici que tous les états sont normalisés, de sorte que selon la procédure **R**, on obtient :

$$\text{probabilité de } \mathbf{OUI} = |w|^2 \;;\; \text{probabilité de } \mathbf{NON} = |z|^2\,.$$

Dans le cas d'un miroir *semi*-argenté (comme dans l'exemple considéré à la section 5.7, où nos $|\alpha\rangle$ et $|\beta\rangle$ étaient respectivement les états $|\mathrm{B}\rangle$ et $\mathrm{i}|\mathrm{C}\rangle$), ces deux probabilités valent chacune $\frac{1}{2}$ et l'on a $|w| = |z| = 1/\sqrt{2}$.

L'état initial $|\Psi\rangle$ du détecteur se transforme en $|\Psi_O\rangle$ (**OUI**) en cas d'absorption du photon (dans l'état $|\alpha\rangle$) et en $|\Psi_N\rangle$ (**NON**) en cas de non-absorption du photon (dans l'état $|\beta\rangle$). Si l'on pouvait ignorer l'environnement, l'état du détecteur serait alors

$$w|\Psi_O\rangle + z|\Psi_N\rangle|\beta\rangle$$

(tous les états étant supposés normalisés) ; supposons cependant que le détecteur, qui est un objet macroscopique, s'engage très rapidement dans des interactions avec son environnement — nous pouvons également supposer que le photon non capturé (initialement dans l'état $|\beta\rangle$) est absorbé par les murs du laboratoire pour devenir partie intégrante de cet environnement. Comme avant, selon qu'il a ou non reçu le photon, le détecteur s'installe dans l'état $|\Psi_O\rangle$ ou $|\Psi_N\rangle$, mais dans le même temps perturbe l'environnement, de manière différente selon qu'il est dans l'un ou l'autre de ces états. Si on désigne par $|\Phi_O\rangle$ et $|\Phi_N\rangle$ les états de l'environnement respectivement associés à $|\Psi_O\rangle$ et $|\Psi_N\rangle$ (en les supposant normalisés mais non nécessairement orthogonaux), l'état total prend la forme emmêlée

$$w|\Phi_O\rangle|\Psi_O\rangle + z|\Phi_N\rangle|\Psi_N\rangle\,.$$

Cet état ne tient pas compte du physicien qui entreprend d'examiner le détecteur pour voir s'il affiche **OUI** ou **NON**. Quel est de son point de vue l'état du détecteur juste avant qu'il ne l'examine ? À l'instar de l'observateur mesurant le spin de la particule de droite dans la discussion de la section précédente, il serait bien inspiré d'utiliser une matrice densité. Tout comme avec la particule de gauche dans la paire EPR décrite plus haut, nous pouvons en effet supposer qu'aucune mesure n'est effectuée sur l'environnement pour savoir si *son* état est $|\Phi_O\rangle$ ou $|\Phi_N\rangle$, et donc que la matrice densité fournit à notre physicien une description quantique adéquate du détecteur.

Quelle est cette matrice densité ? Le raisonnement standard[7] (fondé sur un modèle particulier de cet environnement — et sur quelques hypothèses qui ne sont que partiellement justifiées, telles la non-prise en compte d'un emmêlement de type EPR) conduit à la conclusion que cette matrice densité devrait rapidement approcher de très près la forme

$$\boldsymbol{D} = a|\Psi_O\rangle\langle\Psi_O| + b|\Psi_N\rangle\langle\Psi_N|\,,$$

où

$$a = |w|^2 \text{ et } b = |z|^2.$$

On peut considérer que cette matrice densité représente un mélange statistique des états **OUI** et **NON**, avec comme probabilités respectives $|w|^2$ et $|z|^2$. C'est exactement ce que la procédure **R** nous dit que trouverait le physicien suite à son expérience. Mais est-ce bien vrai ? Cette conclusion n'est-elle pas trop précipitée ?

Si notre physicien pouvait faire l'*hypothèse* que les états accessibles au détecteur sont uniquement $|\Psi_O\rangle$ *ou* $|\Psi_N\rangle$, la matrice densité $\boldsymbol{D}$ lui permettrait effectivement de calculer les probabilités dont il a besoin. Mais cette

hypothèse n'est nullement une conséquence de notre discussion. Nous l'avons vu à la section précédente, les matrices densité admettent de nombreuses interprétations *différentes* en tant que mélanges statistiques d'états. En particulier, dans le cas d'un miroir *semi*-argenté, on obtient une matrice densité dont la forme est exactement celle que nous avons obtenue plus haut pour la particule de spin $\frac{1}{2}$, à savoir

$$\boldsymbol{D} = \tfrac{1}{2}|\Psi_O\rangle\langle\Psi_O| + \tfrac{1}{2}|\Psi_N\rangle\langle\Psi_N| .$$

On peut récrire $\boldsymbol{D}$ sous une autre forme, par exemple

$$\boldsymbol{D} = \tfrac{1}{2}|\Psi_P\rangle\langle\Psi_P| + \tfrac{1}{2}|\Psi_Q\rangle\langle\Psi_Q| ,$$

où $|\Psi_P\rangle$ et $|\Psi_O\rangle$ sont

$$|\Psi_P\rangle = (|\Psi_O\rangle + |\Psi_N\rangle) / \sqrt{2} \text{ et } |\Psi_Q\rangle = (|\Psi_O\rangle - |\Psi_N\rangle) / \sqrt{2} ,$$

deux états du détecteur orthogonaux, totalement différents — et totalement absurdes du point de vue de la physique classique !

Le fait que le physicien considère que l'état de son détecteur est décrit par la matrice densité $\boldsymbol{D}$ n'explique en rien pourquoi il trouve toujours que ce détecteur est soit dans un état **OUI** (donné par $|\Psi_O\rangle$), soit dans un état **NON** (donné par $|\Psi_N\rangle$). Car ce physicien obtiendrait exactement la même matrice densité si l'état du détecteur était une combinaison équiprobable des absurdités classiques $|\Psi_P\rangle$ et $|\Psi_Q\rangle$ (décrivant respectivement les superpositions linéaires quantiques « **OUI** *plus* **NON** » et « **OUI** *moins* **NON** ») !

Pour souligner l'absurdité d'états tels que $|\Psi_P\rangle$ et $|\Psi_Q\rangle$ dans le cas d'un détecteur macroscopique, considérez un « appareil de mesure » composé d'une boîte contenant un chat — un *chat de Schrödinger* (*cf.* §5.1 et Fig. 6.3) — qui se fait tuer par un dispositif *ad hoc* si le détecteur reçoit un photon (dans l'état $|\alpha\rangle$) et reste vivant si le détecteur ne reçoit pas le photon (qui est alors dans l'état $|\beta\rangle$). Les réponses **OUI** et **NON** correspondraient respectivement à « chat mort » et à « chat vivant ». Toutefois, le simple fait de savoir que la matrice densité est une combinaison équiprobable de ces deux états *ne nous dit pas* que le chat est soit mort soit vivant (avec des probabilités égales), car il pourrait aussi bien être « mort plus vivant » ou « mort moins vivant », avec des probabilités égales ! D'elle-même, la matrice densité *ne nous dit pas* que nous ne rencontrerons jamais, dans le monde où nous vivons, ces deux dernières possibilités absurdes du point de vue de la physique classique. À l'instar de l'interprétation de **R** dans une approche de type « mondes multiples », il semble que nous soyons encore acculés à nous demander quels types d'états sont perceptibles par un observateur conscient (en l'occurrence, notre « physicien »). Autrement dit, qu'est-ce qui interdit à un observateur extérieur* conscient de jamais percevoir un état du genre « chat mort plus chat vivant » ?

---

* Bien sûr, il faudrait également considérer le problème de la conscience du chat lui-même ! Ce point a été clairement mis en lumière grâce à une version du paradoxe du chat de Schrödinger imaginée par Eugene P. Wigner (1961). L'« ami de Wigner » est exposé aux mêmes désagréments que le chat de Schrödinger, mais est pleinement conscient de chacun de ses états de superposition !

On pourrait répondre que la « mesure » que notre physicien est sur le point d'effectuer sur le détecteur consiste simplement, en définitive, à déterminer si le détecteur affiche **OUI** ou **NON** — *i.e.* à vérifier, dans cet exemple, si le chat est mort ou vivant. (Cela équivaut, pour l'observateur de la section précédente, à déterminer si le spin de la particule de droite est *up* ou *down*.) Pour cette mesure, la matrice densité donne effectivement les probabilités correctes, quelle que soit la manière dont nous l'écrivions. Mais cela revient en fait à éluder la question. Nous devons nous demander pourquoi un simple *regard* posé sur le chat correspond de fait à la réalisation d'une mesure de ce type. Rien, dans la seule évolution **U** d'un système quantique, n'interdit que lorsque nous « regardons » et par conséquent *percevons* un système quantique, notre conscience ne puisse rencontrer la combinaison « chat mort plus chat vivant ». Nous voici ramenés à notre point de départ. Qu'est-ce que la conscience ? Comment est *réellement* construit notre cerveau ? C'est précisément parce qu'ils voulaient *éviter* ces interrogations que les physiciens ont donné à **R** une interprétation EP !

Certains lecteurs pourraient objecter que l'exemple précédent n'est pas représentatif dans la mesure où les deux probabilités $\frac{1}{2}$ et $\frac{1}{2}$ sont *égales* (cas des « valeurs propres dégénérées ») et que ce n'est que dans de telles situations que la matrice densité admet plusieurs expressions sous forme de mélanges statistiques d'états *orthogonaux*. Cette restriction *n'est pas* importante, car l'interprétation de la matrice densité en termes de mélange statistique n'exige pas l'orthogonalité des états. En fait, un article récent de Hughston *et al.* (1993) a montré que dans les situations analogues à celle considérée ici, où la matrice densité apparaît parce que le système envisagé est emmêlé avec un autre système, il existe toujours, *quel que soit* le mélange statistique adopté pour représenter la matrice densité, une mesure qui, effectuée sur ce second système, donne cette représentation particulière de la matrice densité. En tout cas, la seule présence de cette ambiguïté lorsque les probabilités *sont* égales indique que la matrice densité ne suffit pas à décrire parfaitement les états réels de notre détecteur.

La morale de tout cela est que le simple fait de savoir que la matrice densité est un certain ***D*** *ne nous dit pas* que le système est un mélange statistique des états particuliers qui ont donné naissance à ce ***D*** particulier. Il existe de nombreuses manières totalement différentes d'obtenir le même ***D***, la plupart d'entre elles étant d'ailleurs « absurdes » du point de vue du sens commun. En outre, cette ambiguïté est présente quelle que soit la matrice densité.

D'une manière générale, les discussions standard se bornent à démontrer que la matrice densité est « diagonale ». Cela signifie qu'on peut l'exprimer sous forme d'un mélange statistique d'états *orthogonaux* — ou plutôt qu'il en est ainsi quand les états sont les états classiques auxquels on s'intéresse. (Sans cette dernière restriction, *toutes* les matrices densité seraient diagonales !) Mais nous avons vu que le simple fait que la matrice densité admette une telle expression n'interdit pas en soi que des détecteurs se trouvent dans des superpositions quantiques « absurdes » de **OUI** et de **NON** simultanés.

Ainsi, contrairement à ce que l'on affirme souvent, le raisonnement standard *n'explique pas* comment l'« illusion » de **R** peut être une forme de

description approximative de l'évolution **U** lorsque l'environnement devient inextricablement complexe. Ce que montre *seulement* ce raisonnement, c'est que dans une telle situation, la procédure **R** peut coexister tranquillement avec l'évolution **U**. La théorie quantique ne peut se dispenser de recourir à une procédure **R** distincte de l'évolution **U** (du moins en l'absence d'une théorie définissant les états perceptibles par des êtres conscients).

Ce fait est en lui-même important pour la cohérence générale de la théorie quantique. Mais il importe également d'avoir à l'esprit que le statut de cette coexistence et de cette cohérence est plus EP que rigoureusement logique. Rappelons que, d'après la discussion finale de la section précédente, l'adéquation de la description de la particule de droite en termes de matrice densité résultait uniquement de l'impossibilité de comparer les mesures effectuées sur *chacune* des deux particules. Pour pouvoir effectuer une telle comparaison, il aurait fallu connaître l'état quantique entier, avec ses superpositions *quantiques*, et non pas simplement ses superpositions pondérées par des probabilités. De même dans la présente discussion, la description du détecteur à l'aide de la matrice densité ne convient que si l'on ne peut mesurer les détails fins de l'environnement et comparer le résultat de cette mesure aux observations que l'expérimentateur effectue sur le détecteur. **R** peut coexister avec **U** seulement si les détails fins de l'environnement échappent aux mesures ; et si tel est le cas, on ne peut observer les effets subtils des interférences quantiques qui (selon la théorie quantique standard) sont dissimulés dans l'immense complexité de la description détaillée de l'environnement.

Le raisonnement standard contient bien entendu une bonne part de vérité ; pourtant, il n'apporte pas de réponse complète. Comment peut-on être certain que des progrès technologiques futurs ne permettront pas de découvrir un jour les effets de ces interférences ? Nous aurions besoin d'une nouvelle règle physique nous disant que certaines expériences que l'on ne peut aujourd'hui effectuer concrètement ne pourront jamais être effectuées même *en principe*, qu'il existe un niveau d'action physique auquel il est en principe impossible de déceler ces effets d'interférences. C'est seulement avec la découverte d'un *nouveau* phénomène physique que les superpositions complexes de la physique quantique perdront leur statut de simples alternatives EP pour devenir *réellement* des alternatives classiques. Tel qu'il existe actuellement, le point de vue pragmatique ne nous donne aucune image d'une véritable réalité physique. Ce point de vue n'est qu'une théorie physique provisoire — précieuse certes — et cela jouera un grand rôle dans les propositions que j'émettrai à la section 6.12.

6.7 L'approche EP permet-elle d'expliquer la règle du module au carré ?

Les trois précédentes sections contiennent une hypothèse qui a pu passer inaperçue. La *seule* nécessité de cette hypothèse écarte toute éventualité d'une *déduction* de la règle du module au carré associée à la procédure **R** à partir de l'évolution **U** — même EP. Nous avons en effet implicitement *supposé* qu'une matrice densité décrit correctement un mélange statistique. L'adéquation même d'une expression comme $|\alpha\rangle\langle\alpha|$, de la forme « objet fois son complexe conjugué », est intimement liée à l'hypothèse du carré du module. C'est uniquement parce que la règle du module au carré est inscrite dans la structure même d'une matrice densité que la règle d'obtention des probabilités à partir d'une matrice densité associe correctement probabilités classiques et quantiques.

S'il est effectivement vrai, mathématiquement parlant, que l'évolution unitaire **U** cadre bien avec les notions de matrice densité et de produit scalaire hilbertien ($\langle\alpha|\beta\rangle$), elle ne nous *dit* nullement que les grandeurs que permettent de calculer les modules au carré sont des *probabilités*. Ici encore, c'est une affaire de simple coexistence entre **R** et **U**, et non une explication de **R** à partir de **U**. L'évolution unitaire reste totalement muette sur la notion de probabilité. Le fait qu'on puisse calculer des probabilités quantiques grâce à cette procédure résulte clairement d'une hypothèse *supplémentaire*, indépendante de l'approche — qu'elle soit de type mondes multiples ou de type EP — utilisée pour justifier la compatibilité de **R** avec **U**.

Puisqu'une bonne part du matériau expérimental qui conforte la mécanique quantique résulte de la manière même dont cette théorie nous dit qu'il faut calculer des probabilités, on ne peut impunément ignorer la partie **R** de la mécanique quantique. **R** est une procédure très différente de **U** ; elle n'en est aucunement une conséquence, quelle que soit l'obstination avec laquelle les théoriciens ont tenté de montrer le contraire. **R** n'étant pas une conséquence de **U**, nous devons la traiter comme un processus physique à part entière. Cela ne signifie toutefois pas qu'elle soit une *loi* physique à part entière. Nul doute qu'elle est une approximation de quelque chose d'autre que nous ne comprenons pas encore. Les discussions de la fin de la section précédente ont fortement suggéré que le recours à la procédure **R** lors du processus de mesure est effectivement une approximation.

Acceptons donc que nous ayons besoin de quelque chose de nouveau et aventurons-nous, avec la prudence qui s'impose, sur les diverses voies inconnues qui s'ouvrent dès lors à nous.

6.8 Est-ce la conscience qui réduit le vecteur d'état ?

Certains, parmi ceux qui considèrent sérieusement que $|\psi\rangle$ est une description du monde physique, affirment — pour éluder l'action de **U** à toutes les échelles et donc le concept de mondes multiples — que l'intervention de la conscience d'un observateur provoque l'apparition d'un processus de type **R**. L'éminent physicien Eugene Wigner a émis une théorie de ce genre (Wigner 1961) : la matière inconsciente — ou peut-être seulement la matière inanimée — évolue selon **U**, mais dès qu'une entité consciente (ou la « vie ») s'emmêle physiquement avec l'état, quelque chose de nouveau se produit, et un processus physique aboutissant à **R** se substitue à **U** pour réduire *réellement* le vecteur d'état.

Un tel point de vue ne suggère nullement que l'entité consciente « influencerait » le choix particulier fait à ce stade par la nature. Une telle suggestion nous conduirait dans des eaux extrêmement troubles et, pour autant que je sache, toute suggestion trop simpliste d'une influence exercée par un acte de volonté conscient sur le résultat d'une expérience quantique s'opposerait gravement aux faits observés. Ainsi, nous *supposerons* ici que le « libre arbitre » ne joue aucun rôle actif dans la procédure **R** (voyez toutefois en section 7.1 quelques autres points de vue possibles).

Certains lecteurs penseront probablement que, puisque je cherche à établir un lien entre le problème de la mesure quantique et celui de la conscience, je dois être séduit par des idées de ce genre. Je tiens à dire clairement qu'*il n'en est rien.* La conscience est probablement un phénomène assez rare dans l'Univers. Si elle semble se manifester avec force en de nombreux endroits à la surface de la Terre, les témoignages dont on dispose à ce jour[8] montrent qu'il n'existe aucune conscience hautement développée — voire même aucune conscience du tout — dans les profondeurs de l'Univers, à de nombreux siècles-lumière à la ronde. Ce serait une image très étrange d'un univers physique « réel » que celle dans laquelle les objets évolueraient de manière totalement différente selon qu'ils seraient ou non vus, entendus ou touchés par l'un de ses habitants conscients.

Considérez par exemple la météorologie. Les situations météo qui se développent sur toute planète dépendent de processus physiques chaotiques (*cf.* §1.7) et sont donc sensibles à de multiples événements quantiques. Si le processus **R** ne survient qu'en présence de conscience, aucune situation météo ne peut alors émerger du fatras des superpositions d'états quantiques. Peut-on réellement croire que les situations météo d'une planète lointaine demeurent sous forme de superpositions d'innombrables possibilités distinctes affectées de coefficients complexes — dans une confusion très différente d'une situation météo réelle — jusqu'à ce qu'un être conscient en ait connaissance, et qu'alors *seulement* cette superposition se transforme en une situation météo réelle ?

On pourrait objecter que d'un point de vue opérationaliste — *i.e.* du point de vue opérationaliste d'un être conscient — une telle superposition de « météos » ne différerait pas d'une météo qui serait *réelle* mais incertaine

(EP !). Cela ne constitue toutefois pas une explication satisfaisante du problème de la réalité physique. Nous avons vu que la position EP ne permet pas de résoudre le problème de la « réalité », mais reste un bouche-trou autorisant la coexistence des procédures **U** et **R** de la mécanique quantique actuelle — du moins jusqu'au jour où notre technologie nous obligera à adopter une attitude plus précise et plus cohérente.

Ainsi, je propose de rechercher ailleurs une solution aux problèmes de la mécanique quantique. Si le problème de l'esprit s'avère effectivement lié à celui de la mesure quantique — ou au paradoxe **U**/**R** de la mécanique quantique —, ce n'est pas, selon moi, la conscience (ou la forme de conscience qui nous est familière) qui, par elle-même, résoudra les problèmes physiques internes à la théorie quantique. Je suis convaincu que nous devons attaquer et résoudre le problème de la mesure quantique bien avant d'espérer progresser réellement sur la voie d'une interprétation de la conscience en termes d'action physique — et que le problème de la mesure doit être résolu en termes entièrement *physiques*. C'est seulement une fois que nous disposerons d'une solution satisfaisante que nous pourrons peut-être apporter une forme de réponse au problème de la conscience. J'ai la conviction que la compréhension de la mesure quantique est un *préalable* à la compréhension de l'esprit, et qu'elles ne constituent *nullement* un seul et même problème. Le problème de l'esprit est bien plus ardu que celui de la mesure !

6.9 Où l'on prend $|\psi\rangle$ vraiment au sérieux

Telles que m'apparaissent les choses, les points de vue qui affirment prendre au sérieux la description quantique du monde sont loin de la prendre *réellement* au sérieux. Peut-être le formalisme quantique est-il trop étrange pour que l'on se risque à le prendre au pied de la lettre et que la plupart des physiciens évitent de se définir clairement sur ce point. Car outre un vecteur d'état $|\psi\rangle$ évoluant selon **U** — tant que le système reste au niveau quantique —, on a également l'action troublante de la procédure discontinue et probabiliste **R** qui semble devoir être introduite pour faire « sauter » $|\psi\rangle$ dès que les effets quantiques s'amplifient suffisamment pour avoir une influence au niveau classique. Ainsi, si on considère que $|\psi\rangle$ donne une image de la *réalité*, on doit alors également considérer ces *sauts* comme physiquement réels, quel que soit l'inconfort intellectuel résultant d'un tel jugement. Toutefois, si l'on prend *à ce point* au sérieux la réalité de la description en termes de vecteurs d'état quantiques, on doit alors aussi être prêt à accepter l'introduction d'un changement (de préférence très subtil) dans les règles mêmes de la théorie quantique. La procédure **U** est en effet, strictement parlant, incompatible avec **R**, et la résolution des contradictions entre les descriptions du comportement aux niveaux quantique et classique exige des réflexions assez délicates.

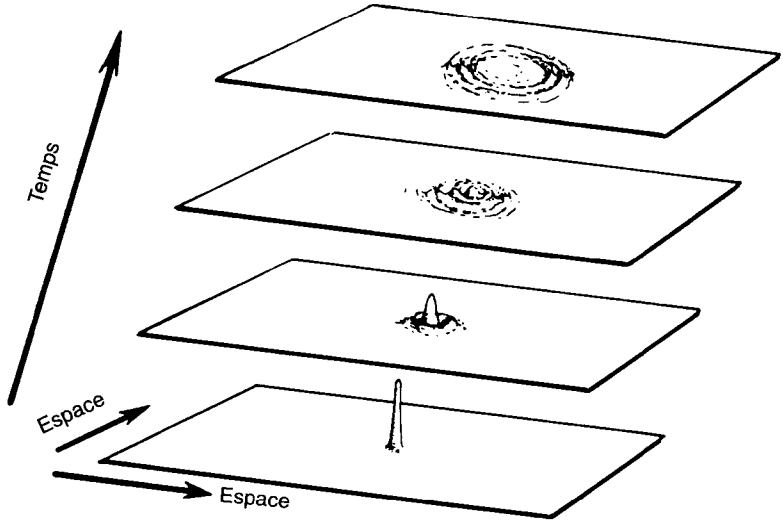


Figure 6.1. La fonction d'onde, initialement fortement localisée en un point, diffuse ensuite dans toutes les directions sous l'action de l'évolution temporelle imposée par l'équation de Schrödinger.

De fait, on a vu apparaître au fil des ans plusieurs tentatives originales pour élaborer une théorie cohérente. Depuis 1966 environ, l'école hongroise, animée à Budapest par Károlyházy, travaille sur une théorie dans laquelle des effets gravitationnels donneraient naissance à un phénomène physique réel équivalent à une procédure **R** (*cf.* aussi Komar 1969). Suivant une voie légèrement différente, Phillip Pearle, du Hamilton College, à Clinton (dans l'État de New York), a proposé vers 1976 une théorie non gravitationnelle dans laquelle **R** intervient pareillement comme phénomène physique réel. Plus récemment, en 1986, Giancarlo Ghirardi, Alberto Rimini et Tullio Weber se sont tournés vers une approche nouvelle et intéressante, puis, après de chaleureux encouragements de la part de John Bell, d'autres chercheurs[9] ont apporté de nombreuses autres suggestions et améliorations.

Mes propres préférences s'inspirant en grande partie de la théorie de Ghirardi, Rimini et Weber (GRW), je vais d'abord la décrire brièvement. L'idée de base est d'accepter la réalité de $|\psi\rangle$ et, dans l'ensemble, la précision des procédures **U** standard. La fonction d'onde d'une particule libre et initialement localisée tend alors, en vertu de l'équation de Schrödinger, à diffuser dans toutes les directions de l'espace à mesure que le temps s'écoule (Fig. 6.1). (Rappelons que la fonction d'onde d'une particule décrit les coefficients de pondération complexes associés aux différentes localisations possibles de cette particule. La figure 6.1 représente schématiquement la partie réelle de ces coefficients.) Ainsi, à mesure que le temps passe, la particule devient de moins en moins localisée. La nouveauté de la théorie GRW consiste à supposer qu'il

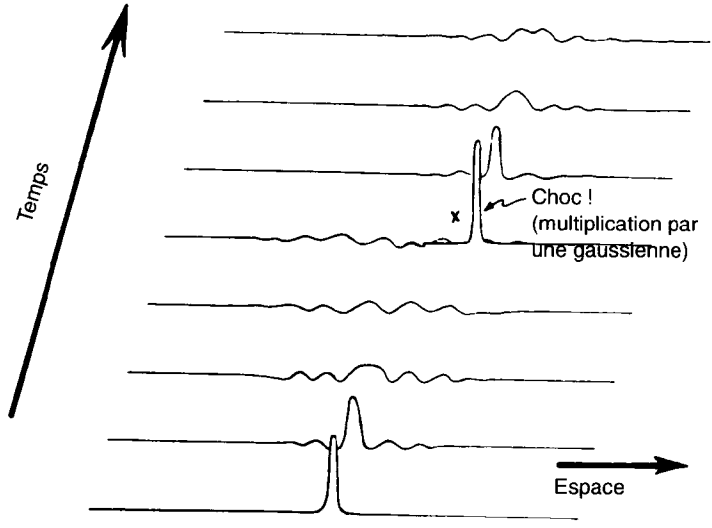


Figure 6.2. Dans la théorie de Ghirardi-Rimini-Weber (GRW), la fonction d'onde évolue normalement en se conformant à l'évolution schrödingerienne **U** ; mais une fois tous les 10^8 ans environ, l'état de chaque particule subit un « choc » qui a pour effet de multiplier la fonction d'onde de la particule par une fonction gaussienne très localisée. C'est la version GRW de **R**.

existe une très faible probabilité pour que cette fonction d'onde se trouve soudainement multipliée par une fonction fortement localisée — une *gaussienne* — dont la largeur est définie par un certain paramètre σ (Fig. 6.2). La fonction d'onde de la particule devient alors instantanément très localisée, et prête à diffuser de nouveau dans toutes les directions. La probabilité de présence de cette gaussienne en un endroit donné étant supposée proportionnelle au carré du module de la valeur de la fonction d'onde en cet endroit, ce schéma est cohérent avec la « règle du module au carré » de la théorie quantique standard.

Cette multiplication est censée intervenir environ une fois tous les cent millions d'années (10^8 ans) ! Désignons cette période par T. Ainsi, en une seconde, la probabilité pour que cette réduction d'état se produise sur une particule est inférieure à 10^{-15} (car il y a environ 3×10^7 secondes dans une année). Pour une seule particule, ce processus n'est donc absolument pas discernable. Supposons cependant que nous ayons un objet relativement gros, dont chacune des particules constitutives est soumise à ce même processus. Si le nombre de ces particules est d'environ 10^{25} (c'est le cas d'une petite souris), la probabilité pour qu'*une* d'entre elles subisse un « choc » de ce type est bien plus considérable que pour une particule isolée, de sorte que les chocs à l'intérieur de l'objet se produisent au rythme de 10^{10} par seconde. Chacun de ces chocs affecte l'état entier de l'objet, car chaque particule « heurtée » est dans un état quantique emmêlé avec le reste de l'objet.

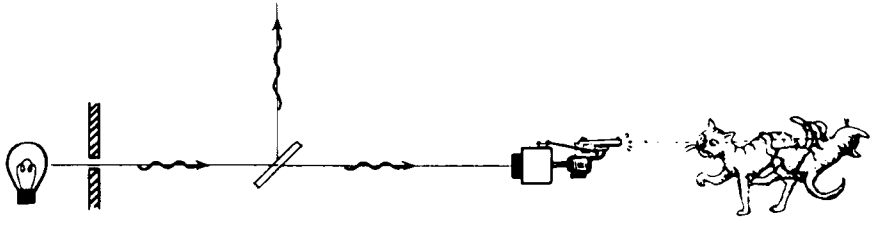


Figure 6.3. *Le chat de Schrödinger.* L'état quantique est une superposition linéaire de photon transmis et de photon réfléchi. La composante transmise déclenche un dispositif qui tue le chat, de sorte que selon l'évolution **U**, le chat existe dans une superposition de vie et de mort. La théorie GRW évite cette contradiction : les fonctions d'onde des particules du chat subissent presque instantanément des chocs gaussiens et le premier d'entre eux a pour effet de localiser l'état du chat, qui devient alors *soit* mort, *soit* vivant.

Voyons comment cette idée s'applique au paradoxe du *chat de Schrödinger*[10], qui symbolise l'énigme-**X** fondamentale de la théorie quantique. Un objet macroscopique, le chat, est placé dans une superposition linéaire quantique de deux états manifestement différents, « chat mort » et « chat vivant » (*cf.* §5.1 et §6.6). D'un point de vue quantique, cette opération ne présente pas de difficulté. Toutefois, la situation résultante n'est pas réellement crédible au niveau du monde *réel*— ainsi que l'a soigneusement démontré Schrödinger (bien que certains tenants de la réalité de la fonction $|\psi\rangle$ aient, nous l'avons vu aux sections 6.2 et 6.8, opté pour d'autres voies : mondes multiples, réduction du vecteur d'état induite par la conscience, etc.). Pour construire un chat de Schrödinger, il suffit d'avoir un événement quantique opérant un changement macroscopique — une *mesure.* Prenons par exemple un photon individuel qui, après avoir été émis par une source de photons, est réfléchi/transmis par un miroir semi-argenté (comme à la section 5.7). Supposons que la partie transmise de la fonction d'onde du photon déclenche un détecteur couplé à un dispositif qui tue le chat, tandis que la partie réfléchie se perd dans la nature et laisse le chat indemne (Fig. 6.3). En reprenant la discussion de la section 6.6 sur les détecteurs, on voit que cette situation engendre un état quantique emmêlé correspondant pour une part à un chat mort et pour l'autre à un chat vivant et un photon libre. Ces deux possibilités se maintiennent *simultanément* dans le vecteur d'état tant qu'aucun processus de réduction (**R**) n'intervient. Ce mystère de la « mesure », répétons-le, est l'énigme-**X** centrale de la théorie quantique.

Dans le schéma GRW , un objet aussi gros qu'un chat, qui contient quelque 10^{27} particules, voit presque instantanément l'une de ses particules « heurtée » par une gaussienne (comme à la figure 6.2), et puisque l'état de cette particule est emmêlé avec ceux des autres particules du chat, la réduction de cette particule « entraîne » les autres avec elle et fait que le chat entier se trouve soit dans l'état vivant, soit dans l'état mort. Ainsi, l'énigme-**X** du chat de Schrödinger — et le problème de la mesure en général — se trouve résolue.

Bien qu'ingénieux, ce schéma a le handicap d'être très *ad hoc*. Rien dans d'autres domaines de la physique n'indique l'existence d'un tel processus, et les valeurs avancées pour T et σ sont simplement choisies de manière à donner des résultats « raisonnables ». (Diósi (1989) a suggéré un schéma semblable, mais dans lequel les paramètres T et σ sont fixés par la constante de gravitation G. Il existe un lien très intime entre ses idées et celles que je vais développer dans un instant.) Une autre difficulté, plus sérieuse celle-là, que soulèvent les schémas de ce type est qu'ils entraînent une (petite) violation du principe de la *conservation de l'énergie*. Ce point aura une importance considérable pour nous à la section 6.12.

6.10 La réduction du vecteur d'état est-elle induite par la gravitation ?

Il y a de bonnes raisons* de penser que la modification de la théorie quantique dont nous avons besoin pour transformer une procédure de type **R** en un processus physique *réel* met effectivement en jeu les effets de la gravitation. Certaines de ces raisons sont liées au fait que la structure même de la théorie quantique standard apparaît tout à fait incompatible avec le concept d'espace courbe de la théorie einsteinienne de la gravitation. Même des concepts tels que l'énergie et le temps — qui jouent un rôle fondamental dans les procédures de la théorie quantique — ne peuvent, dans un contexte gravitationnel totalement général, être définis de manière précise tout en satisfaisant aux exigences ordinaires de la théorie quantique standard. Rappelons également l'« inclinaison » des cônes de lumière (§4.4), phénomène que seule la gravitation a pouvoir de susciter. On pourrait donc penser qu'une modification des principes de base de la théorie quantique est une condition nécessaire à son (éventuelle) unification avec la relativité générale d'Einstein.

Pourtant, la plupart des physiciens semblent manifester une certaine réticence à l'idée que ce soit la théorie *quantique* qui nécessite une modification pour que cette unification soit réussie. Ils affirment au contraire que c'est la théorie d'Einstein que l'on doit modifier, invoquant à juste titre, par exemple, le fait que la relativité générale classique a ses propres problèmes dans la mesure où elle conduit à des *singularités spatio-temporelles* — telles celles associées aux trous noir et au Big Bang — se traduisant par une valeur infinie de la courbure et un évanouissement des notions mêmes d'espace et de temps (*cf.* EOLP, chapitre 7). Personnellement, je ne doute pas que l'on devra modifier la relativité générale pour réaliser son unification avec la théorie quanti-

* Dans EOLP, chapitres 7 et 8, j'ai exposé ces raisons relativement en détail. Je ne les reproduirai donc pas ici. Il me suffira de dire qu'elles sont encore valides — bien que le critère introduit à la section 6.12 diffère de celui donné dans EOLP (p. 391-400).

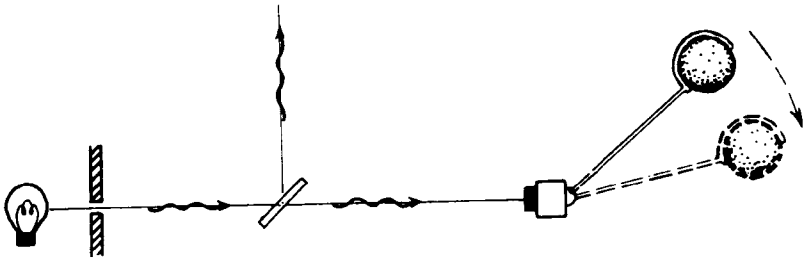


Figure 6.4. Au lieu d'un chat, la mesure pourrait faire intervenir le mouvement d'une sphère matérielle. Quelle doit être la masse minimale de cette sphère, ou de quelle distance minimale doit-elle être déplacée, pour que **R** intervienne ?

que. Cette modification nous sera d'ailleurs précieuse pour comprendre ce qui se passe *réellement* dans les régions qu'on appelle aujourd'hui « singularités ». Mais cela ne signifie pas pour autant qu'il soit inutile de modifier la théorie quantique. Nous l'avons vu à la section 4.5, la relativité générale est une théorie extrêmement précise — non moins précise que la théorie quantique. Il est fort probable que la plupart des idées physiques qui sous-tendent la théorie d'Einstein, tout comme celles qui sous-tendent la théorie quantique, survivront à la fusion de ces deux grandes théories.

Nombre de ceux qui sont prêts à partager ce point de vue objectent néanmoins que les échelles sur lesquelles l'action de *toute* forme de gravitation quantique pourrait s'avérer significative seront totalement inadéquates pour résoudre le problème de la mesure quantique. Ils invoquent la longueur caractéristique de la gravitation quantique, à savoir la *longueur de Planck*, dont la valeur, 10^{-35} m, est de quelque 20 ordres de grandeur inférieure au diamètre d'une simple particule ; ils contestent en outre vigoureusement que la physique opérant sur des distances aussi infimes puisse avoir un rapport avec le problème de la mesure qui, en définitive, concerne des phénomènes qui se situent (à tout le moins) à la frontière du domaine macroscopique. Ces objections reposent sur une vision erronée de l'applicabilité de la gravitation quantique : 10^{-35} m *est* en fait une longueur pertinente pour le problème de la mesure, mais pas dans le sens qui vient spontanément à l'esprit.

Considérons une situation, relativement semblable à celle du chat de Schrödinger, donnant naissance à un état de superposition de deux possibilités macroscopiquement discernables. Considérons par exemple un photon (Fig. 6.4) qui, après avoir heurté un miroir semi-argenté, a pour état une superposition linéaire d'une partie réfléchie et d'une partie transmise. La partie transmise de la fonction d'onde du photon active (ou activerait) un dispositif qui déplace un objet sphérique macroscopique (et non plus un chat) d'un endroit à un autre. Tant que l'évolution du système est régie par la procédure schrödingerienne **U**, la « position » de l'objet sphérique est une superposition quantique de « position initiale » et de « position finale ». Si **R** entre en jeu en tant que processus physique réel, l'objet « saute » sur l'une ou sur l'autre de ces

deux positions — et cela constitue une « mesure » réelle. Comme dans la théorie GRW, l'idée ici est que cette « mesure » est en fait un processus physique totalement objectif, qui intervient chaque fois que la masse de l'objet ou son déplacement sont suffisamment importants. (En particulier, ce processus serait indépendant de la présence ou de l'absence d'un être conscient qui « percevrait » réellement le mouvement — ou quoi que ce soit d'autre — de l'objet.) Je suppose ici que le *dispositif* qui détecte le photon et déplace l'objet est lui-même suffisamment petit pour admettre un traitement purement quantique et que c'est uniquement l'objet qui signale la mesure. Par exemple, on pourrait imaginer un cas extrême où l'objet serait en équilibre suffisamment instable pour que le simple impact du photon suffise à le déplacer de manière significative.

Si l'on applique les procédures **U** standard de la théorie quantique, on trouve que l'état du photon après sa rencontre avec le miroir se compose de deux parties correspondant à deux positions très différentes. L'une de ces parties interagit alors avec le dispositif et finalement avec l'objet sphérique, de sorte que cet objet a pour état quantique une superposition de deux positions tout à fait différentes. Cet objet possédant en outre son propre champ gravitationnel, celui-ci doit donc lui aussi nécessairement figurer dans cette superposition. Ainsi, l'état quantique de l'objet contient une superposition de deux champs gravitationnels différents. En vertu de la théorie d'Einstein, cela signifie que l'on est en présence d'une superposition de deux géométries spatio-temporelles différentes ! La question qui se pose alors est la suivante : existe-t-il un point où ces deux géométries deviennent suffisamment différentes pour obliger à une modification des règles de la théorie quantique et où, plutôt que de forcer ces deux géométries différentes à se superposer, la nature opte pour l'une seulement d'entre elles en opérant *réellement* une procédure de réduction analogue à **R** ?

Le problème est que nous ne savons véritablement pas comment appréhender des superpositions linéaires d'états lorsque ces états mettent en jeu des géométries spatio-temporelles différentes. Une des difficultés fondamentales de la « théorie standard » est que lorsque les géométries deviennent significativement différentes, nous n'avons aucun moyen systématique d'identifier un point de la première géométrie avec un point de la seconde — ces deux géométries étant des espaces strictement *distincts* —, de sorte que l'idée même que l'on puisse former une superposition des états *matériels* au sein de ces deux espaces distincts est profondément obscure.

Nous devons à ce stade nous demander quand il faut considérer que deux géométries sont « significativement différentes » l'une de l'autre. C'est ici, en fait, que la longueur de Planck (10^{-33} cm) entre en scène. Schématiquement parlant, l'idée est que la réduction se produit uniquement si la différence entre ces géométries, mesurée selon une échelle appropriée, est d'environ 10^{-33} cm ou plus. On pourrait par exemple imaginer (Fig. 6.5) que ces deux géométries ont tendance à coïncider, mais que lorsque, sur cette échelle idoine, la différence devient trop importante, la réduction **R** se produit — de sorte que, plutôt que de maintenir la superposition régie par **U**, la nature choisit une géométrie au détriment de l'autre.

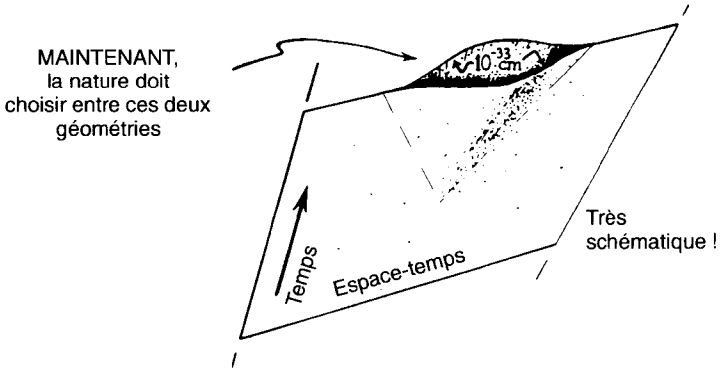


Figure 6.5. Quel lien y a-t-il entre l'échelle de Planck (10^{-33} cm) et la réduction de l'état quantique ? Très schématiquement, la réduction intervient lorsque le déplacement de masse associé aux deux états en superposition est tel que les deux espaces-temps résultants diffèrent d'environ 10^{-33} cm.

À quelle échelle de masse ou de longueur correspond un changement aussi infime de géométrie ? En fait, à cause de la faiblesse des effets gravitationnels, cette échelle s'avère tout à fait conséquente et l'on peut très bien s'en servir de ligne de démarcation entre les niveaux classique et quantique. Afin de préciser tout cela, je vais d'abord dire quelques mots du *système d'unités absolues* (ou *unités de Planck*).

6.11 Les unités absolues

L'idée (initialement* due à Max Planck (1906) et reprise notamment par John A. Wheeler (1975)) consiste à utiliser les trois constantes naturelles les plus fondamentales, à savoir la vitesse de la lumière c, la constante de Planck (divisée par 2π) $\hbar$ et la constante de la gravitation de Newton G, comme unités pour convertir toutes les mesures physiques en nombres purs (sans dimension). Cela revient à adopter des unités de longueur, de masse et de temps telles que ces trois constantes soient toutes égales à l'unité :

$$c = 1, \hbar = 1, G = 1.$$

* Une idée très semblable avait été émise 25 ans auparavant par le physicien irlandais George Johnstone Stoney (1881). Celui-ci avait pris, à la place de la constante de Planck (inconnue à cette époque), la charge de l'électron. (Je remercie John Barrow de m'avoir communiqué cette précision.)

La longueur de Planck, 10^{-33} cm ou 10^{-35} m, qui, en unités ordinaires, est donnée par $(G\hbar/c^3)^{1/2}$, prend alors la valeur 1, de sorte qu'*elle* est l'unité absolue de *longueur*. L'unité absolue de *temps*, qui correspond au temps mis par la lumière pour parcourir la longueur de Planck, est le *temps de Planck* ($= (G\hbar/c^5)^{1/2}$), soit environ 10^{-43} seconde. Il existe également une unité absolue de *masse*, appelée *masse de Planck* ($= (\hbar c/G)^{1/2}$) ; elle vaut environ 2×10^{-5} gramme, ce qui représente une masse très importante à l'échelle des phénomènes quantiques, mais relativement petite à l'échelle ordinaire — à peu près la masse d'une puce.

Ce ne sont manifestement pas là des unités très pratiques, sauf éventuellement la masse de Planck, mais elles s'avèrent très utiles lorsqu'on considère des effets pouvant être associés à la gravitation quantique. Voici les valeurs (très arrondies) de certaines grandeurs physiques importantes, une fois exprimées en unités absolues :

$$\begin{aligned}
\text{seconde} &= 1{,}9 \times 10^{43} \\
\text{jour} &= 1{,}6 \times 10^{48} \\
\text{année} &= 5{,}9 \times 10^{50} \\
\text{mètre} &= 6{,}3 \times 10^{34} \\
\text{centimètre} &= 6{,}3 \times 10^{32} \\
\text{micron} &= 6{,}3 \times 10^{28} \\
\text{fermi (« taille de l'interaction forte »)} &= 6{,}3 \times 10^{19} \\
\text{masse d'un nucléon} &= 7{,}8 \times 10^{-20} \\
\text{gramme} &= 4{,}7 \times 10^{4} \\
\text{erg} &= 5{,}2 \times 10^{-17} \\
\text{degré Kelvin} &= 4 \times 10^{-33} \\
\text{densité de l'eau} &= 1{,}9 \times 10^{-94}
\end{aligned}$$

6.12 Le nouveau critère

Je vais maintenant donner un nouveau critère[11] d'apparition d'une réduction du vecteur d'état induite par la gravitation. Ce critère diffère substantiellement de celui proposé dans EOLP et se rapproche de certaines idées récentes dues notamment à Diósi. Les *raisons* qui, dans EOLP, incitaient à soupçonner un lien entre la procédure **R** et la gravitation restent toujours valides, mais mon nouveau critère repose sur des idées théoriques différentes. Il évite en outre certains des problèmes conceptuels soulevés par son prédécesseur et est d'un emploi plus simple. Le critère proposé dans EOLP définissait le stade où l'on pouvait considérer que deux états étaient trop différents (par rapport à leurs champs gravitationnels respectifs — *i.e.* leurs espaces-temps respectifs) pour pouvoir coexister dans une superposition linéaire quantique, et donc où intervenait la procédure **R**. L'idée que j'expose ici est légèrement différente. Elle consiste non pas à chercher une mesure absolue de la différence gravitationnelle au-delà de laquelle les états sont trop différents pour se maintenir en superposition, mais à considérer que des états superposés qui sont très différents sont *instables* — un peu, par exemple, comme un noyau d'uranium instable — et à associer à cette différence une *vitesse* de réduction du vecteur d'état. Plus cette différence est grande, plus la vitesse de réduction est importante.

Par souci de clarté, je vais dans un premier temps appliquer ce nouveau critère au cas particulier décrit à la section 6.10 (mais on peut facilement le généraliser à de nombreuses autres situations). Évaluons l'*énergie* nécessaire pour séparer deux exemplaires de la sphère en tenant compte uniquement des effets *gravitationnels*. Pour cela, supposons que nous ayons initialement deux sphères qui coïncident et s'interpénètrent l'une l'autre (Fig. 6.6), puis que nous les éloignions l'une de l'autre, lentement, en diminuant ainsi progressivement leur degré d'interpénétration, jusqu'à ce que toutes deux atteignent la distance qui les sépare dans la superposition d'états considérée. L'inverse de l'énergie gravitationnelle — mesurée en unités absolues — dépensée au cours de cette opération donne approximativement le temps — mesuré lui aussi en unités absolues* — écoulé avant que ne se produise la réduction du vecteur d'état, autrement dit, avant que l'état de superposition de la sphère ne saute spontanément sur l'un ou l'autre des états localisés.

Pour une sphère de masse m et de rayon a, on obtient une énergie de l'ordre de m^2/a. En fait, la valeur réelle de cette énergie dépend de la distance dont les deux sphères ont été éloignées, mais cette distance n'est pas vraiment déter-

---

* Peut-être préférera-t-on exprimer ce temps de réduction dans des unités plus ordinaires que les unités absolues adoptées ici. En fait, ce temps de réduction a simplement pour expression $\hbar/E$, où E est l'énergie de séparation gravitationnelle mentionnée et où, on le voit, la seule constante absolue présente est $\hbar$. Le fait que la vitesse de la lumière c soit absente de cette expression suggère qu'une théorie « newtonienne » de ce type mériterait d'être étudiée (*cf.* Christian 1994).

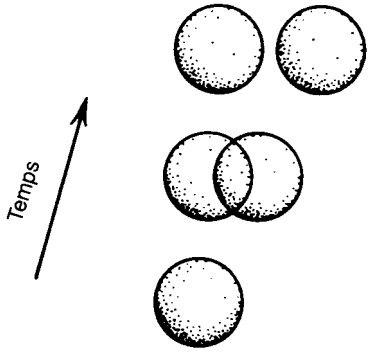


Figure 6.6. Pour déterminer le temps de réduction $\hbar / E$, imaginez que l'on sépare deux exemplaires de la sphère. E est alors l'énergie associée à ce mouvement. Elle est calculée en ne tenant compte que de l'attraction gravitationnelle des deux sphères.

minante si les deux sphères ne se chevauchent pas (trop) lorsqu'elles atteignent leurs positions finales. L'énergie supplémentaire qui serait nécessaire pour les séparer de la position de contact, même pour les éloigner d'une distance infinie, est du même ordre que celle mise en jeu pour passer de la coïncidence à la position de contact (elle vaut $\frac{5}{7}$ fois autant). Ainsi, tant qu'on s'en tient aux ordres de grandeur, on peut ignorer la contribution due au déplacement des sphères une fois dépassée la position de contact, à condition que les sphères soient en fait (clairement) dissociées. Le temps de réduction — mesuré en unités absolues — est donc de l'ordre de

$$\frac{a}{m^2}$$

soit, très grossièrement,

$$\frac{1}{20\rho^2 a^5},$$

où ρ est la densité de la sphère. Pour un objet de densité ordinaire (par exemple, une gouttelette d'eau), cela donne environ $10^{186}/a^5$.

On constate avec satisfaction que ce critère donne des réponses très « raisonnables » dans certaines situations simples. Par exemple, dans le cas d'un nucléon (un neutron ou un proton), pour lequel a est la « taille de l'interaction forte », soit 10^{-13} cm (environ 10^{20} en unités absolues), et m vaut environ 10^{19}, on obtient un temps de réduction d'environ 10^{58}, ce qui est supérieur à 10 millions d'années. Une valeur aussi énorme est rassurante, car les neutrons individuels engendrent des effets d'interférence quantique directement observables[12]. Si nous avions obtenu un temps de réduction très bref, nous aurions été en contradiction avec ces observations.

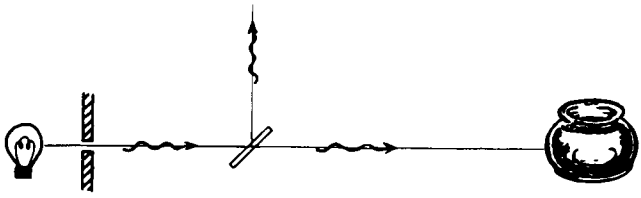


Figure 6.7. On peut imaginer qu'au lieu d'avoir une masse en mouvement, la partie transmise de l'état du photon est simplement absorbée par une masse de fluide.

Si l'on considère un corps plus « macroscopique », par exemple une infime poussière d'eau de 10^{-5} cm de diamètre, on obtient un temps de réduction qui se mesure en heures ; si cette poussière a un rayon de 10^{-4} cm (1 micron), son temps de réduction vaut environ un vingtième de seconde ; si elle a un rayon de 10^{-3} cm, son temps de réduction est inférieur à un millionième de seconde. D'une manière générale, lorsqu'on considère un objet se trouvant dans une superposition de deux états spatialement distants, on calcule simplement l'énergie nécessaire pour les éloigner de cette distance en tenant compte uniquement de l'interaction gravitationnelle entre ces deux objets. L'inverse de cette énergie mesure une sorte de « demi-vie » de l'état superposé. Plus cette énergie est grande, plus le temps durant lequel cet état de superposition se maintient est bref.

Dans une situation expérimentale réelle, il serait extrêmement difficile de maintenir une superposition d'états à l'abri des perturbations de — et des emmêlements avec — la matière composant l'environnement. Il faudrait alors tenir également compte des effets gravitationnels associés à cet environnement. Ces effets gravitationnels joueraient un rôle non négligeable même si la perturbation ne provoquait dans l'environnement aucun mouvement de masse significatif à l'échelle macroscopique. Les mouvements, même infimes, des particules individuelles pourraient s'avérer importants — bien qu'il en serait normalement ainsi pour des déplacements de masses totales plus élevées que dans le cas de « poussière » macroscopique.

Afin de clarifier l'effet d'une telle perturbation, remplaçons le dispositif qui, dans la situation idéalisée de la figure 6.4, déplaçait les sphères par une masse fluide dont la seule action est d'*absorber* le photon dans le cas où il est transmis par le miroir (Fig. 6.7). Cette masse fluide joue donc ici le rôle de l'« environnement ». Au lieu de considérer une superposition linéaire de deux états macroscopiquement distincts — rappelons-le, un exemplaire de la masse sphérique avait subi un déplacement par rapport à l'autre —, nous nous intéressons uniquement ici à la différence survenant entre les deux configurations de positions atomiques lorsqu'une de ces configurations subit un déplacement aléatoire par rapport à l'autre. Pour une masse fluide ordinaire de rayon a, on obtient un temps de réduction égal, non plus à $10^{186}/a^5$ comme dans le cas des deux masses en déplacement relatif, mais à $10^{130}/a^3$ — cette dernière relation dépendant dans une certaine mesure des hypothèses adoptées. Cela

indique que pour que la réduction se produise, il faut une masse de fluide plus grande que dans le cas du déplacement des deux sphères. Mais la réduction se produit néanmoins, bien qu'il n'y ait aucun mouvement macroscopique d'ensemble.

Revenons sur l'obstacle matériel qui interceptait le faisceau de photons lors de la discussion des interférences quantiques de la section 5.8. La simple *absorption* — ou absorption potentielle — d'un photon par cet obstacle suffisait pour déclencher **R**, en dépit du fait qu'il ne se produisait aucun événement macroscopiquement observable. Cela montre également comment une perturbation suffisamment importante d'un environnement *emmêlé* avec le système considéré déclenche par elle-même la procédure **R** — ce qui nous ramène aux approches plus conventionnelles EP.

En fait, la quasi-totalité des processus de mesure concrets devrait perturber un grand nombre de particules de l'environnement. En vertu des idées avancées ici, *cette* perturbation serait souvent l'effet dominant, et non le mouvement macroscopique d'objets, comme dans le cas du « déplacement de masse » considéré au début de cette section. En l'absence d'une maîtrise rigoureuse de la situation expérimentale, tout mouvement macroscopique d'un objet de taille appréciable perturberait une grande partie de l'environnement, et il est probable que le temps de réduction *de cet environnement* — qui se comporterait comme $10^{130}/b^3$, où b est le rayon de la région de l'environnement (de densité égale à celle de l'eau) emmêlée — l'emporterait sur (*i.e.* serait bien plus bref que) le temps de $10^{186}/a^5$ qui serait associé à l'objet lui-même. Par exemple, pour un rayon b de seulement un dixième de millimètre, la réduction devrait survenir — et du fait de cette seule perturbation — en l'espace d'un millionième de seconde.

Si un tel schéma a beaucoup en commun avec la description conventionnelle que j'ai examinée à la section 6.6, il fournit toutefois une condition précise pour l'intervention *réelle* de **R** dans cet environnement. N'oublions pas les objections opposées, dans la section 6.6, à l'approche conventionnelle EP en tant qu'elle prétend être une description de la réalité physique. Avec le critère proposé ici, ces objections disparaissent, car une fois l'environnement suffisamment perturbé, la réduction y intervient *réellement* et rapidement — et s'accompagne d'une réduction dans tout « appareil de mesure » emmêlé avec cet environnement. Rien ne peut inverser cette réduction et permettre à l'état initial de se reformer, même en imaginant d'énormes progrès technologiques. Il n'y a donc aucune contradiction dans le fait que l'appareil de mesure affiche réellement *soit* **OUI**, *soit* **NON** — ainsi qu'il le ferait dans la présente interprétation.

J'imagine qu'une telle description devrait s'appliquer à nombre de processus biologiques et expliquer de manière plausible la raison pour laquelle les structures biologiques de diamètre très inférieur au micron ont souvent un comportement manifestement classique. Un système biologique extrêmement emmêlé avec son environnement — à l'instar de ce que nous venons d'examiner — verrait son *propre* état constamment réduit du fait de la réduction permanente de cet *environnement.* On peut imaginer, inversement, que dans

certaines circonstances, il soit plus propice à un système biologique de demeurer longtemps, pour une raison ou une autre, dans un état *non* réduit. Cela surviendrait uniquement si ce système était, dans un certain sens, très efficacement isolé de son environnement. Ces considérations s'avéreront importantes pour nous (§7.5).

Soulignons que l'énergie qui définit la durée de vie de l'état superposé est une *différence* d'énergie, et non l'énergie *totale* (la masse-énergie) mise en jeu dans l'ensemble de la situation. Ainsi, pour un corps de taille conséquente mais à l'amplitude de mouvement limitée — imaginons que ce corps soit un cristal, de sorte que ses atomes ne subissent pas de mouvements aléatoires —, les superpositions quantiques se maintiendraient longtemps. Ce corps pourrait être bien plus gros que les gouttelettes d'eau considérées plus haut. Son voisinage pourrait contenir d'autres masses bien plus grosses, à condition qu'elles ne soient pas trop emmêlées avec son état superposé. (Ces considérations pourraient se révéler importantes pour les dispositifs à circuits intégrés, tels les détecteurs d'ondes gravitationnelles, qui utilisent des oscillations cohérentes de corps solides[13] — voire cristallins.)

Jusqu'ici, les ordres de grandeur semblent tout à fait plausibles, mais il faut bien sûr voir si cette idée survivra à un examen plus rigoureux. Un test crucial consisterait à trouver des situations expérimentales pour lesquelles la théorie standard prédirait des effets dépendant de superpositions quantiques à grande échelle, mais à un niveau où le présent critère affirme que ces superpositions ne peuvent se maintenir. Si, dans de telles situations, les prédictions quantiques conventionnelles étaient confortées par l'observation, il faudrait alors abandonner les idées que j'avance — ou, du moins, les modifier radicalement. Le cas contraire tendrait à les confirmer. Malheureusement, je n'ai pour l'instant aucune idée d'expérience concrète correspondant à ce genre de situation. Les supraconducteurs, ainsi que les dispositifs appelés SQUID (qui dépendent des superpositions quantiques à grande échelle apparaissant dans les supraconducteurs) semblent offrir à cet égard des possibilités d'expériences prometteuses (voir Leggett 1984). Mais les idées que je propose exigent d'être affinées avant de pouvoir être appliquées directement à ces situations. Dans les supraconducteurs, les différents états superposés correspondent à de très faibles déplacements de masse. Il y a en revanche un important transfert de *quantité de mouvement*, et mon critère doit être amélioré pour prendre en compte ce transfert.

Ce critère doit en outre être légèrement reformulé pour traiter ne serait-ce que le simple contexte d'une chambre à brouillard — dans laquelle la présence d'une particule se signale par la condensation de petites gouttelettes dans la vapeur environnante. Supposez que l'état quantique d'une particule chargée soit une superposition linéaire de deux états, correspondant à l'état de la particule localisée respectivement à l'intérieur et à l'extérieur de la chambre à brouillard. Seule la partie du vecteur d'état associée à la localisation de la particule à l'intérieur de la chambre déclenche la formation d'une gouttelette, de sorte que l'état de la particule est une superposition de deux états macroscopiquement différents. L'un correspond à une gouttelette de vapeur, l'autre à une

vapeur uniforme. Le problème est donc d'évaluer l'énergie gravitationnelle mise en jeu lors de la séparation des molécules de vapeur associées aux deux états composant la superposition. Cette évaluation se trouve compliquée par le fait qu'apparaît en outre ici une différence entre les *self*-énergies gravitationnelles de la gouttelette et de la vapeur non condensée, et probablement faut-il formuler *différemment* le critère donné plus haut pour en tenir compte. Considérons donc la *self-énergie gravitationnelle* de cette distribution de masse, à savoir la *différence* entre les distributions de masse des deux états participant à la superposition linéaire quantique. L'inverse de cette self-énergie donne une autre valeur du temps de réduction (Penrose 1994*b*). Pour les situations considérées jusqu'ici, cette nouvelle formulation redonne les mêmes résultats que précédemment, mais indique un temps de réduction relativement différent (plus rapide) dans le cas d'une chambre à brouillard. En fait, il existe diverses méthodes de détermination du temps de réduction ; elles donnent des réponses différentes dans certaines situations, mais concordent pour la superposition élémentaire de deux états associée au déplacement d'une masse rigide considéré au début de cette section. La méthode originale est celle de Diósi (1989) (elle a rencontré certaines difficultés, comme l'ont souligné Ghirardi, Grassi et Rimini (1990) ; ceux-ci ont également suggéré un remède à ces difficultés). Je ne distinguerai pas ici entre ces diverses méthodes et dans les chapitres qui suivent, je les désignerai globalement par une expression du type « théorie de la section 6.12 ».

Quelles raisons incitent à faire intervenir expressément le « temps de réduction » ? Mes motivations initiales (Penrose 1993*a*) sont un peu trop techniques pour être décrites ici et se sont, en définitive, révélées peu convaincantes et incomplètes[14]. J'exposerai dans un instant un autre argument en faveur de la pertinence du temps de réduction. Bien qu'il soit lui aussi incomplet, il semble cependant indiquer avec force une cohérence sous-jacente dont l'existence donne une raison supplémentaire de croire que la réduction d'état est bien en définitive un phénomène gravitationnel du type que je propose ici.

Nous avons évoqué à la section 6.9 le problème de la *conservation de l'énergie* dans les théories de type GRW. Les « chocs » subis par les particules (lorsque leurs fonctions d'onde sont spontanément multipliées par des gaussiennes) engendrent de petites violations de la conservation de l'énergie. En outre, ce type de processus semble impliquer un transfert d'énergie non local. Cette violation apparaît comme une caractéristique — inévitable semble-t-il — des théories qui considèrent la procédure **R** comme un effet physique *réel*. À mon sens, cela constitue un solide témoignage supplémentaire en faveur des théories dans lesquelles les effets *gravitationnels* participent de manière cruciale au processus de réduction, car en relativité générale, la conservation de l'énergie est un problème subtil et difficile à cerner. Le champ gravitationnel lui-même contient de l'énergie, et cette énergie contribue de manière appréciable à l'énergie totale (et donc à la masse, en vertu de la relation d'Einstein $E = mc^2$) d'un système. Pourtant, c'est une énergie diffuse, qui occupe l'espace vide d'une manière mystérieuse et non locale[15]. Souvenons-nous, notamment, de la masse-énergie émise, sous forme d'ondes gravitationnelles,

par le pulsar binaire PSR 1913 + 16 (*cf.* §4.5) ; ces ondes créent des « rides » dans la structure même de l'espace vide. L'énergie contenue dans les champs mutuellement attractifs de ces deux étoiles à neutrons joue également un rôle important dans leur dynamique, rôle que l'on ne peut ignorer. Mais ce type d'énergie, situé dans l'espace vide, est particulièrement insaisissable. Il ne peut s'obtenir par « addition » de contributions locales de la densité d'énergie ; et l'on ne peut même pas le localiser dans une région particulière de l'espace-temps (*cf.* EOLP, p. 239). Il est alors tentant de relier le problème tout aussi délicat de la non-localité de l'énergie de la procédure **R** à celui de la gravitation classique, et de les mettre tous deux en balance pour obtenir une vision globale cohérente.

Les suggestions que j'émets ici permettront-elles d'atteindre cet objectif ? Je crois qu'elles ont de bonnes chances d'y parvenir — même si l'on ne dispose pas encore du cadre théorique adéquat. Certains indices montrent d'ailleurs que l'on peut être optimiste car, ainsi que je l'ai mentionné plus haut, le processus de réduction a des traits analogues à celui de la désintégration d'un atome ou d'un noyau instables. De même qu'au terme d'une « demi-vie » caractéristique, un noyau instable se désintègre en quelque chose de plus stable, de même une superposition de localisations de masse forme un état quantique instable qui, après une durée de vie caractéristique (donnée, en gros, par l'inverse de l'énergie gravitationnelle de la distance), se désintègre en un état où la masse est soit en un endroit, soit en un autre — ce qui équivaut à deux modes de désintégration possibles.

Pour les particules ou les noyaux instables, la durée de vie (plus exactement, la demi-vie) du processus de désintégration est égale à l'inverse d'une petite *incertitude* sur la masse-énergie de la particule initiale — c'est une conséquence du principe d'indétermination de Heisenberg. (Par exemple, la masse d'un noyau de polonium 210 (noyau instable qui se désintègre en émettant une particule α pour se transformer en un noyau de plomb) est connue avec une incertitude égale à l'inverse du temps de désintégration — en l'occurrence, environ 138 jours, ce qui donne une incertitude très faible, de l'ordre de 10^{-34} fois la masse du noyau de polonium ! Pour des particules individuelles instables, le rapport de cette incertitude à la masse est toutefois bien plus grand.) Ainsi, la « désintégration » associée au processus de réduction devrait *également* mettre en jeu une incertitude fondamentale sur l'énergie de l'état initial. Selon le schéma proposé ici, cette incertitude est essentiellement celle de la self-énergie gravitationnelle de l'état superposé. Cette self-énergie gravitationnelle est associée à l'insaisissable énergie de champ non locale qui cause tant de soucis en relativité générale et ne se réduit pas à l'addition de contributions de densités d'énergie locales. Elle est également associée à l'incertitude fondamentale liée à l'identification des points de deux géométries spatio-temporelles différentes en état de superposition (*cf.* §6.10). Si l'on considère réellement que cette contribution gravitationnelle constitue une « incertitude » inhérente à l'énergie de l'état superposé, on obtient un accord avec la durée de vie de cet état telle qu'elle est donnée par le schéma proposé ici. Ainsi, ce schéma paraît lier de manière très cohérente les deux problèmes d'énergie et il

semble bien qu'il pourrait permettre de parvenir à une théorie totalement cohérente.

Deux questions particulièrement importantes se posent maintenant à nous. Premièrement, quel est l'intérêt de ces considérations pour le fonctionnement du *cerveau* ? Et deuxièmement, a-t-on des raisons de penser, d'un point de vue purement physique, que ce processus de réduction induit par la gravitation pourrait se caractériser par une (forme adéquate de) *non-calculabilité* ? Le prochain chapitre va montrer qu'il existe effectivement quelques possibilités surprenantes.

7
Théorie quantique et fonctionnement du cerveau

7.1 Le fonctionnement du cerveau met-il en jeu une action quantique macroscopique ?

Selon le point de vue conventionnel, c'est avant tout la physique classique qui devrait nous donner la clé de l'activité cérébrale. Les signaux neuronaux sont perçus comme des phénomènes fonctionnant sur le principe du « tout ou rien » — à l'instar du courant qui, dans les circuits d'un ordinateur électronique, est *soit* présent, *soit* absent —, sans faire intervenir les mystérieuses *superpositions* d'états caractéristiques des actions quantiques. S'ils reconnaissent qu'aux niveaux *sous-jacents*, certains effets quantiques jouent probablement un rôle, les biologistes semblent généralement penser que rien, dans l'analyse des conséquences macroscopiques de ces éléments quantiques, n'oblige véritablement à sortir du cadre classique. Certes les forces chimiques qui contrôlent les interactions atomiques et moléculaires sont d'origine quantique, et ce sont bien des actions chimiques qui, pour une large part, gouvernent le comportement des *neurotransmetteurs* propageant les signaux entre neurones — à travers d'infimes intervalles appelés *espaces synaptiques* ; certes les potentiels d'action qui régissent la transmission proprement dite des messages neuronaux ont eux aussi une origine quantique. Pourtant, il semble communément admis qu'un modèle purement classique devrait suffire pour simuler le comportement des neurones et leurs relations mutuelles, et de même qu'un système *classique*, dans lequel les caractéristiques mystérieuses et plus subtiles de la physique quantique n'interviennent pas de façon significative, devrait suffire pour simuler le fonctionnement physique de l'ensemble du cerveau.

Un tel point de vue implique que toute activité cérébrale est censée *soit* « survenir », *soit* « ne pas survenir ». Les étranges *superpositions* quantiques qui autorisent une « occurrence » *et* une « non-occurrence » simultanées — avec des coefficients de pondération complexes — n'auraient donc ici aucun rôle significatif. Si l'on reconnaît que de telles superpositions quantiques interviennent « réellement » à un certain niveau d'activité submicroscopique, on considère toutefois que les effets d'interférence caractéristiques des phénomènes quantiques ne jouent aucun rôle aux échelles plus importantes. Ainsi, il conviendrait de traiter toutes ces superpositions comme de simples mélanges statistiques, et la modélisation de l'activité cérébrale en termes classiques serait parfaitement satisfaisante « EP ».

Cette opinion n'est cependant pas unanimement partagée. En particulier, le célèbre neurophysiologiste John Eccles souligne l'importance d'effets quantiques dans l'activité synaptique (voir notamment Beck et Eccles 1992, Eccles 1994) et affirme que la grille vésiculaire présynaptique — un réseau hexagonal paracristallin situé dans les cellules pyramidales du cerveau — serait un site quantique. De même, certaines personnes (dont je suis ; *cf.* EOLP, p. 435, et Penrose 1987) pensent que si les cellules rétiniennes (qui font techniquement partie du cerveau) peuvent réagir à un petit nombre de photons (Hecht *et al.* 1941) — jusqu'à être sensibles, dans des circonstances appropriées, à *un seul* photon (Baylor *et al.* 1979) —, il se pourrait que le cerveau contienne lui aussi des neurones qui seraient pareillement des dispositifs de détection essentiellement quantiques.

Si le cerveau est effectivement le siège d'effets macroscopiques déclenchés par des processus quantiques, certaines personnes pensent que, grâce au phénomène d'*indétermination quantique*, l'*esprit* est en mesure d'influencer le cerveau. Ces personnes adoptent probablement ici un point de vue *dualiste*, soit explicitement, soit implicitement : le « libre arbitre » d'un « esprit extérieur » influencerait les choix quantiques résultant de processus non déterministes. Ce serait au moyen de la procédure quantique **R** que l'« entité pensante », selon les dualistes, influencerait le comportement du cerveau.

Je ne sais que penser de telles suggestions, notamment parce qu'en théorie quantique standard, l'indétermination quantique *ne se situe pas* au niveau quantique dans la mesure où celui-ci est toujours régi par l'évolution déterministe **U**. C'est seulement lors de l'amplification du niveau quantique au niveau classique que l'indétermination de **R** semble se manifester. Du point de vue standard EP, cette indétermination ne « survient » que lorsque des domaines suffisamment vastes de l'environnement interagissent avec l'événement quantique. En fait — nous l'avons vu à la section 6.6 —, on ne sait même pas clairement, du point de vue standard, ce qu'il faut entendre par « survient ». On peut difficilement soutenir que la théorie quantique conventionnelle autorise une indétermination au niveau d'une particule quantique individuelle — un photon, un atome ou une petite molécule. Lorsque (par exemple) la fonction d'onde d'un photon rencontre une cellule photosensible, elle déclenche une succession d'événements qui est déterministe (action de **U**) tant que l'on peut considérer que le système reste « au niveau quantique ». Cette succession

d'événements finit par perturber de vastes régions de l'environnement et c'est alors que l'on considère que, EP, **R** s'est produite. Ce serait donc seulement à ce stade assez flou que l'« entité pensante » influencerait le système.

Selon mon point de vue sur la réduction du vecteur d'état (*cf.* 6.12), pour trouver le niveau où la procédure **R** entre *réellement* en action, il faut examiner les très grandes échelles mises en jeu lorsque des quantités de matière considérables se trouvent dans un état quantique emmêlé. Il s'agit de diamètres se mesurant en microns ou en millimètres, voire peut-être davantage en l'absence de mouvement de masse significatif. (À partir de maintenant, je désignerai cette procédure assez précise mais hypothétique par **RO**, pour *réduction objective**). En tout cas, si, adhérant au point de vue dualiste mentionné plus haut, on cherche l'endroit où un « esprit » extérieur pourrait influencer des processus physiques — probablement en remplaçant le caractère purement aléatoire de la théorie quantique par quelque chose de plus subtil —, on doit alors expliquer comment cette influence pourrait se manifester à une échelle bien plus grande que celle des particules quantiques individuelles, et donc se concentrer sur le point de rencontre entre les niveaux quantique et classique. Comme nous l'avons vu au précédent chapitre, les scientifiques ne s'accordent pas sur l'existence, la nature ni la localisation d'un tel point.

À mon avis, envisager l'existence d'un « esprit » dualiste, qui serait (en toute logique) *extérieur* au corps et qui influencerait les choix apparemment associés à l'action de **R**, ne présente guère d'avantages d'un point de vue scientifique. Si la « volonté » pouvait influencer le choix de la nature face aux alternatives qui se traduit par **R**, pourquoi un expérimentateur ne pourrait-il dès lors, par le « pouvoir de sa volonté », influencer le résultat d'une expérience quantique ? S'il en était ainsi, on constaterait certainement nombre de violations des probabilités quantiques ! Je ne peux croire, pour ma part, qu'une telle image soit proche de la vérité. Poser l'existence d'une « entité pensante » extérieure non soumise aux lois de la physique n'est pas ce que l'on pourrait appeler, même au sens large, une explication scientifique et ressortit au point de vue $\mathscr{D}$ (*cf.* §1.3).

Il est toutefois difficile d'opposer des arguments rationnels à un tel point de vue car, de par sa nature même, sa formulation n'obéit pas aux règles précises qui permettraient de le soumettre à une analyse scientifique. Aux lecteurs qui,

---

* Dans EOLP, j'avais désigné ce genre de chose par l'expression « gravité quantique correcte » (GQC). Cette nouvelle désignation veut marquer une idée légèrement différente : moins que le lien unissant cette procédure au profond problème de la formulation d'une théorie pleinement cohérente de la gravitation quantique, elle vise surtout à souligner que nous sommes à la recherche d'une procédure, inspirée des suggestions exposées à la section 6.12, qui contiendrait un élément fondamentalement irréductible au calcul et encore absent de la théorie physique.†

† Mentionnons à part la dernière phrase de cette note : Penrose y invoque un autre avantage de cette nouvelle désignation, qui ne peut être rendu en français : elle rappelle qu'appliquée à une superposition d'états, une réduction objective [*objective reduction (**OR**)* en anglais] donne un résultat physique qui est effectivement une chose *ou* [***OR*** en anglais] une autre [*one thing* or *the other*]. (N.d.T.)

pour quelque raison que ce soit, gardent la conviction (point de vue $\mathscr{D}$) que la science demeurera à jamais incompétente pour aborder le problème de l'esprit, je demande encore un peu de patience. Nous allons explorer les perspectives qu'offrirait une science dont on pourrait faire reculer les limites étroites qui sont aujourd'hui les siennes. Si l'« esprit » est une entité totalement extérieure au corps humain, il est difficile de comprendre pourquoi tant de ses attributs sont si intimement liés aux propriétés du cerveau physique. Pour aborder le problème de l'esprit, nous devons à mon sens mener des recherches plus approfondies sur les structures « matérielles », physiques, qui constituent le cerveau — et nous demander ce qu'*est* une structure « matérielle » au niveau quantique ! Il n'y a selon moi aucune autre issue que de s'interroger plus profondément sur ces vérités qui fondent l'existence même de la nature.

Quoi qu'il en soit, une chose au moins semble claire. Nous devons examiner non seulement le comportement quantique des particules et des atomes individuels — voire des petites molécules —, mais aussi celui de systèmes qui conservent leur nature manifestement quantique à une échelle bien plus grande. Si ces systèmes ne présentent aucune cohérence quantique à grande échelle, il n'y a aucun espoir que l'un quelconque des effets quantiques subtils — tels la non-localité, le parallélisme quantique (plusieurs actions superposées se déroulant simultanément) — ou les effets de la contrafactualité aient un impact significatif une fois atteint le niveau classique de l'activité cérébrale. Sans un « écran » protégeant convenablement l'état quantique de son environnement, ces effets se trouvent immédiatement noyés dans le désordre inhérent à cet environnement — en l'occurrence, dans les mouvements aléatoires des substances et fluides biologiques constituant l'ensemble du cerveau.

Qu'est-ce que la *cohérence quantique* ? Ce phénomène se réfère aux situations dans lesquelles un nombre assez élevé de particules coopèrent au sein d'un même état quantique non emmêlé avec son environnement. (D'une manière générale, le mot « cohérence » signifie que des oscillations situées en des points différents ont lieu au même rythme. Dans le cas de la cohérence *quantique*, les oscillations sont celles de la fonction d'onde et la cohérence traduit le fait que l'on est en présence d'un seul état quantique.) De tels états apparaissent de manière spectaculaire avec la supraconductivité (lorsque la résistance électrique devient nulle) et la superfluidité (lorsque les forces de frottement dans un fluide — la viscosité — deviennent nulles). La caractéristique de ces phénomènes est la présence d'une *barrière d'énergie* que l'environnement doit franchir pour pouvoir perturber l'état quantique. Si la température de l'environnement est très élevée, de sorte que l'énergie de beaucoup des particules ambiantes est suffisamment grande pour franchir cette barrière, la cohérence quantique est détruite. C'est pourquoi les phénomènes de type supraconductivité et superfluidité ne se manifestent normalement qu'aux très basses températures, à quelques degrés seulement au-dessus du zéro absolu. Pour ces raisons — entre autres —, les scientifiques doutent généralement qu'un objet aussi « chaud » qu'un cerveau humain — comme, d'ailleurs, tout autre système biologique — puisse être le siège d'une telle cohérence quantique.

Ces dernières années cependant, des expériences remarquables ont montré qu'avec certaines substances, la supraconductivité peut se manifester à des températures bien plus élevées, même jusqu'à 115 K (*cf.* Sheng *et al.* 1988). C'est encore très froid du point de vue biologique, environ −158°C, à peine un peu plus chaud que l'azote liquide. Mais des observations encore plus remarquables de Laguës *et al.* (1993) semblent indiquer la présence de la supraconductivité à des températures simplement « sibériennes », −23°C. Bien qu'ayant encore lieu dans le « froid », en termes biologiques, cette *supraconductivité à haute température* conforte sérieusement l'hypothèse d'une éventuelle cohérence quantique au sein de systèmes biologiques.

En fait, bien avant que l'on ait observé le phénomène de supraconductivité à haute température, l'éminent physicien Herbert Fröhlich (qui, dans les années trente, accomplit une percée fondamentale dans la compréhension de la supraconductivité « ordinaire », à basse température) avait émis l'hypothèse de l'existence d'effets quantiques collectifs dans des systèmes biologiques. Étudiant un phénomène troublant observé dès 1938 dans les membranes cellulaires, il suggéra en 1968 (en utilisant un concept dû à mon frère Oliver Penrose et à Lars Onsager (1956) — ainsi que je l'ai découvert à ma grande surprise en me documentant sur ce sujet), l'existence d'effets vibratoires au sein des cellules actives, effets qui devaient entrer en résonance avec un rayonnement électromagnétique hyperfréquence de 10^{11} Hz à la suite d'un phénomène de cohérence quantique biologique. Ces effets, qui ne nécessitent pas de basses températures, sont dus à l'existence d'une grande énergie d'origine métabolique. On dispose aujourd'hui de témoignages observationnels fiables confirmant la présence, dans de nombreux systèmes biologiques, de l'effet prédit par Fröhlich en 1968. Nous verrons plus loin (§7.5) quelles conséquences cela peut avoir au niveau de l'activité cérébrale.

7.2 Neurones, synapses et ordinateurs

S'il est encourageant de constater que la cohérence quantique semble pouvoir jouer un rôle significatif dans les systèmes biologiques, aucun lien direct n'a été à ce jour clairement établi entre ce rôle et l'activité cérébrale. Une bonne part de nos connaissances sur le cerveau, bien qu'encore très rudimentaires, nous a conduits à une image classique (essentiellement celle proposée par McCullogh et Pitts en 1943) dans laquelle les neurones et les synapses qui les relient semblent se comporter pratiquement comme les transistors et les fils (les circuits imprimés) des ordinateurs actuels. Pour être plus précis, l'image fournie par la biologie est celle de signaux neuronaux classiques partant du bulbe central (le soma) du neurone et se propageant dans la très longue fibre appelée *axone*, laquelle se divise en divers endroits en brins distincts (Fig. 7.1). Chacun de ces brins se termine par une *synapse* — la jonction où s'opère, à

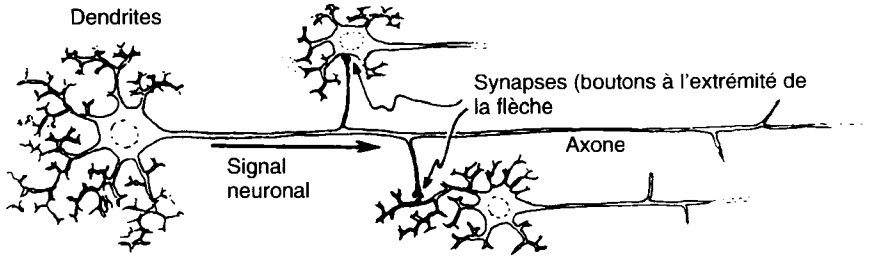


Figure 7.1. Représentation schématique d'un neurone et de ses liaisons synaptiques.

travers un espace synaptique, le transfert du signal, habituellement à un neurone voisin. C'est ici que les substances chimiques que sont les neurotransmetteurs véhiculent le message émis par le neurone précédent, en passant d'une cellule (neurone) à la suivante. Cette jonction synaptique se trouve souvent sur une *dendrite* du neurone suivant, ou sur son soma. Certaines synapses sont excitatrices (leurs neurotransmetteurs tendent à renforcer l'activité du neurone suivant) ; d'autres sont inhibitrices (leurs neurotransmetteurs — différents des précédents — tendent à inhiber l'activité du neurone). L'effet résultant des diverses actions synaptiques sur le neurone suivant est additif (« plus » pour une synapse excitatrice, « moins » pour une synapse inhibitrice), et lorsqu'un certain seuil est atteint, le neurone entre en activité*. Plus exactement, il y aurait une forte *probabilité* pour qu'il entre en activité. Tous ces processus comporteraient également des éléments aléatoires.

Si l'on suppose que les liaisons synaptiques et leurs intensités sont invariantes, il ne fait aucun doute, du moins jusqu'ici, que ce modèle du cerveau peut être en principe correctement simulé sur ordinateur. (Les éléments aléatoires ne poseraient, bien entendu, aucun problème de programmation, *cf.* §1.9.) En fait, il n'est pas difficile de voir que le schéma neurone-synapse présenté ici (avec des synapses et des intensités fixes) est en gros *équivalent* à celui d'un ordinateur (*cf.* EOLP, p. 426-430). Toutefois, du fait de l'existence d'un phénomène appelé *plasticité cérébrale*, les intensités de certaines de ces liaisons — si ce n'est toutes —, et ces liaisons elles-mêmes, peuvent varier de temps en temps, peut-être même sur des échelles de temps inférieures à la seconde. La question importante qui se pose alors est la suivante : quelles procédures gouvernent ces variations synaptiques ?

* Du moins, selon l'image conventionnelle. On pense aujourd'hui que ce schéma « additif » est peut-être une simplification excessive et que les dendrites des neurones seraient le siège d'un « traitement de l'information ». Cette possibilité résulte notamment des travaux de Karl Pribram (*cf.* Pribram 1991). Certaines suggestions avaient déjà été faites dans ce sens par Alwyn Scott (1973, 1977) ; sur l'éventualité d'une « intelligence » au sein des cellules individuelles, voir par exemple Albrecht-Buehler (1985). Cette éventualité d'un « traitement dendritique » complexe au sein de neurones individuels est compatible avec la discussion exposée à la section 7.4.

Dans les modèles connexionnistes (tels ceux adoptés pour les réseaux de neurones formels), on inclut une *règle de calcul* qui régit ces variations. Cette règle permet au système d'améliorer ses résultats sur la base de certains critères prédéfinis liés aux données qu'il reçoit de l'extérieur. Dès 1949, Donald Hebb suggéra une règle simple de ce type, que les modèles connexionnistes modernes[1] ont depuis considérablement modifiée. Il est évident que ces modèles *doivent* comporter une règle de calcul de ce type — puisqu'ils sont programmés sur ordinateur (*cf.* §1.5). Mais l'idée maîtresse des raisonnements que j'ai exposés dans la première partie est qu'aucune de ces procédures de calcul ne peut convenablement expliquer toutes les manifestations mises en jeu lors du processus de compréhension consciente chez l'homme. Nous devons donc chercher quelque chose de différent, par exemple le type adéquat de « mécanisme » de contrôle — du moins dans le cas des modifications synaptiques pouvant être associées à l'activité *consciente.*

D'autres idées ont été suggérées, telles celles émises par Gerald Edelman qui, dans son dernier livre *Bright Air, Brilliant Fire* (1992) (mais aussi dans trois ouvrages antérieurs, *cf.* Edelman 1987, 1988, 1989) propose, au lieu d'un système de règles de type hebbien, l'existence d'une sorte de principe « darwinien » qui, opérant sur ces liaisons à la manière d'une sélection naturelle, permettrait au cerveau d'améliorer en permanence ses résultats. Ce modèle rappelle fortement la manière dont le système immunitaire développe son aptitude à « reconnaître » des substances. Edelman met l'accent sur le rôle complexe joué par les neurotransmetteurs et d'autres substances chimiques dans les communications entre neurones. Toutefois, tels qu'ils sont actuellement compris, ces processus sont encore traités de manière algorithmique et classique. Edelman et ses collègues ont construit une série de dispositifs numériquement contrôlés (baptisés DARWIN I, II, III, IV, etc.) destinés à simuler, par ordre de complexité croissante, les types de procédures censés constituer selon eux le fondement de l'activité mentale. Le simple fait que les opérations de contrôle soient exécutées par un ordinateur standard montre que ce modèle est algorithmique — avec un système de règles « ascendantes ». Peu importent les différences de détail qui le distinguent d'autres procédures de calcul. Il n'en appartient pas moins à la famille que j'ai étudiée dans la première partie de ce livre — voir notamment les sections 1.5, 3.9 et les arguments résumés dans le dialogue imaginaire de la section 3.23. Ces seuls arguments rendent excessivement improbable qu'un modèle de cette nature puisse réellement simuler l'activité de l'esprit conscient.

Pour échapper à l'emprise algorithmique, il faut trouver une autre procédure de contrôle des liaisons synaptiques — et quelle qu'elle soit, elle repose probablement sur un processus physique dans lequel la cohérence quantique joue un rôle déterminant. Si ce processus est essentiellement semblable à l'action du système immunitaire, ce système doit alors lui-même dépendre d'effets quantiques. Peut-être y a-t-il effectivement un élément fondamentalement quantique dans le mécanisme de reconnaissance du système immunitaire — ainsi que l'affirme notamment Michael Conrad (1990, 1992, 1993). Cela ne me surprendrait pas ; mais l'existence de cet élément ne joue aucun rôle central dans le modèle cérébral d'Edelman.

Même si les liaisons synaptiques sont contrôlées par des effets quantiques cohérents, on voit difficilement comment la transmission des signaux neuronaux elle-même pourrait reposer sur un ingrédient fondamentalement quantique. Plus précisément, on peut se demander quel avantage il y aurait à considérer une superposition quantique de *déclenchement* et *non-déclenchement* simultanés d'un même neurone. Les signaux neuronaux semblent suffisamment macroscopiques pour rendre peu crédible une telle image, en dépit du fait que la transmission soit assez bien isolée par la présence de la gaine graisseuse de myéline enveloppant la fibre nerveuse. Selon le critère (**RO**) présenté à la section 6.12, lors d'un déclenchement neuronal, la réduction d'état objective devrait survenir rapidement, non qu'il y ait un important mouvement de masse macroscopique (il est loin d'être suffisant d'après le critère proposé), mais parce que le champ électrique — dû au signal — qui se propage le long du nerf serait probablement détectable par la matière environnante du cerveau. Ce champ perturberait, de manière aléatoire, des domaines relativement vastes de cet environnement — apparemment suffisamment pour que le critère de la section 6.12 sur l'intervention de **RO** soit satisfait presque immédiatement après l'émission du signal. Ainsi, la persistance de superpositions quantiques de déclenchement et de non-déclenchement d'un neurone semble une éventualité très peu plausible.

7.3 Le calcul quantique

Cette perturbation de l'environnement par le déclenchement d'un neurone est selon moi la caractéristique la plus gênante de l'ébauche de proposition que j'avais faite dans EOLP et qui exigeait apparemment des superpositions quantiques de déclenchements et non-déclenchements de familles de neurones. Avec le critère **RO** de réduction d'état, le processus **R** intervient même lorsque la perturbation de l'environnement est inférieure à ce que demandait EOLP et rend encore moins crédible l'éventualité d'une persistance durable de telles superpositions. Dans EOLP, l'idée était que si différentes configurations de déclenchements neuronaux pouvaient exécuter simultanément de nombreux « calculs » distincts superposés, cela signifierait que le cerveau peut accomplir des opérations plus proches d'un *calcul quantique* que d'un simple calcul de type machine de Turing. En dépit de l'apparente improbabilité d'un calcul quantique à ce niveau d'activité cérébrale, il nous sera utile ici d'examiner certains aspects de cette notion.

Le calcul quantique est un concept théorique dont les fondements furent posés par David Deutsch (1985) et Richard Feynman (1985, 1986) — *cf.* aussi Benioff (1982) et Albert (1983) — et qui fait aujourd'hui l'objet de recherches très actives. L'idée est d'étendre au niveau quantique la notion classique de machine de Turing. Toutes les opérations exécutées par cette

« machine » sont donc soumises aux lois quantiques, ce qui autorise des superpositions. Ainsi, l'évolution de cette machine est la plupart du temps régie par **U** dont l'action a fondamentalement pour effet de préserver ces superpositions. La procédure **R** n'intervient qu'à la *fin* de l'exécution, lorsqu'on « mesure » le système pour obtenir le résultat du calcul. En fait (bien que ceci ne soit pas toujours admis), **R** intervient aussi ponctuellement — de façon mineure — lors de l'exécution pour voir si le calcul n'est pas déjà terminé.

Il se trouve que bien qu'un ordinateur quantique ne puisse faire mieux que ce que peut accomplir *en principe* une machine de Turing conventionnelle, il existe certaines classes de problèmes pour lesquelles un calcul quantique s'avère supérieur, au sens de la *théorie de la complexité* (*cf.* Deutsch 1985), à un calcul de type Turing. Autrement dit, pour ces classes de problèmes, l'ordinateur quantique est en principe bien *plus rapide* — mais *seulement* plus rapide — qu'un ordinateur conventionnel. Voir Deutsch et Jozsa (1992) pour une classe de problèmes intéressants (bien qu'un peu artificiels) dans laquelle l'ordinateur quantique se distingue particulièrement. Signalons également que d'après un raisonnement récent de Peter Shor, le calcul quantique permet de résoudre (en un temps polynomial) l'important problème de la factorisation des grands nombres entiers.

Rappelons-le, le calcul quantique « standard » obéit aux règles usuelles de la théorie quantique — le système évolue selon la procédure **U** durant la majeure partie de l'opération — tout en étant, à certains endroits précis, soumis à l'intervention de **R**. Rien dans l'ensemble de cette procédure n'est « irréductible au calcul », au sens ordinaire du mot « calcul », car **U** est une opération calculable et **R** une procédure purement probabiliste. Tout calcul réalisable sur un ordinateur quantique pourrait être en principe effectué par une machine de Turing associée à un dispositif de randomisation. Ainsi, selon les arguments de la première partie de ce livre, aucun ordinateur quantique ne pourrait effectuer les opérations intervenant dans le processus de compréhension humaine. Reste alors l'espoir que les subtilités propres à la *vraie* réduction du vecteur d'état — par opposition à la la procédure aléatoire bouche-trou **R** — nous fassent découvrir un processus authentiquement *non calculable*. Ainsi, la théorie complète de l'hypothétique processus **RO** serait une théorie *fondamentalement non calculable*.

Dans EOLP, l'idée était que l'on ne peut exécuter d'un trait des superpositions de calculs de type machine de Turing mais que l'on doit faire intervenir dans cette exécution une action non calculable, interprétable uniquement en termes d'une physique nouvelle (*e.g.* **RO**) amenée à se substituer à **R**. Mais si ces superpositions de calculs neuronaux nous sont interdites — parce que chaque signal neuronal provoque une perturbation d'un trop grand domaine de l'environnement —, on voit difficilement comment on pourrait ne serait-ce qu'exploiter le concept de calcul quantique standard, voire une modification de ce concept qui tirerait avantage d'une hypothétique substitution de **R** par une procédure non calculable telle que **RO**. Nous verrons toutefois dans un instant qu'il existe une autre possibilité bien plus prometteuse. Afin de

comprendre pourquoi il en est ainsi, nous devons d'abord examiner plus en détail l'anatomie des cellules cérébrales.

7.4 Cytosquelettes et microtubules

Si l'on croit que le comportement sophistiqué des animaux est régi par leurs seuls neurones, on se trouve confronté à un problème profond dès que l'on envisage l'humble paramécie (Fig. 7.2). Celle-ci se déplace dans son étang grâce à d'innombrables cils l'entourant tels une chevelure, se précipite sur toute nourriture bactérienne dont elle détecte la présence à l'aide de mécanismes variés, et bat en retraite au moindre danger, prête à filer dans une autre direction. Elle sait aussi triompher d'un obstacle en le contournant et tirerait même des leçons de ses expériences passées[2] — bien que cette remarquable aptitude soit contestée par certains[3]. Comment un animal totalement dépourvu de neurones et de synapses — qui n'est en vérité qu'une simple cellule (même pas un neurone) et n'a donc pas la place de loger de tels accessoires — peut-il accomplir tout cela ?

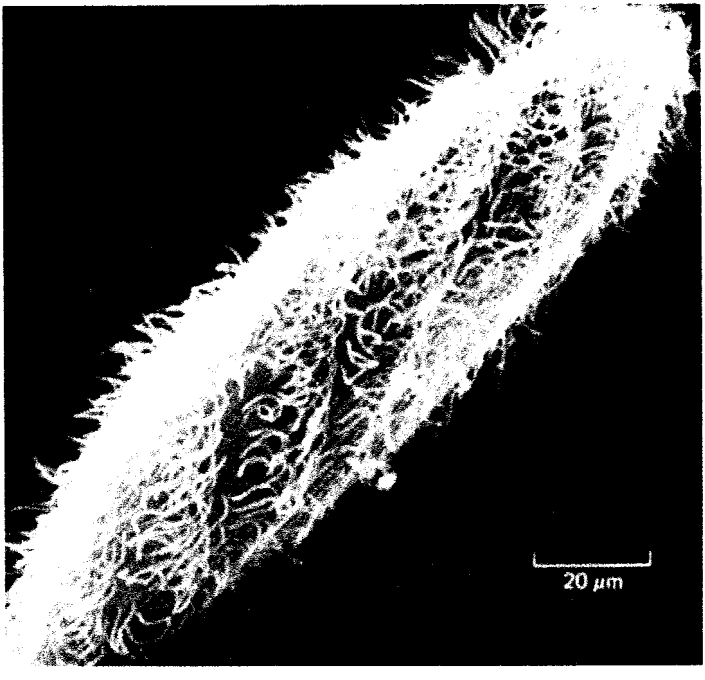


Figure 7.2. Une paramécie. Remarquez les cils périphériques utilisés comme organes de locomotion. Ces cils sont les extrémités externes du cytosquelette de la paramécie.

Le comportement de la paramécie, à l'instar de celui de tout autre animal unicellulaire — par exemple, une amibe —, est certes régi par un système de contrôle complexe, mais ce n'est pas un système nerveux. La structure responsable de ce contrôle fait apparemment partie de ce que l'on appelle le *cytosquelette.* Si, comme le suggère son nom, le cytosquelette est la charpente qui maintient la forme de la cellule, il remplit bien d'autres fonctions — il sert de « tapis roulant » pour le transport de diverses molécules au sein de la cellule ; les cils eux-mêmes sont les terminaisons de fibres cytosquelettiques. En résumé, le cytosquelette d'une cellule est plutôt une combinaison de squelette, de système musculaire, de système locomoteur, de système circulatoire et de système nerveux, tout cela en un seul bloc !

C'est surtout le cytosquelette en tant que « système nerveux » cellulaire qui va nous intéresser ici, car nos propres neurones sont eux-mêmes des cellules individuelles et chaque neurone possède son *propre* cytosquelette ! Cela signifie-t-il que l'on pourrait en un sens considérer que chaque neurone est doté d'une sorte de « système nerveux personnel » ? C'est là un problème fascinant, et de nombreux scientifiques ont à un moment ou un autre envisagé une idée de ce genre. (Voir l'ouvrage pionnier de Stuart Hameroff, *Ultimate Computing : Biomolecular Consciousness and NanoTechnology* (1987) ; voir aussi Hameroff et Watt (1982) et nombre d'articles de la revue *Nanobiology.*)

Afin de pouvoir examiner ces problèmes, nous allons d'abord jeter un coup d'œil sur l'organisation de base du cytosquelette. Celle-ci consiste en molécules protéiques arrangées selon divers types de structures : actine, microtubules et filaments intermédiaires. Nous nous intéresserons particulièrement aux *microtubules.* Ce sont des tubes creux cylindriques d'environ 25 nm de diamètre extérieur et 14 nm de diamètre intérieur (« nm » = « nanomètre » = 10^{-9} m), parfois groupés en fibres tubulaires — composées de 9 doublets, triplets ou triplets partiels de microtubules — dont la section évoque un manège (Fig.7.3), avec parfois, au centre, une paire de microtubules. Les cils de la paramécie ont une structure de ce type. Chaque microtubule est en lui-même un polymère protéique composé de sous-unités appelées *tubulines.* Chaque tubuline est un « dimère » : elle se compose d'une paire de protéines globulaires — deux parties pratiquement disjointes appelées α-tubuline et β-tubuline, chacune d'elles se composant d'environ 450 acides aminés —, ressemble plus ou moins à une cacahuète et s'inscrit dans un réseau hexagonal oblique couvrant toute la longueur du tube (Fig. 7.4). Chaque microtubule comprend généralement 13 colonnes de dimères. Chaque dimère mesure environ 8 nm × 4 nm × 4 nm et son nombre atomique (le nombre de nucléons qui le composent) vaut environ 11×10^4 (ce qui signifie que sa masse, en unités absolues, vaut environ 10^{-14}).

Ces dimères peuvent se présenter selon (au moins) deux configurations géométriques différentes — appelées *conformations.* Dans l'une d'elles, ils s'inclinent de 30 degrés dans la direction du microtubule. Ces deux conformations correspondent à deux états de polarisation électrique différents dus au fait qu'un électron, situé au centre de la jonction α-tubuline/β-tubuline, peut se décaler d'une position à une autre.

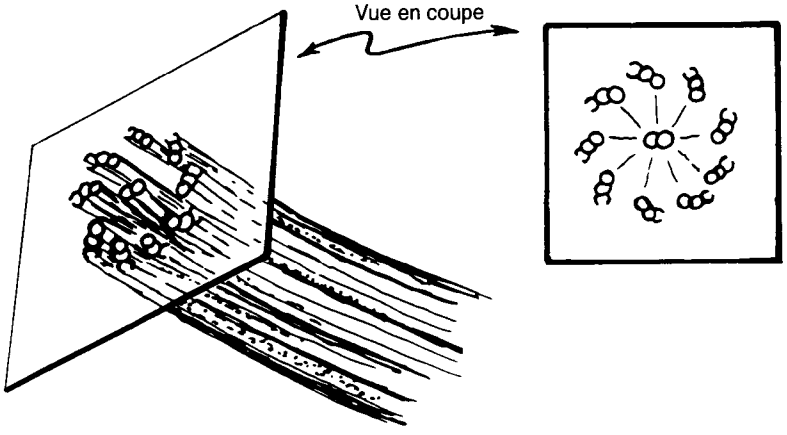


Figure 7.3. Une part importante du cytosquelette se compose de faisceaux de tubes minuscules (les microtubules). Vus en coupe, ces tubes évoquent un manège. Les cils de la paramécie sont composés de ce type de faisceaux.

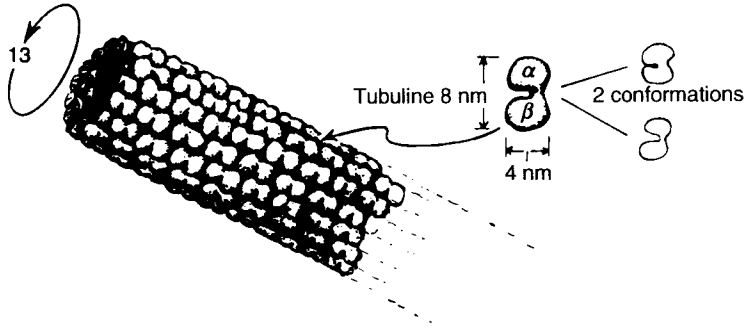


Figure 7.4. Un *microtubule.* C'est un tube creux, composé normalement de 13 colonnes de tubulines (qui sont des dimères). Chaque tubuline peut prendre (au moins) deux conformations.

Le « centre de contrôle » du cytosquelette (si le terme est vraiment approprié) est une structure appelée *centriole.* Celle-ci semble constituée essentiellement de deux cylindres de neuf triplets de microtubules, formant une sorte de « T » disjoint (Fig. 7.5). (D'une manière générale, ces cylindres ressemblent à ceux présents dans les cils ; *cf.* Fig. 7.3.) Le centriole constitue l'élément clé d'une structure appelée *centre organisateur des microtubules* ou *centrosome.* En dehors de son rôle durant les phases ordinaires de la vie cellulaire, le centriole accomplit au moins une tâche fondamentale. À un moment critique, chacun des deux cylindres qui le constituent se dédouble pour donner naissance à *deux* « T » qui ensuite se *séparent* l'un de l'autre, chaque « T » entraînant avec lui un faisceau de microtubules — il serait cependant plus précis de dire que chacun devient un point focal autour duquel se rassemblent ces microtubules. Ces

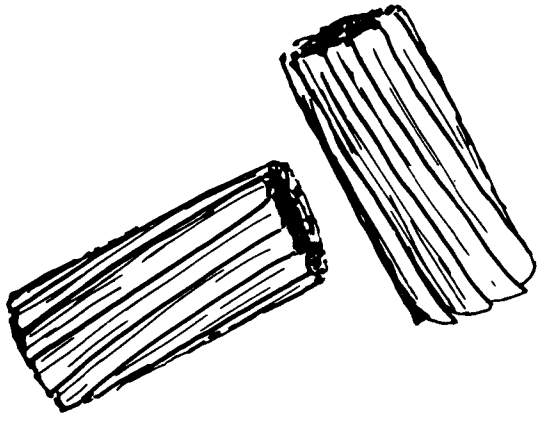


Figure 7.5. Le *centriole* (qui semble être le centre de contrôle — s'il existe — du cytosquelette) se compose essentiellement d'un « T » disjoint formé de deux faisceaux de microtubules très semblables à ceux présentés à la figure 7.3.

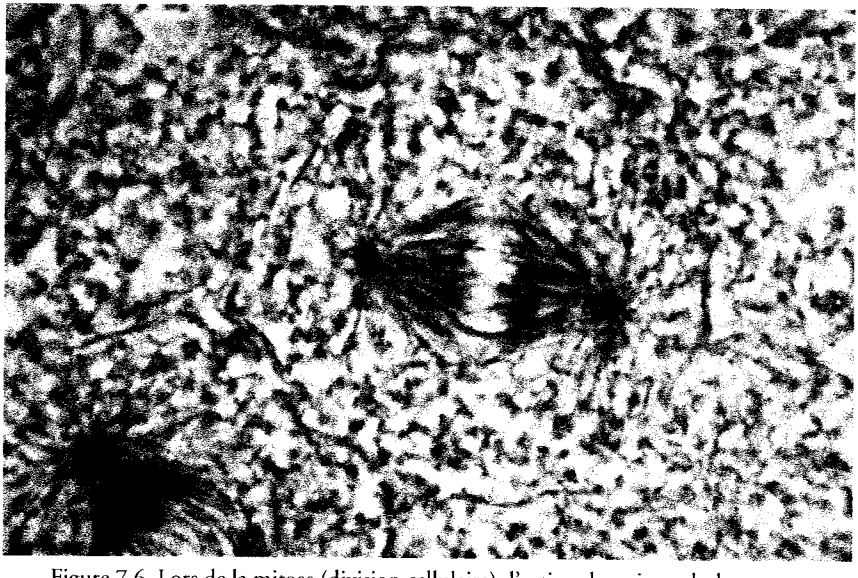


Figure 7.6. Lors de la mitose (division cellulaire), l'action des microtubules provoque la division des chromosomes.

fibres microtubulaires connectent le centriole aux divers brins d'ADN présents dans le noyau (aux points centraux appelés centromères), puis les brins d'ADN se séparent, initiant le fabuleux processus de la *mitose* — la *division cellulaire* (Fig. 7.6).

La présence de deux « quartiers généraux » au sein d'une même cellule peut paraître curieuse. D'une part, il y a le *noyau*, dans lequel réside le matériau

génétique de la cellule, matériau fondamental qui non seulement contrôle l'hérédité et l'identité propre de la cellule, mais aussi régit la production des substances protéiques cellulaires. D'autre part, il y a le *centrosome* avec son constituant clé, le *centriole*, qui semble être le point focal du cytosquelette, structure qui apparemment contrôle les mouvements de la cellule et le détail de son organisation. On pense que la présence de ces deux structures différentes dans les cellules eucaryotes (les cellules de tous les animaux et de presque toutes les plantes de notre planète — à l'exception des bactéries, des algues bleu-vert et des virus) serait due à une « infection » survenue il y a des milliers de millions d'années. Les cellules qui habitaient antérieurement la Terre étaient des cellules procaryotes qui existent encore aujourd'hui sous forme de bactéries et d'algues bleu-vert, dépourvues de cytosquelette. Selon Carl Sagan (1976), certaines cellules procaryotes primitives se seraient mêlées à — ou auraient été « infectées » par — un type de spirochète, organisme se déplaçant à l'aide d'un long flagelle composé de protéines cytosquelettiques. Ces organismes mutuellement étrangers auraient par la suite appris à vivre ensemble, en relation symbiotique, sous forme de cellules *eucaryotes*. Ainsi, ces « spirochètes » sont finalement devenus les cytosquelettes des cellules — avec toutes les conséquences que cela a entraîné pour l'évolution qui *nous* a permis d'exister !

L'organisation des microtubules de mammifères est intéressante d'un point de vue mathématique. Le nombre 13, qui semble n'avoir aucune importance mathématique particulière, appartient en fait à la célèbre *suite de Fibonacci* :

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \ldots$$

dont chaque élément est égal à la somme de ses deux prédécesseurs. Les nombres de Fibonacci sont célèbres car ils apparaissent fréquemment (à une échelle bien plus grande) dans les systèmes biologiques. Par exemple, sur les pommes de pin, les fleurs de tournesol et les troncs de palmiers, on observe des interpénétrations de spirales ou d'hélices tournant tantôt vers la droite, tantôt vers la gauche, et dans lesquelles le nombre de rangées correspondant à une orientation et celui correspondant à l'autre sont deux nombres de Fibonacci consécutifs (Fig. 7.7). (Lorsqu'on examine ces structures d'une extrémité à l'autre, on constate parfois un « glissement » ; les nombres prennent alors deux autres valeurs voisines.) Curieusement, le motif hexagonal oblique présent sur les microtubules exhibe une caractéristique très semblable — qui est même généralement plus précise — et semble (du moins, habituellement) constitué de 5 et 8 arrangements hélicoïdaux orientés respectivement vers la droite et vers la gauche (Fig. 7.8). À la figure 7.9, j'ai tenté de montrer comment se présente cette structure « vue » de l'intérieur d'un microtubule. Le nombre 13 correspond à la somme 5 + 8. Autre curiosité, la frontière extérieure de microtubules relativement fréquents, à savoir les microtubules doubles, semble normalement contenir un total de 21 — le nombre suivant dans la suite de Fibonacci ! — colonnes de tubulines. (Il ne faut toutefois pas s'emballer avec de telles considérations ; par exemple, le

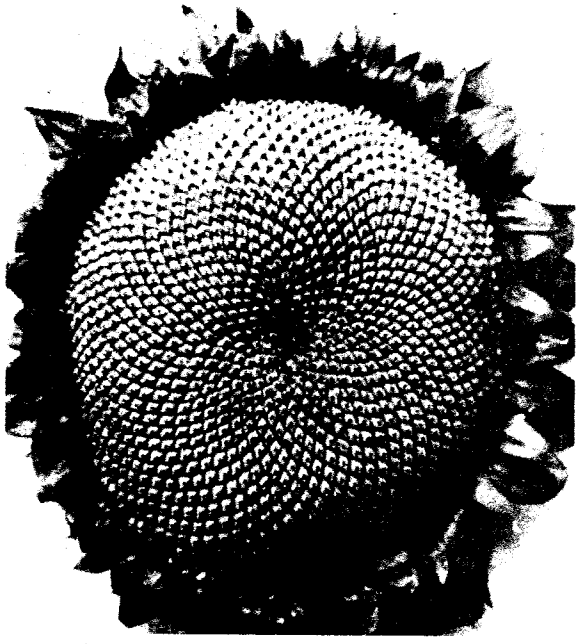


Figure 7.7. Une fleur de tournesol. Comme de nombreuses autres plantes, cette fleur a une structure fortement marquée par les nombres de Fibonacci. Ses régions périphériques contiennent 89 spirales orientées dans le sens des aiguilles d'une montre, et 55 orientées en sens inverse. La région centrale se caractérise par d'autres nombres de Fibonacci.

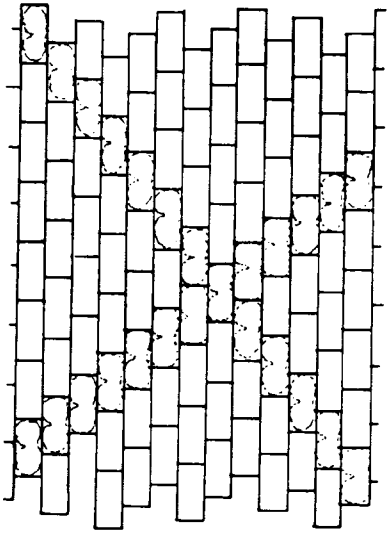


Figure 7.8. Imaginez un microtubule fendu sur toute sa longueur puis étalé sur un plan. On voit alors que les tubulines sont disposées selon des droites inclinées qui se rejoignent d'un bord à l'autre 5 ou 8 niveaux plus loin (selon que les pentes sont ascendantes ou descendantes).

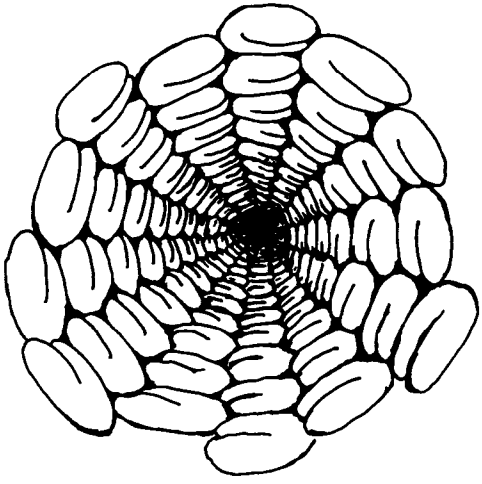


Figure 7.9. L'intérieur d'un microtubule ! On y voit la disposition spiralée (5 + 8) des tubulines.

« 9 » apparaissant dans les faisceaux de microtubules des cils et des centrioles *n'est pas* un nombre de Fibonacci.)

Pour quelle raison les nombres de Fibonacci interviennent-ils dans la structure microtubulaire ? Pour les pommes de pin, les fleurs de tournesol, etc., diverses théories plausibles ont été avancées — Alan Turing lui-même y a longuement réfléchi (Hodges 1983, p. 437) —, mais rien n'indique que ces théories soient adaptées aux microtubules et qu'à ce niveau, il ne faille pas recourir à des idées différentes. Selon Koruga (1974), les nombres de Fibonacci permettraient au microtubule d'améliorer son aptitude à « traiter l'information ». De fait, Hameroff et ses collègues affirment depuis plus d'une décennie[4] que les microtubules joueraient un rôle en tant qu'*automates cellulaires* traitant et transmettant, le long des tubes, des signaux complexes. Ces signaux seraient des ondes dues aux différences de polarisation des tubulines. Rappelons que les tubulines peuvent exister sous (au moins) deux conformations différentes pouvant se transformer l'une en l'autre, apparemment parce qu'elles admettent des états de polarisation électrique différents. L'état de chaque dimère serait influencé par les états de polarisation de chacun de ses six voisins (à cause des forces de van der Waals qui les unissent), ce qui donnerait naissance à certaines règles spécifiques gouvernant la conformation de chaque dimère en fonction de celle de ses voisins. Cela permettrait la transmission et le traitement de tous types de signaux le long de chaque microtubule. Ces signaux semblent dépendre d'une part de la manière dont les microtubules véhiculent diverses molécules, d'autre part des diverses interconnexions entre microtubules voisins. Ces interconnexions sont des ponts protéiques appelés PAM, pour « protéines associées aux microtubules » (Fig. 7.10). Selon Koruga, l'efficacité serait particulièrement élevée dans le cas d'une structure liée par des nombres

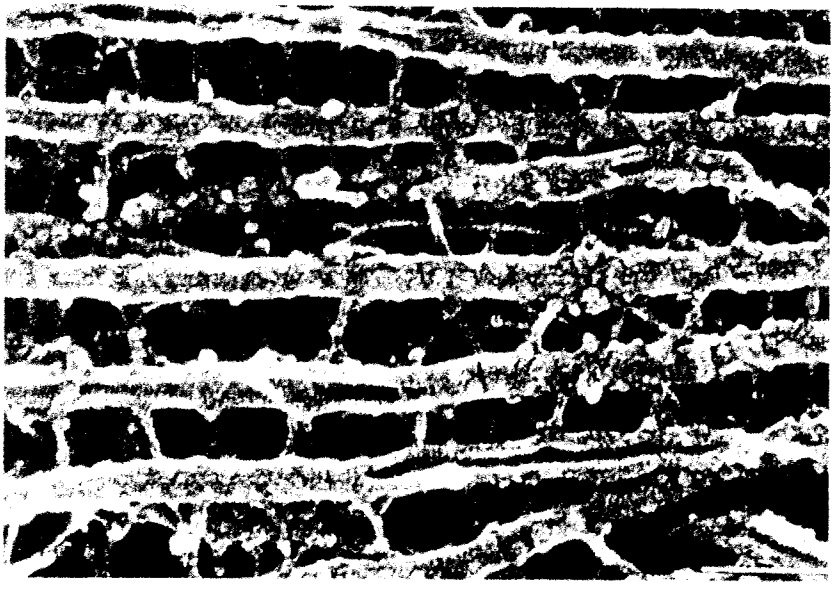


Figure 7.10. Les microtubules sont en général reliés entre eux par des ponts de *protéines associées aux microtubules* (PAM).

de Fibonacci, telle celle observée pour les microtubules. Cette organisation des microtubules semble d'ailleurs avoir une bonne raison d'être, car bien qu'il y ait en général quelques variations pour les nombres concernant les cellules eucaryotes, 13 colonnes semble être une valeur pratiquement universelle parmi les microtubules de mammifères.

Examinons maintenant l'importance des microtubules pour les neurones. Chaque neurone, nous l'avons dit, possède son propre cytosquelette. Je suis persuadé que bien des choses restent à découvrir sur le cytosquelette, mais il ne nous est pas du tout inconnu. En particulier, nous savons que les microtubules neuronaux peuvent atteindre des longueurs assez considérables — de l'ordre de plusieurs millimètres — comparées à leur diamètre — qui vaut seulement 25 à 30 nm. En outre, ils peuvent, selon les circonstances, se dilater ou se contracter, et véhiculent des molécules neurotransmettrices. Ils sont présents tout le long des axones et des dendrites. Bien qu'individuellement ils ne semblent pas couvrir toute la longueur de l'axone, ils forment certainement des réseaux de communication qui eux, couvrent tout l'axone, chaque microtubule communiquant avec ses voisins par l'intermédiaire des PAM mentionnées à l'instant. Les microtubules semblent être responsables du maintien des intensités des liaisons synaptiques ; nul doute que ce sont eux également qui modifient ces intensités si besoin est. En outre, ils semblent régir la croissance des nouvelles terminaisons nerveuses et guider leur cheminement pour établir des liaisons avec d'autres cellules nerveuses.

Les neurones ne se divisant plus une fois le cerveau totalement formé, leurs centrioles n'ont aucun rôle de ce type. En fait, les centrioles semblent absents

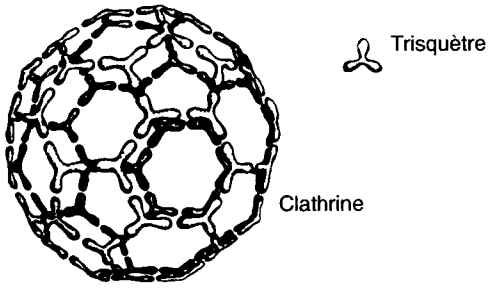


Figure 7.11. Une clathrine. La forme d'ensemble évoque celle d'un fullerène qui serait composé de sous-structures plus complexes — des trisquètres protidiques au lieu d'atomes de carbone. La structure de la clathrine représentée ici rappelle un ballon de football.

des centrosomes neuronaux — qui se trouvent près des noyaux neuronaux. Les microtubules s'étendent depuis les centrosomes jusqu'au voisinage des terminaisons présynaptiques de l'axone ; ils pénètrent également, dans l'autre sens, à l'intérieur des dendrites et, grâce à l'actine contractile, à l'intérieur des épines dendritiques qui forment souvent la terminaison post-synaptique d'un espace synaptique (Fig. 7.12). Ces épines dendritiques connaissent une croissance et une dégénérescence, deux processus qui semblent constituer une part importante de la plasticité cérébrale responsable des subtiles variations affectant en permanence l'ensemble des interconnexions cérébrales. Il semble bien que les microtubules participent de manière importante au contrôle de la plasticité cérébrale.

On peut également mentionner, à titre d'apparente curiosité, que les terminaisons présynaptiques des axones contiennent, associées aux microtubules, certaines substances géométriquement fascinantes qui jouent un rôle important dans l'émission des neurotransmetteurs. Ces substances — appelées *clathrines* — sont constituées de trimères protéiques appelés « trisquètres ». Ces trisquètres, des étoiles (polypeptidiques) à trois branches, s'assemblent pour former de belles configurations géométriques identiques, par leur structure d'ensemble, aux molécules de carbone appelées « fullerènes »[5]. Les clathrines sont toutefois bien plus grosses que les fullerènes, car constituées non pas de simples atomes de carbone, mais de trisquètres qui contiennent eux-mêmes plusieurs acides aminés. Les clathrines liées à l'émission des substances neurotransmettrices survenant aux jonctions synaptiques semblent avoir une structure d'*icosaèdre tronqué* — autrement dit, ressemblent à des ballons de football (*cf.* Fig. 7.11 et 7.12) !

À la section 7.2, nous avons été conduits à nous demander quel processus gouverne les variations d'intensités synaptiques et décide des liaisons synaptiques qui doivent être activées. Nous l'avons dit, tout indique que le *cytosquelette* joue un rôle central dans ce processus, mais cela nous aide-t-il dans notre recherche d'un rôle non calculable de l'esprit ? Tout ce que nous avons obtenu

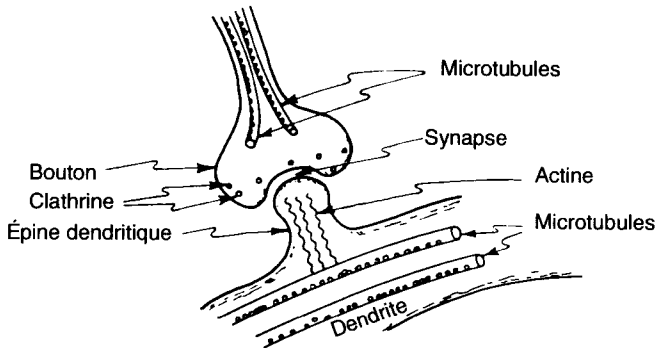


Figure 7.12. Les clathrines (*cf.* Fig. 7.11) et les terminaisons microtubulaires se situent au niveau des boutons synaptiques de l'axone et semblent participer au contrôle de l'intensité de la liaison synaptique ; cette intensité serait également influencée par les filaments contractiles d'actine présents dans les épines dendritiques contrôlées, elles, par les microtubules.

pour l'instant est un énorme accroissement de la puissance de calcul par rapport à la situation dans laquelle les entités de base étaient les neurones.

D'ailleurs, si les tubulines sont les unités de calcul, le cerveau pourrait avoir une puissance excédant immensément celle envisagée par les tenants de l'IA. Dans son livre *Mind Children* (1988), Hans Moravec a supposé, en se fondant sur un modèle où n'entrent en jeu que les seuls neurones, que le cerveau humain pourrait en principe accomplir quelque 10^{14} opérations fondamentales par seconde — mais pas plus —, dans le sens où il y aurait quelque 10^{11} neurones opérationnels, chacun émettant environ 10^3 signaux par seconde (*cf.* §1.2). Supposons en revanche que l'unité fondamentale de calcul soit la tubuline. Comme il existe environ 10^7 dimères par neurone et que les opérations élémentaires sont, dans cette hypothèse, exécutées quelque 10^6 fois plus vite, on obtient un total d'environ 10^{27} opérations par seconde. Si les ordinateurs actuels sont en mesure d'approcher la première valeur de 10^{14} opérations par seconde, ainsi que le soutiennent avec force Moravec et d'autres chercheurs, il n'y a aucun espoir d'atteindre les 10^{27} opérations par seconde dans un avenir prévisible.

Bien sûr, on pourrait légitimement objecter que le rendement du cerveau est loin d'atteindre 100 %. Il est clair cependant que la possibilité d'un « calcul microtubulaire » (*cf.* Hameroff 1987) jette une lumière totalement nouvelle sur la prétendue imminence d'une intelligence artificielle égalant le niveau humain. Peut-on même croire que l'on a atteint, comme certains le prétendent[6], les facultés mentales des nématodes pour la simple raison que l'on semble être parvenu à cartographier et à simuler numériquement leur organisation neuronale ? Nous l'avons remarqué à la section 1.15, les aptitudes actuelles d'une fourmi semblent surpasser de loin celles des procédures standard de l'IA. On est en droit de se demander quels avantages retire une fourmi de son énorme collection de « processeurs d'information

microtubulaires », par rapport à ce qu'elle pourrait faire si elle ne disposait que de « commutateurs neuronaux ». Sans parler de la paramécie.

Pourtant, les arguments que j'ai développés dans la première partie affirment bien plus, à savoir que la faculté de compréhension humaine est irréductible à tout schéma numérique, quel qu'il soit. Si nous admettons que les microtubules contrôlent l'activité cérébrale, il nous faut donc rechercher dans leur comportement un mécanisme différent d'un simple calcul. J'ai affirmé qu'une telle action non calculable est alors probablement le résultat d'un phénomène de cohérence quantique se situant à une échelle relativement grande et couplé de manière subtile au comportement macroscopique, de sorte que le système utilise un processus physique encore inconnu et qui devrait remplacer la procédure bouche-trou **R** de la physique actuelle. Examinons dans un premier temps ce que pourrait être le rôle de la *cohérence quantique* dans l'activité du cytosquelette.

7.5 Une cohérence quantique au sein des microtubules ?

Quels sont les indices ? Rappelons les idées qui ont suggéré à Fröhlich (1975) la possibilité d'une cohérence quantique dans les systèmes biologiques (*cf.* §7.1). Il affirma que si l'énergie métabolique est suffisamment grande et si les propriétés diélectriques des matériaux concernés sont suffisamment extrêmes, il existe alors une possibilité de cohérence quantique à grande échelle, similaire à celle qui se manifeste en supraconductivité et en superfluidité — parfois appelée *condensation* de *Bose-Einstein* —, même aux températures relativement élevées qui sont celles des systèmes biologiques. Non seulement il s'avère que l'énergie métabolique est suffisamment élevée et les propriétés diélectriques anormalement extrêmes (c'est là l'observation frappante réalisée dans les années trente qui inspira ses idées à Fröhlich), mais on dispose aujourd'hui de la preuve directe de la présence — prédite par Fröhlich — d'oscillations à 10^{11} Hz au sein des cellules (Grundler et Keilmann 1983).

Un condensat de Bose-Einstein (qui se forme également lors de l'action d'un laser) met en présence un grand nombre de particules se trouvant dans un même état quantique. La fonction d'onde associée à cet état ressemble à celle de chacune des particules, mais s'applique maintenant à leur ensemble. Dans un condensat de Bose-Einstein, l'état quantique du système entier est essentiellement une amplification de l'état quantique d'une seule particule. Il y a donc une cohérence à grande échelle, et cette cohérence maintient, à un niveau macroscopique, nombre des étranges caractéristiques des fonctions d'onde quantiques.

Si l'idée originelle de Fröhlich semble avoir été que ces états quantiques à grande échelle existeraient dans les membranes cellulaires*, nous sommes aujourd'hui face à une autre éventualité — peut-être plus plausible —, à savoir qu'un tel comportement quantique existerait aussi dans les *microtubules*. Certains indices semblent montrer que c'est effectivement le cas[7]. Dès 1974, Hameroff (1974) suggéra que les microtubules pouvaient se comporter comme des « guides d'ondes diélectriques ». Il est en effet tentant de penser que la nature ait eu de bonnes raisons de choisir les tubes creux de ces structures cytosquelettiques. Peut-être ces tubes sont-ils l'isolant efficace qui permet à l'état quantique à l'intérieur du tube de rester, durant un temps appréciable, non emmêlé avec son environnement. À cet égard, il est intéressant de noter qu'Emilio del Giudice et ses collègues de l'université de Milan (del Giudice *et al.* 1983) ont affirmé qu'une action quantique autoconvergente d'ondes électromagnétiques à l'intérieur du matériau cytoplasmique provoque le confinement des signaux dans un volume dont le diamètre est exactement égal au diamètre interne des microtubules. Cela conforterait la théorie des guides d'ondes. Cet effet pourrait aussi participer à la formation même des microtubules.

Un autre aspect intéressant concerne la nature de l'*eau*. Les tubes sont apparemment vides — ce qui est à la fois curieux et éventuellement important, dans la mesure où ils devraient fournir des conditions favorables à des oscillations quantiques collectives. « Vide » signifie ici qu'ils ne contiennent pratiquement que de l'eau (sans même d'ions dissous). On pourrait penser que l'eau, avec ses molécules animées de mouvements aléatoires, n'est pas une structure suffisamment organisée pour être le siège d'oscillations quantiques cohérentes. Toutefois, l'eau des cellules n'a absolument rien à voir avec l'eau ordinaire que l'on trouve dans les océans — eau désordonnée, dont les molécules se déplacent de manière aléatoire et incohérente. Une part de l'eau cellulaire — dans une proportion qui est toutefois l'objet de controverse — existe dans un état *ordonné* (cette eau est parfois qualifiée de « vicinale » ; *cf.* Hameroff (1987), p. 172). Cette eau ordonnée s'étendrait jusqu'à au moins 3 nm hors du cytosquelette. On peut légitimement supposer que l'eau contenue dans les microtubules est également dans un état ordonné, ce qui favoriserait sérieusement la présence éventuelle d'oscillations quantiques cohérentes à l'intérieur, ou dans le voisinage, de ces tubes (*cf.* notamment Jibu *et al.* 1994).

Quel que soit le sort réservé à ces idées fascinantes, une chose me semble claire : il y a peu de chances qu'une description entièrement classique du cytosquelette puisse expliquer correctement son comportement. Il en va tout

* S'inspirant des idées de Fröhlich, Ian Marshall (1989) affirme avec force que la condensation de Bose-Einstein fournirait le « sens unitaire du moi » qui semble caractéristique de la conscience. Voir aussi Zohar (1990), Zohar et Marshall (1994) et Lockwood (1989). Avant eux, Karl Pribram (1966, 1975, mais aussi 1991) avait soutenu l'existence, dans l'ensemble du cerveau, d'une activité « hologrammique » cohérente à grande échelle (et essentiellement quantique).

à fait autrement avec les neurones, pour lesquels une description entièrement classique semble suffire amplement. De fait, l'examen de la littérature montre que l'étude du comportement cytosquelettique fait continuellement appel à des concepts quantiques, et je suis persuadé qu'il en sera de plus en plus ainsi.

Il est cependant également clair que nombre de chercheurs continueront de douter de l'existence d'un effet quantique significatif dans l'activité cérébrale ou cytosquelettique. Même si le fonctionnement des microtubules et l'activité cérébrale consciente reposent de manière cruciale sur des effets quantiques, ceux-ci risquent fort de se dérober à toute mise en évidence expérimentale. Si la chance nous sourit, certaines des procédures standard qui permettent déjà de démontrer la présence de condensats de Bose-Einstein — telle la supraconductivité à haute température — dans des systèmes physiques seront peut-être également applicables aux microtubules. Dans le cas contraire, il nous faudra chercher des idées d'expériences nouvelles. Une possibilité fascinante consisterait à démontrer que les excitations des microtubules présentent une non-localité analogue à celle exhibée par les phénomènes EPR (inégalités de Bell, etc. ; *cf.* §5.3, §5.4, §5.17), car ces excitations n'admettent aucune interprétation classique (locale). On pourrait par exemple imaginer des mesures effectuées en deux points d'un même microtubule — voire de microtubules distincts –, mesures qui ne pourraient s'expliquer à l'aide d'actions classiques indépendantes survenant en chacun de ces deux points.

Quoi qu'il advienne de ces suggestions, les recherches sur les microtubules en sont encore relativement à leurs débuts et je suis persuadé qu'elles nous réservent des surprises de taille.

7.6 Microtubules et conscience

Il existe toutefois un témoignage direct indiquant un lien entre le phénomène de la *conscience* et l'activité du cytosquelette et, en particulier, des microtubules. Ce témoignage, que nous allons maintenant examiner, concerne paradoxalement le problème de l'*absence* de conscience !

L'une des approches les plus fécondes du problème des fondements physiques de la conscience consiste à déterminer quels agents provoquent sa disparition. L'*anesthésie générale* a précisément cette propriété — totalement réversible si les concentrations ne sont pas trop élevées — et, fait remarquable, est induite par un large éventail de substances totalement différentes qui semblent n'avoir aucun lien chimique entre elles. Parmi les anesthésiques, on trouve des substances aussi diverses que le protoxyde d'azote (N_2O), l'éther ($CH_3CH_2OCH_2CH_3$), le chloroforme ($CHCl_3$), l'halothane ($CF_3CHClBr$), l'isofluorane ($CHF_2OCHClCF_3$) et même le xénon, gaz chimiquement inerte !

Si la chimie n'est pas responsable de l'anesthésie générale, quel en est le responsable ? Il existe en fait d'autres forces intermoléculaires, bien plus faibles que les forces chimiques. L'une d'elles, la force de van der Waals, est une attraction faible entre molécules dotées de *moments dipolaires électriques* (l'équivalent « électrique » des moments dipolaires magnétiques qui mesurent l'intensité des aimants ordinaires). Rappelons que les tubulines peuvent avoir deux conformations différentes, dues à la présence d'un électron — situé au centre de chaque dimère dans une région dépourvue d'eau — pouvant occuper deux positions distinctes. La forme d'ensemble du dimère, ainsi que son moment dipolaire électrique, dépendent de la position de cet électron. L'aptitude du dimère à « basculer » d'une conformation à l'autre dépendant de la force de van der Waals exercée par les substances environnantes, Hameroff et Watt (1983) ont suggéré que les anesthésiques pourraient agir par l'intermédiaire de leurs interactions de van der Waals (dans les régions « hydrophobes » — d'où l'eau a été expulsée — ; *cf.* Franks et Lieb (1982)), interactions qui interféreraient avec les « basculements » normaux de la tubuline. Lorsque des gaz anesthésiants diffusent dans des cellules nerveuses individuelles, leurs propriétés dipolaires électriques (qui n'ont pas à être liées à leurs propriétés chimiques ordinaires) pourraient donc interrompre l'action des microtubules. C'est certainement là une explication plausible du fonctionnement des anesthésiques. Bien qu'aucun modèle détaillé de ce fonctionnement ne fasse l'unanimité des chercheurs, ceux-ci semblent reconnaître que cette action opère par l'intermédiaire des forces de van der Waals exercées par ces substances sur la conformation des protéines cérébrales. Il existe en outre une forte possibilité pour que les protéines concernées soient les tubulines constituant les microtubules neuronaux — et que l'interruption du fonctionnement des microtubules qui en résulte se traduise par une perte de conscience.

Cette action directe des anesthésiques sur le *cytosquelette* se trouve confortée par le fait que ces substances n'immobilisent pas seulement les « animaux supérieurs » tels les mammifères ou les oiseaux. Une paramécie, une amibe, voire certaines moisissures (ainsi que Claude Bernard l'a noté dès 1875) sont pareillement affectées par les anesthésiques, et pour une concentration à peu près équivalente. Quel que soit l'endroit de la paramécie — ses cils, son centriole, etc. — où s'exerce l'action immobilisatrice des anesthésiques, il semble qu'il se situe nécessairement *sur son cytosquelette.* Si l'on admet que le système de contrôle d'un animal unicellulaire est effectivement son cytosquelette, il semble que l'on soit logiquement porté à reconnaître que les anesthésiques généraux agissent sur ce dernier.

Cela ne signifie pas que les animaux unicellulaires soient dotés d'une conscience. C'est là un tout autre problème. Il faut en effet probablement bien plus qu'un cytosquelette fonctionnant correctement pour susciter un état conscient. Toutefois, les arguments que j'ai présentés ici suggèrent fortement que notre état (ou nos états) de conscience *exige* un cytosquelette en action. Sans système cytosquelettique en état de marche, la conscience disparaît ; elle s'éteint instantanément dès que le fonctionnement de ce système est inhibé — et revient instantanément dès que ce fonctionnement est rétabli, à condition

bien sûr que d'autres dommages ne soient pas survenus entre-temps. On est bien sûr en droit de se demander si une paramécie — ou, par exemple, une cellule individuelle de foie humain — ne pourrait pas posséder une forme rudimentaire de conscience, mais les présentes considérations ne visent pas à répondre à cette question. Il est en tout cas probable que les détails de l'organisation neuronale du cerveau influent également sur la *forme* prise par la conscience. D'ailleurs, si cette organisation n'avait qu'un rôle mineur, notre foie susciterait autant de conscience que notre cerveau. Quoi qu'il en soit, les arguments qui précèdent suggèrent fortement que ce n'est pas la *seule* organisation neuronale de notre cerveau qui importe. L'assise cytosquelettique des neurones semble être un élément essentiel à la manifestation de la conscience.

Il est probable que la manifestation de la conscience exige moins la présence d'un cytosquelette en soi que d'une *action physique déterminante* que la biologie s'est si intelligemment arrangée pour intégrer à l'activité des microtubules. Quelle est cette action ? Selon les arguments de la première partie de ce livre, si nous désirons trouver un fondement physique à nos actes conscients, nous devons rechercher une action irréductible à toute simulation numérique. Et les arguments présentés jusqu'ici dans la deuxième partie montrent que nous devons pour cela concentrer notre attention sur la frontière entre les niveaux quantique et classique, frontière pour laquelle la physique actuelle nous prescrit d'utiliser la procédure bouche-trou **R**, alors que je prétends que nous avons besoin d'une *nouvelle* théorie physique de la **RO**. Dans le présent chapitre, après avoir tenté de localiser l'endroit du cerveau où l'action quantique pourrait s'avérer importante pour le comportement classique, nous avons apparemment été conduits à considérer que c'est par l'intermédiaire du *contrôle cytosquelettique des liaisons synaptiques* que cette interface quantique/classique exerce son influence fondamentale sur le comportement du cerveau. Essayons maintenant d'explorer cette idée un peu plus en détail.

7.7 Un modèle pour l'esprit ?

Comme nous l'avons vu à la section 7.1, il semble qu'une description totalement classique suffise à rendre compte des signaux neuronaux — il est probable en effet que ces signaux perturbent suffisamment leur environnement pour y détruire toute cohérence quantique. Si les liaisons synaptiques et leurs intensités restent fixes, l'influence du déclenchement d'un neurone sur le neurone suivant admet elle aussi un traitement classique — à l'exception des composants aléatoires présents à ce niveau. Le comportement du cerveau serait alors entièrement réductible à un calcul, en ce sens qu'il serait en principe possible d'en construire une simulation numérique. Je ne veux pas dire par là qu'une telle simulation imiterait parfaitement le comportement d'un cerveau

particulier câblé selon le protocole de cette simulation, mais — à cause de ces composants aléatoires — qu'elle serait une simulation d'un comportement *typique* de ce cerveau et donc d'un comportement typique d'un individu contrôlé par un tel cerveau (*cf.* §1.7). C'est d'ailleurs là une pure affirmation *de principe*, car la technologie actuelle ne permet nullement d'envisager une telle simulation. Je suppose également que les éléments aléatoires sont *vraiment* aléatoires. Autrement dit, j'écarte l'éventualité d'un « esprit » extérieur dualiste intervenant pour influencer ces probabilités (*cf.* §7.1).

Admettons donc (du moins provisoirement) que si ses liaisons synaptiques sont *fixes*, le cerveau se comporte comme un *ordinateur* — un ordinateur intégrant toutefois des ingrédients aléatoires. Les raisonnements de la première partie ont montré qu'il est hautement improbable qu'un tel schéma puisse jamais fournir un modèle de la compréhension consciente humaine. En revanche, si les liaisons synaptiques particulières qui définissent l'ordinateur neuronal en question sont en permanence soumises à des variations gouvernées par une action *non* calculable, rien n'interdit qu'un tel modèle étendu puisse simuler le comportement d'un cerveau conscient.

Que pourrait être cette action non calculable ? Souvenons-nous que la conscience a un caractère *global*. Si quelque 10^{11} cytosquelettes individuels fournissaient chacun une donnée d'entrée non calculable, on voit difficilement en quoi cela nous aiderait dans notre recherche. Selon les arguments de la première partie, le comportement non calculable est en effet lié à l'action de la conscience — du moins dans la mesure où *certaines* actions conscientes, en particulier la *compréhension*, sont supposées non calculables. Cela exclut donc les cytosquelettes individuels ou les microtubules individuels au sein d'un cytosquelette. Rien ne suggère qu'un cytosquelette ou un microtubule individuels « comprennent » une partie quelconque du raisonnement gödelien ! La compréhension opère sur une échelle bien plus globale ; si elle implique des cytosquelettes, elle doit alors consister en un phénomène collectif mettant en jeu un très grand nombre de ces cytosquelettes.

Rappelons l'idée de Fröhlich : l'existence — peut-être sous une forme analogue aux condensats de Bose-Einstein — d'effets quantiques collectifs à grande échelle dans des systèmes biologiques est une possibilité réelle, même à l'intérieur d'un objet aussi « chaud » que le cerveau (*cf.* aussi Marshall 1989). Dans notre contexte, cela se traduit par l'existence d'une cohérence quantique qui non seulement est associée aux microtubules individuels, mais en outre s'étend d'un microtubule à son voisin. Ainsi, cette cohérence quantique englobe non seulement toute la longueur d'un microtubule (et, nous l'avons vu, les microtubules peuvent avoir des longueurs considérables), mais aussi un très grand nombre de microtubules — si ce n'est leur totalité — du cytosquelette d'un neurone. En outre, cette cohérence quantique doit franchir la barrière synaptique séparant deux neurones. On ne peut en effet véritablement parler de globalité si cet état cohérent n'implique que des cellules individuelles ! Dans cette description, l'unité de l'esprit ne peut survenir que s'il existe une forme de cohérence quantique s'étendant au minimum à une partie appréciable du cerveau.

Accomplir cela à l'aide de moyens biologiques constituerait, de la part de la nature, une prouesse remarquable — presque inconcevable. Pourtant, je pense que tout, en particulier l'existence de notre esprit, montre qu'elle a réussi cet exploit. Il nous reste beaucoup à apprendre des systèmes biologiques et de leur magie. La biologie contient bien des processus qui surpassent, et de loin, ce que l'on pourrait réaliser avec les techniques actuelles de la physique. (Pensez par exemple à la toile très sophistiquée que tisse une minuscule araignée.) Rappelons en outre que les expériences conduites par Aspect et d'autres chercheurs ont (*physiquement*) montré l'existence de certains effets de cohérence quantique se manifestant sur des distances de plusieurs mètres — les emmêlements EPR associés à des paires de photons (*cf.* §5.4). Les difficultés techniques que soulève la détection de tels effets quantiques à grande distance ne nous autorisent pas à exclure l'éventualité que la nature ait trouvé des moyens biologiques pour accomplir des exploits encore plus admirables. Son « ingéniosité » n'a pas fini de nous surprendre.

Toutefois, les arguments que j'ai présentés exigent plus qu'une simple cohérence quantique à grande échelle. Ils supposent que le système biologique qu'est notre cerveau s'est arrangé pour exploiter les particularités d'une physique encore inconnue des physiciens humains ! Cette physique est la « théorie de la réduction objective » (**RO**), située à cheval sur les niveaux classique et quantique, et qui, ainsi que je le prétends, substitue au procédé bouche-trou **R** une procédure physique extrêmement subtile et non calculable (mais incontestablement mathématique).

(Le fait que les physiciens humains ignorent encore largement ce qu'est cette théorie ne prouve bien sûr pas que la nature ne l'ait pas utilisée. Elle utilisait les principes de la dynamique newtonienne bien avant Newton, les phénomènes électromagnétiques bien avant Maxwell, et la mécanique quantique bien avant Planck, Einstein, Bohr, Heisenberg, Schrödinger et Dirac. Elle l'a fait pendant des milliers de millions d'années ! Mais l'arrogance de notre époque actuelle nous incite à penser que nous connaissons tous les principes fondamentaux sous-jacents à toutes les subtilités des processus biologiques.)

Lorsqu'un organisme a la chance de tomber sur un tel processus, il peut profiter des avantages que cela lui apporte. La nature lui sourit alors, à lui et à ses descendants, et autorise la transmission de ce processus de génération en génération — à travers le puissant mécanisme de la sélection naturelle. Les premières créatures cellulaires eucaryotes ont probablement jugé qu'elles tiraient un grand avantage de la présence en leur sein de microtubules primitifs. Selon les idées que j'expose ici, cela donna naissance à une sorte de supériorité organisationnelle qui peut-être leur permit d'adopter un comportement téléonomique rudimentaire qui à son tour les aida à mieux survivre aux dépens de leurs compétiteurs. Il ne convient sans doute pas à ce stade de qualifier d'« esprit » cette supériorité ; je pense néanmoins qu'elle est apparue grâce à une interaction subtile entre des processus classiques et quantiques. La subtilité de cette interaction dut son existence même à la procédure physique sophistiquée **RO** — dont les détails nous sont encore inconnus — qui, dans des contextes organisationnels moins subtils, se manifeste sous la forme du processus quantique

grossier **R** que nous adoptons aujourd'hui. Les lointains descendants de ces créatures cellulaires — nos paramécies, nos amibes, nos fourmis, nos arbres, nos grenouilles, nos boutons d'or, et nous-mêmes êtres humains — ont conservé les avantages que cette procédure sophistiquée a accordés à ces anciennes créatures cellulaires et les ont détournés pour les utiliser à toutes sortes de fins différentes. Ce n'est qu'une fois intégrée dans un système nerveux hautement développé que cette procédure a pu réaliser ses énormes potentialités — donnant alors naissance à ce que nous appelons aujourd'hui l'« esprit ».

Acceptons donc que tous les microtubules des cytosquelettes d'une vaste famille de neurones de notre cerveau puissent participer à un phénomène de cohérence quantique globale — ou du moins qu'il y ait un emmêlement quantique suffisant entre les états de différents microtubules appartenant au cerveau —, de sorte que l'action collective de ces microtubules *n'admette pas* de description *classique*. Par exemple, l'intérieur de ces microtubules pourrait être le siège d'« oscillations quantiques » complexes (*e.g.* celles de del Giudice *et al.* 1983 ou Jibu *et al.* 1994), et l'isolant que constituent les tubes suffirait à empêcher la destruction totale de la cohérence quantique. Il est alors tentant de supposer que les calculs de type automates cellulaires qui, selon Hameroff et ses collègues, surviennent *le long* des microtubules sont couplés à ces hypothétiques oscillations quantiques.

Notons que la fréquence envisagée par Fröhlich pour ses oscillations quantiques collectives — qui est confortée par les observations de Grundler et Keilmann (1983) et se situe autour de 5×10^{10} Hz (5×10^{10} oscillations par seconde) — correspond à celle-là même adoptée par Hameroff et ses collègues comme « temps de basculement » des tubulines dans leurs automates cellulaires microtubulaires. Ainsi, si le mécanisme de Fröhlich opère effectivement au sein des microtubules, cela indique fortement l'existence d'un couplage entre ces deux types d'activité*.

Toutefois, si ce couplage s'avère trop fort, il sera impossible de maintenir la nature quantique des oscillations internes sans traiter de façon quantique les calculs effectués le long des microtubules eux-mêmes. Les microtubules seront alors le siège de *calculs quantiques* (*cf.* §7.3) ! Voyons si cette éventualité est sérieusement envisageable.

La difficulté est que cela semble exiger que les changements de conformation du dimère ne perturbent pas de manière significative le matériau ambiant extérieur. Il se trouve cependant, nous l'avons vu, qu'un microtubule semble être entouré d'une région contenant de l'eau *ordonnée*, à l'exclusion de toute

* L'existence d'un lien direct entre ces fréquences relativement élevées et le phénomène plus familier d'« ondes cérébrales » (tel le rythme α à 8-12 Hz) est cependant moins évident. Si l'on peut admettre que ces plus basses fréquences correspondent à des « fréquences de battement », aucun lien cependant n'a été établi. Signalons aussi les observations récentes d'oscillations à 35-75 Hz apparemment associées aux zones cérébrales mises en jeu par l'attention consciente. Ces oscillations semblent avoir de curieuses propriétés non locales. (Voir Eckhorn *et al.* 1988, Gray et Singer 1989, Crick et Koch 1990, 1992, et Crick 1994.)

autre matière (*cf.* Hameroff 1987, p. 172), région qui pourrait alors constituer une sorte d'écran quantique. D'un autre côté, les ponts PAM (*cf.* §7.4), qui émergent des microtubules et dont certains ont pour rôle de transporter d'autres matériaux, semblent être influencés par le déplacement des signaux le long des tubes (*cf.* Hameroff, p. 122). Ce dernier fait donne à penser que les « calculs » effectués par le microtubule perturbent suffisamment l'environnement pour devoir être traités en termes classiques. Certes, par rapport au critère **RO** présenté à la section 6.12, ces perturbations restent relativement faibles en termes de déplacement de masse. Cependant elles devraient, pour que le système demeure au niveau quantique, ne pas diffuser longtemps dans la cellule, puis à l'extérieur, au-delà de la membrane cellulaire. Selon moi, tant la situation physique réelle que la manière dont on doit appliquer le critère **RO** laissent place à suffisamment d'incertitudes pour que l'on ne puisse être certain de l'adéquation, à ce stade, d'une description entièrement classique.

Supposons cependant, pour les besoins du raisonnement, que les calculs effectués par les microtubules doivent réellement être traités en termes essentiellement classiques — dans le sens où l'on considère que les superpositions quantiques des divers calculs ne jouent pas un rôle significatif. Supposons également que les tubes soient le siège d'oscillations purement quantiques et qu'existe un subtil couplage entre les aspects internes quantiques et les aspects externes classiques de chacun de ces tubes. Ce serait alors au niveau de ce couplage qu'interviendraient de façon très significative les *détails* de la nouvelle théorie **RO** que nous recherchons. Les « oscillations » internes quantiques influenceraient les calculs externes, mais cela n'est guère gênant eu égard aux mécanismes que l'on croit responsables du comportement de type automate cellulaire affiché par les microtubules — à savoir, aux forces de van der Waals peu intenses agissant entre tubulines voisines.

On serait donc en présence d'un état quantique global couplant de manière cohérente les activités se déroulant à l'intérieur des tubes, et cela pour tous les microtubules de vastes régions du cerveau. Cet état (qui peut ne pas être un « état quantique » au sens du formalisme quantique standard) influencerait les calculs se déroulant le long des microtubules, et cette influence tiendrait délicatement et scrupuleusement compte de la physique **RO** non calculable dont j'affirme avec force la nécessité. Les « calculs » associés aux variations de conformation des tubulines contrôleraient la manière dont les tubes transportent les matériaux à leur surface externe (Fig. 7.13) et en définitive influeraient sur l'intensité des liaisons synaptiques aux terminaisons pré- et post-synaptiques. Ainsi, une petite partie de l'organisation quantique cohérente présente *au sein* des microtubules serait détournée pour influencer les variations des liaisons synaptiques.

Une telle image peut donner lieu à diverses spéculations. Quel est par exemple le rôle de la mystérieuse non-localité des effets de type EPR associés aux emmêlements quantiques ? Quel est celui de la contrafactualité ? Peut-être que l'ordinateur neuronal est réglé pour effectuer un calcul qu'en fait il n'exécute pas, mais (comme dans le problème d'Elitzur-Vaidman) que le simple fait qu'il *pourrait* avoir exécuté ce calcul produit un effet différent de ce qu'il serait

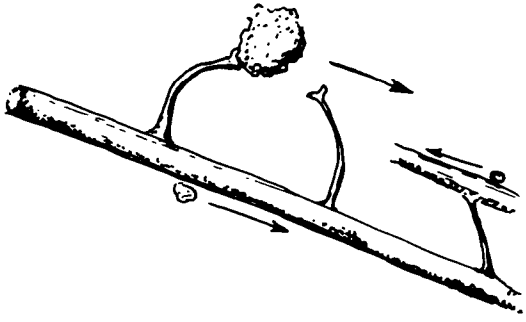


Figure 7.13. Les PAM transportent également de grosses molécules ; d'autres molécules se déplacent directement le long des microtubules.

s'il ne pouvait l'effectuer. Ainsi, tel qu'il est à chaque instant, le « câblage » classique de cet ordinateur neuronal influencerait l'état interne du cytosquelette, même si les déclenchements neuronaux qui activeraient cet ordinateur ne surviennent pas réellement. On pourrait envisager des analogies de ce type dans nombre de nos activités mentales familières — mais j'ai le sentiment qu'il est préférable ici de ne pas poursuivre plus avant ce genre de spéculations !

En vertu des idées proposées dans ce livre, la conscience serait une manifestation de l'état interne du cytosquelette — état quantique emmêlé — et de son implication dans l'interaction (**RO**) entre les niveaux d'activité classique et quantique. Le système de neurones, interconnecté classiquement à la manière d'un ordinateur, serait en permanence influencé par cette activité cytosquelettique, et cette influence serait la manifestation de ce que nous appelons le « libre arbitre ». Les neurones opèrent peut-être surtout comme des *dispositifs d'amplification* transmettant l'action cytosquelettique à petite échelle à quelque chose qui peut influencer d'autres organes du corps — par exemple, les muscles. Ainsi, le niveau neuronal auquel se situe la description conventionnelle du cerveau et de l'esprit ne serait qu'une simple *ombre* du niveau plus profond de l'action cytosquelettique — et c'est à ce niveau plus profond que nous devons rechercher les fondements physiques de l'*esprit* !

Si cette vision contient certes une part de spéculation, elle n'est pas pour autant incompatible avec nos connaissances scientifiques actuelles. Nous avons vu au chapitre précédent qu'il existe de très sérieuses raisons — liées à des considérations sur la physique actuelle — de penser que les idées physiques actuelles nécessitent une modification qui mettrait en lumière des effets nouveaux situés justement au niveau des processus microtubulaires, voire à l'interface cytosquelette/neurone. Les arguments de la première partie ont montré que si nous voulons trouver un cadre physique au phénomène de la conscience, nous devons le rechercher dans une action physique non calculable, et j'ai affirmé dans la deuxième partie que le seul moyen plausible d'y parvenir consiste à substituer une procédure pertinente **RO** à la procédure de réduction quantique **R**. Voyons maintenant s'il existe des raisons purement *physiques* de

penser que **RO** est effectivement irréductible au calcul. En vertu des suggestions émises à la section 6.12, ces raisons existent bel et bien.

7.8 La non-calculabilité en gravitation quantique : 1

L'une des exigences cruciales de la discussion précédente est que la nouvelle physique appelée à remplacer la procédure probabiliste **R** utilisée en théorie quantique standard soit une physique non calculable. À la section 6.10, j'ai montré que cette nouvelle physique (**RO**) devra associer les principes de la théorie quantique à ceux de la relativité générale d'Einstein — qu'elle devra être une théorie quantique et *gravitationnelle*. Existe-t-il un indice montrant que la non-calculabilité sera une caractéristique fondamentale de la théorie, quelle qu'elle soit, qui parviendra à unifier (et à modifier de manière appropriée) la théorie quantique et la relativité générale ?

Dans leur approche de la gravitation quantique, Robert Geroch et James Hartle (1986) ont rencontré un problème algorithmiquement insoluble, à savoir le *problème de l'équivalence topologique pour les variétés à quatre dimensions*. Schématiquement parlant, ce problème consiste à déterminer quand deux espaces quadridimensionnels sont topologiquement identiques (*i.e.* quand on peut déformer de manière continue — sans le déchirer ni le coller — l'un de ces deux espaces pour le faire coïncider avec l'autre). La figure 7.14 illustre ce problème dans le cas à deux dimensions : on voit que, contrairement à celle d'une balle, la surface d'une tasse à thé est topologiquement équivalente à celle d'un anneau. Si, à deux dimensions, le problème de l'équivalence topologique est algorithmiquement soluble, A. A. Markov a montré en 1958 qu'aucun algorithme ne permet de résoudre ce problème dans le cas à *quatre* dimensions. Plus précisément, Markov a démontré que si un tel algorithme existait, on pourrait le convertir en un autre qui résoudrait le *problème de l'arrêt* (qui pourrait décider si l'action d'une machine de Turing s'arrête ou non). Puisqu'un tel algorithme n'existe pas — nous l'avons vu à la section 2.5 —, il s'ensuit que le problème de l'équivalence topologique pour les variétés à quatre dimensions est insoluble algorithmiquement.

Il existe de nombreuses autres familles de problèmes mathématiques insolubles algorithmiquement. Deux d'entre eux, le dixième problème de Hilbert et le problème du pavage, ont été examinés à la section 1.9. Pour un autre exemple, le problème du mot (pour les semi-groupes), voir EOLP p. 140-142.

J'insiste sur le fait que l'expression « insoluble algorithmiquement » ne signifie pas que tout problème de la famille considérée soit en principe individuellement insoluble. Elle signifie simplement qu'il n'existe aucun moyen (algorithmique) systématique de résoudre tous les problèmes de cette famille. Il peut s'avérer que dans un cas particulier, on puisse parvenir à une solution en recourant à l'intuition et à l'ingéniosité humaines, voire en s'aidant d'un

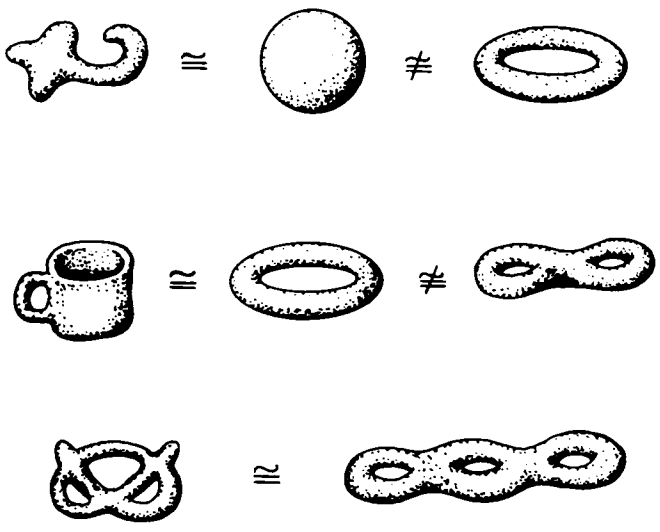


Figure 7.14. Les surfaces fermées à deux dimensions admettent une classification algorithmique (schématiquement, cette classification s'obtient en comptant le nombre de « poignées »). En revanche, les « surfaces » fermées à quatre dimensions *n'admettent pas* de classification algorithmique.

ordinateur. Il *peut* aussi advenir que certains membres de la famille soient inaccessibles au jugement humain (ou même au jugement humain aidé par l'ordinateur). Il semble que l'on ne connaisse pas grand-chose sur ce point, de sorte que c'est à chacun de se forger son opinion. Toutefois, le raisonnement gödelien de la section 2.5 et les arguments du chapitre 3 montrent que les problèmes d'une telle famille, même *accessibles* à la compréhension et à l'intuition humaines (aidées d'un ordinateur, au besoin) forment une classe qui est elle-même inaccessible algorithmiquement. (Dans le cas du problème de l'arrêt, par exemple, la section 2.5 montre une famille de calculs dont un être humain peut affirmer qu'ils ne s'arrêtent pas. Cependant, pour cette famille de calculs, aucun algorithme A que l'on sait sûr ne peut vérifier ce fait — on peut à partir de là reprendre les arguments du chapitre 3.)

Dans l'approche de la gravitation quantique adoptée par Geroch et Hartle, le problème de l'équivalence topologique pour les variétés à quatre dimensions intervient parce qu'en vertu des règles *standard* de la théorie quantique, un état quantique gravitationnel serait une superposition de toutes les géométries possibles — en l'occurrence, des géométries *spatio-temporelles*, qui sont des entités à quatre dimensions — affectées de facteurs de pondération complexes. Pour spécifier cette superposition de manière unique (*i.e.* sans « redondances »), il faut savoir quand on peut considérer que deux de ces espaces-temps sont différents et quand on peut considérer qu'ils sont identiques. Cette décision est donc un problème d'équivalence topologique.

On peut se demander si l'éventuel succès d'une approche de la gravitation quantique — telle celle suivie par Geroch et Hartle — signifierait que l'évolution d'un système physique contient un élément fondamentalement non calculable. Selon moi, la réponse à cette question est loin d'être claire. Il ne m'apparaît même pas clairement que l'insolubilité algorithmique de l'équivalence topologique entraîne nécessairement celle du problème plus complet de l'équivalence *géométrique*. En outre, je ne vois pas comment (ni si) une telle approche pourrait être reliée au concept **RO** que je défends ici et selon lequel la structure de la théorie quantique subirait une modification lorsque entrent en jeu des effets gravitationnels. Toutefois, l'approche de Geroch et Hartle indique clairement que la non-calculabilité pourrait jouer un rôle réel dans toute théorie gravitationnelle quantique qui s'avérera physiquement correcte.

7.9 Machines-oracles et lois physiques

On peut néanmoins se poser une autre question : supposons que la théorie gravitationnelle quantique recherchée s'avère effectivement non calculable, dans le sens précis où elle permettrait de construire un dispositif physique capable de résoudre le problème de l'arrêt. Cela suffirait-il pour résoudre tous les problèmes découlant des considérations sur l'argument de Gödel-Turing mentionnées dans la première partie ? Curieusement, la réponse à cette question est *non* !

Voyons pourquoi l'aptitude à résoudre le problème de l'arrêt ne nous aide en rien. Méditant sur ce problème, Turing introduisit en 1939 un concept important qu'il baptisa *oracle*. Un oracle est une chose (vraisemblablement, dans l'esprit de Turing, une chose fictive qui n'est pas nécessairement physiquement constructible) qui résoudrait le problème de l'arrêt. Ainsi, si l'on présente à l'oracle un couple d'entiers naturels q et n, il répond, au bout d'un temps fini, soit **OUI** soit **NON** selon que le calcul $C_q(n)$ s'arrête ou non (*cf.* §2.5). Si l'argumentation de la section 2.5 démontre le résultat de Turing, à savoir qu'un tel oracle ne fonctionnerait pas de manière entièrement algorithmique, rien ne dit qu'on ne pourrait en construire un physiquement. Pour répondre à cette question, il faudrait savoir si les lois physiques sont calculables — ce qui est en définitive le thème central de toute cette deuxième partie. Je dois souligner également que, pour autant que je puisse en juger, le point de vue que je défends dans ce livre n'implique pas que l'on puisse physiquement construire un oracle. Nous l'avons dit, les problèmes de l'arrêt ne sont pas nécessairement tous accessibles à la compréhension et à l'intuition humaines, de sorte que nous ne pouvons conclure non plus qu'ils soient accessibles à un dispositif physiquement constructible.

Turing, dans son raisonnement, a envisagé une notion de calculabilité modifiée autorisant l'invocation d'un oracle à chaque fois qu'on le souhaite.

Cela permet de définir une *machine-oracle* (exécutant un *algorithme-oracle*) qui serait une machine de Turing ordinaire à laquelle on aurait adjoint, en plus de ses opérations numériques ordinaires, l'opération : « Invoque l'oracle et demande-lui si $C_q(n)$ s'arrête ou non ; lorsque tu auras reçu sa réponse, continue tes calculs en utilisant cette réponse. » L'oracle peut être invoqué autant de fois qu'on le désire. Notez qu'une machine-oracle est tout aussi *déterministe* qu'une machine de Turing ordinaire, ce qui illustre le fait que la calculabilité est totalement différente du déterminisme. Il serait tout aussi possible, en principe, d'avoir un univers fonctionnant de manière déterministe à l'instar d'une machine-oracle que d'avoir un univers fonctionnant de manière déterministe à l'instar d'une machine de Turing. (Les « modèles d'univers » décrits à la section 1.9 et p. 182 de EOLP sont, en fait, des univers de type machines-oracles.)

Se pourrait-il que *notre* Univers fonctionne comme une machine-oracle ? Curieusement, les arguments de la première partie de ce livre qui réfutaient tout modèle de compréhension mathématique de type machine de Turing réfutent pareillement — presque sans changement — tout modèle de type machine-oracle. Il suffit en effet de reprendre la discussion donnée à la section 2.5 en considérant que « $C_q(n)$ » signifie maintenant la « $q^{\text{ième}}$ machine-oracle appliquée à l'entier naturel n ». Notons-la $C_q'(n)$. Comme les machines de Turing, les machines-oracles admettent elles aussi une classification (algorithmique). La seule différence au niveau de leurs spécifications est que l'on doit noter les étapes où l'oracle intervient, mais cela ne pose aucun problème nouveau. On remplace alors l'*algorithme* $A(q, n)$ de la section 2.5 par un *algorithme-oracle* $A'(q, n)$ censé représenter la totalité des moyens dont disposent la compréhension et l'intuition humaines pour décider avec certitude que l'opération-oracle $C_q'(n)$ ne s'arrête pas. En reprenant exactement les arguments de la section 2.5, on aboutit à la conclusion :

$\mathcal{G}'$ Ce n'est pas en utilisant un algorithme-oracle qu'ils savent sûr que les mathématiciens humains établissent la vérité mathématique.

Nous en concluons qu'une physique qui fonctionne comme une machine-oracle ne résoudra pas, elle non plus, nos problèmes.

De fait, on peut reproduire tout le raisonnement en l'appliquant à des « machines-oracles du deuxième degré » invoquant, lorsque c'est nécessaire, un *oracle du deuxième degré* — qui dit si une machine-oracle ordinaire s'arrête ou non. Comme à l'instant, on parvient alors à la conclusion :

$\mathcal{G}''$ Ce n'est pas en utilisant un algorithme-oracle du deuxième degré qu'ils savent sûr que les mathématiciens humains établissent la vérité mathématique.

Il est clair que l'on peut itérer indéfiniment ce processus, à l'instar de la gödelisation utilisée lors de la réponse à **Q19**. À tout ordinal α récursif (calculable) est alors associée une machine-oracle de degré α, et nous sommes amenés à conclure :

$\mathcal{G}^{\alpha}$ Quel que soit l'ordinal calculable α, ce n'est pas en utilisant un algorithme-oracle de degré α qu'ils savent sûr que les mathématiciens humains établissent la vérité mathématique.

Cette dernière conclusion est assez alarmante, car elle suggère que nous devons rechercher une théorie physique non calculable qui soit irréductible à toute machine-oracle de degré calculable (voire davantage).

Certains lecteurs vont penser que mes arguments viennent de perdre leur dernier vestige de crédibilité ! Je ne reprocherai à personne d'avoir une telle pensée, mais ce n'est pas une excuse pour ne pas examiner tous les raisonnements que j'ai détaillés. En particulier, il faut reprendre tous ceux des chapitres 2 et 3 en remplaçant les machines de Turing par des machines-oracles de degré α. Je pense que tous ces raisonnements en sortiront indemnes, mais j'avoue que je ne me sens pas le courage de tout réexposer avec les machines-oracles. Il faut aussi remarquer, d'un autre côté, que la compréhension mathématique humaine n'a pas à être, en principe, aussi puissante que *toutes* les machines-oracles. Nous l'avons vu en effet, la conclusion $\mathcal{G}$ *n'entraîne pas* nécessairement que l'intuition humaine soit, en principe, suffisamment puissante pour résoudre tous les problèmes de l'arrêt. Ainsi, nous n'avons pas à rechercher des lois physiques qui soient irréductibles, en principe, à toute machine-oracle de degré calculable (ni même à une machine-oracle du premier degré). Il suffit de rechercher une physique qui ne soit équivalente à *aucune* machine-oracle particulière (y compris les machines de degré *zéro* que sont les machines de Turing). Peut-être les lois physiques débouchent-elles sur quelque chose de simplement *différent.*

7.10 La non-calculabilité en gravitation quantique : 2

Revenons au problème de la gravitation quantique. Soulignons qu'il n'existe pas actuellement de théorie reconnue — voire de théorie candidate acceptable. Il existe cependant de nombreuses et passionnantes conjectures[8]. Celle que je voudrais mentionner ici envisage, comme l'approche de Geroch et Hartle, des superpositions quantiques de différents *espaces-temps.* (Nombre d'approches n'envisagent que des superpositions de géométries spatiales tridimensionnelles, ce qui est légèrement différent.) L'idée, due à David Deutsch[9], est que l'on doit superposer, outre les géométries spatio-temporelles « raisonnables » dans lesquelles le *temps* se comporte de manière « sensée », des espaces-temps « déraisonnables » contenant des *courbes fermées du genre temps.* La figure 7.15 donne un exemple d'un tel espace-temps. Une *courbe du genre temps* représente une histoire possible d'une particule (classique), « du genre temps » signifiant que la courbe reste toujours à l'intérieur du cône de lumière local attaché à chacun de ses points, de sorte que la vitesse de la particule ne

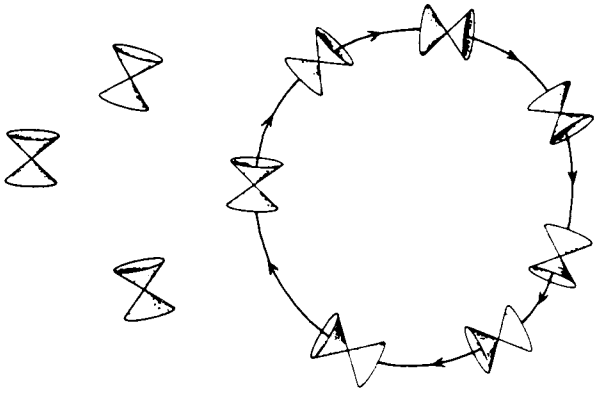


Figure 7.15. Si les cônes de lumière sont suffisamment inclinés, les courbes du genre temps peuvent se refermer sur elles-mêmes.

dépasse pas la vitesse absolue locale — comme l'exige la théorie de la relativité (*cf.* §4.4). Une courbe *fermée* du genre temps pourrait être la ligne d'univers d'un « observateur », *i.e.* la courbe qui décrit, dans l'espace-temps, l'histoire de son propre corps. Au bout d'un temps fini mesuré par une horloge qui lui serait attachée, cet observateur se retrouverait dans son propre passé (voyage dans le temps !). Il pourrait alors (en supposant qu'il ait son « libre arbitre ») accomplir des choses qu'il n'avait pas accomplies auparavant, ce qui conduit à une contradiction. (Habituellement, dans ce genre de discussion, on lui fait tuer son grand-père avant d'être né — ou d'autres choses tout aussi alarmantes.)

Ce genre d'arguments rend peu crédibles les modèles de l'Univers réel (classique) fondés sur des espaces-temps contenant des courbes fermées. (Curieusement, ce fut Gödel qui, en 1949, introduisit ce type de modèle. Selon lui, les aspects paradoxaux de tels espaces-temps ne justifiaient aucunement qu'on les exclût en tant que modèles cosmologiques. Pour diverses raisons, on est toutefois aujourd'hui moins permissif à leur égard — *cf.* cependant Thorne (1994). Il aurait été intéressant de connaître la réaction de Gödel face à l'utilisation qui sera faite dans un instant de ces espaces-temps !) S'il semble effectivement raisonnable de considérer que les géométries spatio-temporelles à courbes du genre temps fermées ne représentent pas des descriptions de l'Univers *classique*, il ne faut toutefois pas les exclure en tant qu'occurrences potentielles au sein d'une *superposition quantique*. C'est là en fait le point de vue de Deutsch. Bien que les contributions de ces géométries au vecteur d'état total puissent s'avérer infimes, leur présence potentielle (selon Deutsch) a des conséquences surprenantes. Il semble en effet que l'exécution d'un calcul quantique dans une telle situation rende possibles des opérations *non calculables* ! Cela résulte du fait que dans les géométries spatio-temporelles contenant des courbes du genre temps fermées, une machine de Turing peut prendre comme donnée d'entrée le résultat qu'elle vient de trouver et ainsi

tourner indéfiniment, si nécessaire, de sorte que la réponse à la question : « Ce calcul s'arrête-t-il ? » a une influence réelle sur le résultat final du calcul quantique. Selon Deutsch, cette approche de la gravitation quantique autorise l'existence de machines-oracles quantiques. À mon avis, son raisonnement devrait s'appliquer aux machines-oracles de n'importe quel degré.

Nombre de lecteurs auront bien entendu le sentiment que tout cela est à prendre avec des pincettes. Rien ne dit que l'approche de Deutsch débouchera effectivement sur une théorie cohérente (voire plausible) de la gravitation quantique. Néanmoins, ces idées présentent une indéniable logique et sont à la fois intéressantes et stimulantes — je crois en outre que nombre d'entre elles s'intégreront à la théorie définitive de la gravitation quantique. À mon avis, ainsi que l'ont notamment souligné les sections 6.10 et 6.12, l'unification de la relativité générale et de la théorie quantique suppose que l'on modifie les lois de cette dernière (au niveau de la **RO**). Mais je considère que la présence de la non-calculabilité — même au degré apparemment exigé pour $\mathcal{G}^{\alpha}$ — dans l'approche de la gravitation quantique de Deutsch conforte considérablement la possibilité d'une action non calculable.

Remarquons enfin que c'est précisément l'inclinaison potentielle des cônes de lumière en relativité générale (*cf.* §4.4) qui donne les effets non calculables prédits par Deutsch. Il suffit en effet que ces cônes *puissent* s'incliner, même d'un angle infime comme il advient dans les circonstances ordinaires, pour qu'ils puissent le faire *potentiellement* d'un angle tel que les lignes d'univers du genre temps se referment sur elles-mêmes. Il suffit que cette possibilité soit contrafactuelle — au sens quantique — pour avoir un effet *réel* !

7.11 Temps et perceptions conscientes

Revenons au problème de la conscience. C'est effectivement le rôle particulier joué par la conscience dans la perception de la vérité mathématique qui nous a engagés sur la voie menant à l'étrange contrée dans laquelle nous nous trouvons maintenant. Mais la conscience est manifestement loin de se réduire à la perception des mathématiques. Nous n'avons suivi cette voie que parce qu'elle semblait conduire quelque part. Nul doute que nombre de lecteurs n'aimeront pas tellement ce « quelque part » où nous sommes plus ou moins arrivés. Toutefois, si nous regardons le chemin parcouru, nous constatons que certains de nos vieux problèmes apparaissent sous un jour nouveau.

L'une des caractéristiques les plus immédiates et les plus frappantes de la perception consciente est l'*écoulement du temps.* Cet écoulement nous est si familier qu'on est stupéfait de constater à quel point nos théories merveilleusement précises sur le comportement du monde physique ont, à ce jour, peu à dire sur lui. Pire, ce que *disent* nos meilleures théories est presque en contradiction flagrante avec la perception que nous avons du temps.

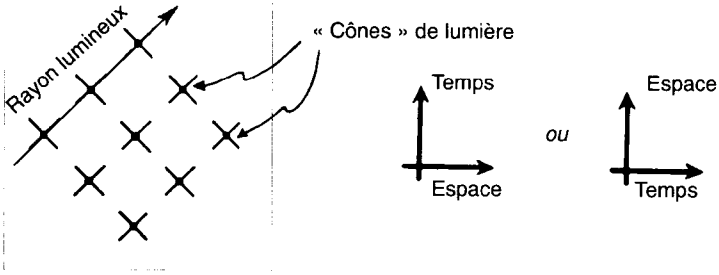


Figure 7.16. Dans un espace-temps à deux dimensions, rien ne distingue le temps de l'espace — pourtant, personne ne dirait que l'espace « s'écoule » !

Selon la relativité générale, le « temps » n'est qu'un choix parmi d'autres d'une coordonnée destinée à localiser un événement spatio-temporel. Rien dans les descriptions spatio-temporelles des physiciens n'associe le « temps » à un « écoulement ». De fait, les physiciens travaillent très souvent sur des modèles d'espace-temps ne contenant qu'*une* dimension spatiale et une dimension temporelle ; et dans ces espaces-temps à deux dimensions, il n'y a aucun moyen de distinguer la dimension spatiale de la dimension temporelle (Fig. 7.16). Pourtant, personne ne considère que l'*espace* « s'écoule » ! Il est vrai que lorsque, dans les problèmes physiques, on cherche à déterminer l'état futur d'un système à partir de son état présent, on considère souvent des évolutions temporelles (*cf.* §4.2). Mais cette procédure n'est pas du tout nécessaire. Nous opèrons habituellement ainsi *parce que* nous voulons modéliser mathématiquement notre vision du monde en termes d'un « écoulement » temporel que nous semblons percevoir — et parce que nous voulons prédire l'avenir[10]. Ce sont les expériences que nous avons conscience de vivre dans la vie quotidienne qui nous incitent à formuler (fréquemment, mais pas invariablement) nos modèles numériques du monde en termes d'évolutions temporelles, tandis que les lois physiques ne contiennent en elles-mêmes rien qui oblige à de telles formulations.

En fait, c'est *uniquement* le phénomène de la conscience qui nous oblige à penser en termes d'« écoulement » temporel. La relativité ne connaît qu'un espace-temps quadridimensionnel « statique », sans aucun « écoulement ». L'espace-temps est simplement *là*, et le temps ne « s'écoule » pas plus que l'espace. Seule la conscience semble exiger un écoulement du temps. Ne soyons donc pas surpris si la relation entre temps et conscience présente d'autres particularités étranges.

Gardons-nous par conséquent d'associer le phénomène de la conscience — et son « écoulement » temporel apparent — à l'utilisation par les physiciens d'un paramètre numérique réel t désignant ce que l'on appelle une « coordonnée temporelle ». Premièrement, la relativité enseigne que si on veut l'appliquer à l'espace-temps tout entier, ce paramètre t n'est pas défini de manière unique. Il existe de nombreux choix possibles et incompatibles, et

rien ne permet d'en privilégier un au détriment des autres. Deuxièmement, il est clair que le concept précis de « nombre réel » n'est pas totalement pertinent au niveau de notre perception consciente de l'écoulement du temps, ne serait-ce que parce que les très petites échelles temporelles — par exemple, un centième de seconde — échappent à nos aptitudes sensorielles, tandis que les échelles de temps des physiciens vont allègrement jusqu'à quelque 10^{-25} seconde (ainsi que le démontre la précision de l'électrodynamique quantique, cette théorie quantique des champs électromagnétiques en interaction avec des électrons et d'autres particules chargées), voire même jusqu'au temps de Planck, 10^{-43} seconde. En outre, la représentation mathématique du temps par des nombres réels exige qu'il n'y ait *aucune* limite à la petitesse pour laquelle cette représentation reste valide, que cette représentation soit ou non physiquement pertinente à toutes les échelles.

Peut-on préciser davantage la relation existant entre l'expérience consciente du temps et le paramètre t que les physiciens utilisent pour le représenter ? Existe-t-il un moyen expérimental de savoir, à l'aide de ce paramètre physique, « quand » une expérience subjective a « réellement » lieu ? Cela a-t-il même un *sens* d'affirmer qu'un événement conscient est survenu à un instant particulier ? Certaines expériences ont effectivement eu pour but de répondre à ces questions, mais leurs résultats sont très déconcertants et leurs conséquences sont presque paradoxales. Quelques-unes de ces expériences ont été décrites dans EOLP, p. 478-482, mais elles méritent d'être réexaminées ici.

Au milieu des années soixante-dix, H. H. Kornhuber et ses collègues (*cf.* Deeke *et al.* 1976) enregistrèrent, à l'aide d'électroencéphalogrammes (EEG), des signaux électriques en divers points du cerveau de plusieurs sujets humains afin de tenter de chronométrer l'activité cérébrale éventuellement associée à un acte déclenché par *libre arbitre* (l'aspect *actif* de la conscience). On demanda aux sujets de plier brusquement l'index d'une main, à divers instants entièrement choisis par eux. On espérait ainsi chronométrer l'activité cérébrale associée à ce mouvement « volontaire » de l'index. Pour obtenir des résultats significatifs, il fallut faire la moyenne des enregistrements EEG sur plusieurs mouvements. On découvrit alors un résultat surprenant : le potentiel électrique enregistré s'accroît progressivement pendant environ une seconde, une seconde et demie, *avant* la flexion de l'index proprement dite. Cela signifie-t-il qu'un acte conscient de la volonté met au moins une seconde pour être effectué ? Si l'on admet que les sujets étaient conscients, la décision de plier l'index aurait dû survenir immédiatement avant la flexion effective, et non une bonne seconde avant. (Rappelons qu'un temps de réaction « préprogrammé », en réponse à un signal extérieur (un « réflexe conditionné »), est bien plus bref — environ un cinquième de seconde.)

Cette expérience montre apparemment que *soit* (i) l'acte conscient décidé « librement » est une pure illusion dans la mesure où il est, en un certain sens, préprogrammé dans l'activité inconsciente antérieure du cerveau ; *soit* (ii) la volonté a peut-être un rôle de « dernière minute », de sorte qu'elle peut parfois (mais pas toujours) inverser la décision qui avait inconsciemment mûri durant une seconde, une seconde et demie ; *soit* (iii) le sujet, en fait, veut consciem-

ment plier l'index à un instant antérieur à une seconde, une seconde et demie avant que la flexion ne survienne, mais perçoit à tort, et à chaque fois, que l'acte conscient intervient bien plus tard, juste avant que l'index ne soit effectivement plié.

Plus récemment, Benjamin Libet et ses collaborateurs ont reproduit cette expérience en y introduisant des perfectionnements destinés à chronométrer plus directement l'acte effectif consistant à vouloir fléchir l'index. Pour cela, ils demandèrent aux sujets de noter la position de l'aiguille d'une horloge à l'instant de leur prise de décision (*cf.* Libet 1990, 1992). Si les conclusions de Libet semblent confirmer les résultats de Kornhuber, elles tendent à infirmer la version (iii). Libet lui-même semble privilégier l'option (ii).

En 1979, Libet et Feinstein ont chronométré, lors d'une autre série d'expériences, les aspects *sensoriels* (*passifs*) de la conscience. Les sujets testés avaient accepté l'implantation d'électrodes dans une région du cerveau liée à la réception de signaux sensoriels provenant de certains points de la peau. À cette stimulation directe se substituait parfois une stimulation sur la peau. La conclusion générale de cette expérience fut qu'il fallait environ une demi-seconde d'activité neuronale (avec des variations selon les circonstances) avant que les sujets prennent conscience de la moindre sensation. Dans le cas de la stimulation de la peau, ils avaient l'impression d'avoir eu conscience du stimulus avant que la peau ne soit effectivement stimulée.

Si elles peuvent paraître troublantes, ces expériences ne contiennent en elles-mêmes rien de paradoxal. Peut-être les décisions apparemment conscientes d'un individu sont-elles en fait prises *inconsciemment* à un instant antérieur, au moins une seconde avant. Peut-être les sensations conscientes d'un individu exigent-elles effectivement une demi-seconde d'activité cérébrale avant d'être réellement suscitées. Mais si l'on met bout à bout ces deux résultats, on est obligé de conclure que dans toute situation où un stimulus extérieur provoque une décision consciente, un laps de temps d'une seconde et demie semble nécessaire avant que cette réaction intervienne. La prise de conscience du stimulus n'a lieu qu'au bout d'une demi-seconde et, pour agir, il faut mettre en branle la machine apparemment léthargique du libre arbitre, ce qui nécessite une autre seconde.

Nos réactions conscientes sont-elles vraiment si lentes ? Ce ne semble pas être le cas, par exemple, dans la conversation ordinaire. Accepter (ii) amènerait à conclure que la plupart des réactions sont entièrement inconscientes mais que l'on peut, de temps en temps, leur substituer une réaction consciente, ce qui prend environ une seconde. Mais si la réaction est habituellement *in*consciente, elle n'a alors, sauf si elle est aussi lente qu'une réaction consciente, aucune chance d'être annulée par la conscience — sinon, lorsque l'acte conscient entre en jeu, la réaction inconsciente a déjà eu lieu et il est trop tard pour que la conscience puisse l'affecter ! Ainsi, à moins que les actes conscients ne soient *parfois* rapides, la réaction inconsciente met *elle-même* environ une seconde pour intervenir. (Rappelons encore une fois qu'une réaction inconsciente « préprogrammée » survient bien plus rapidement — en environ un cinquième de seconde.)

Bien entendu, la possibilité (i) n'interdit pas d'avoir une réaction inconsciente rapide (disons en un cinquième de seconde) intervenant alors que le système de réaction inconscient est dans l'ignorance totale d'une activité consciente (sensorielle) qui pourrait survenir plus tard. Dans ce cas (et la situation avec (iii) est encore pire), lors d'une conversation assez rapide, notre conscience n'est qu'une simple spectatrice, assistant pour ainsi dire à une « reconstitution » de la scène.

Il n'y a là aucune véritable contradiction. Il se peut que la sélection naturelle ait produit la conscience pour lui assigner un vrai rôle dans la pensée raisonnée, tandis que dans toute activité rapide elle constituerait un poids mort. Toute la discussion de la première partie tournait en définitive autour de cette forme de contemplation consciente (la compréhension mathématique) qui est effectivement d'une lenteur notoire. Peut-être la conscience a-t-elle évolué dans le seul but d'accomplir ce type d'activité mentale lente et contemplative, tandis que les temps de réaction plus rapides correspondent à des actes inconscients — accompagnés toutefois de leur perception consciente retardée, perception qui ne joue aucun rôle actif.

On ne peut nier que la conscience trouve sa justification lorsqu'on la laisse travailler longtemps. Mais j'avoue mon incrédulité face à l'idée qu'elle puisse n'avoir *aucun* rôle à jouer dans des activités raisonnablement rapides telles que la conversation ordinaire — ou encore le ping-pong, le squash, la course automobile. La discussion précédente me semble contenir au moins un grave point faible : elle suppose que le chronométrage précis d'événements conscients a réellement un sens. Existe-t-il *vraiment* un « instant » auquel survient une expérience consciente et qui précède celui où se manifeste l'effet provoqué par une réaction volontaire à cette expérience ? Eu égard au lien étrange unissant la conscience à la notion de temps (voir le début de cette section), l'existence d'un « temps » précis mesurant l'instant où survient un événement conscient me semble pour le moins contestable[11].

On pourrait envisager un étalement non local dans le temps, associant un certain flou à la relation entre expérience consciente et temps physique. Je pense toutefois que la véritable explication est bien plus subtile. Si la conscience est physiquement inexplicable sans une intervention, à un niveau fondamental, de la théorie quantique, il se pourrait que les énigmes-**Y** de cette théorie défient nos conclusions apparemment inébranlables sur les relations de causalité, de non-localité et de contrafactualité entre la conscience et le libre arbitre. Par exemple, peut-être le type de contrafactualité intervenant dans le problème d'Elitzur-Vaidman (*cf.* §5.2 et §5.9) a-t-il un rôle à jouer : le simple fait qu'un acte ou une pensée *pourraient* avoir lieu, même s'ils n'ont pas réellement lieu, affecte peut-être l'activité cérébrale. (Cela infirmerait certaines déductions apparemment logiques, comme celle qui consistait, dans l'argumentation ci-dessus, à éliminer la possibilité (ii).)

D'une manière générale, il faut rester très prudent à l'égard des conclusions apparemment logiques que l'on peut tirer sur l'ordre temporel d'événements comportant des effets quantiques (ainsi que le montreront, à la section suivante, les considérations sur l'effet EPR). Réciproquement, *si*, dans certaines

manifestations de la conscience, un raisonnement classique sur l'ordre temporel des événements conduit à une conclusion contradictoire, cela indique fortement la présence d'actions quantiques !

7.12 EPR et le temps : la nécessité d'une nouvelle vision du monde

Notre notion physique de temps, nous venons de le voir, est critiquable dès qu'on envisage le phénomène de la conscience. Mais elle laisse aussi à désirer en physique même, lorsque interviennent la non-localité et la contrafactualité quantiques. Si, dans les situations EPR, on adopte une vision résolument « réaliste » du vecteur d'état $|\psi\rangle$ — et, aux sections 6.3 et 6.5, j'ai amplement examiné les difficultés que l'on rencontre *en ne l'adoptant pas* —, on se trouve confronté à de profondes énigmes, affectant non seulement toute théorie de type GRW (*cf.* §6.9), mais aussi — potentiellement — le schéma **RO** (*cf.* §6.12) que je défends ici.

Reprenons nos dodécaèdres magiques de la section 5.3 et l'explication qui en a été donnée à la section 5.18. Demandons-nous laquelle des deux possibilités suivantes correspond à la situation « réelle » : sont-ce les pressions exercées par *mon collègue* sur ses boutons qui réduisent (et « démêlent ») instantanément l'état total initialement emmêlé — de sorte que l'état de l'atome de mon dodécaèdre se trouve alors instantanément créé, démêlé, et que c'est *cet* état réduit qui définit les possibilités pouvant résulter des pressions que je peux exercer sur mes propres boutons ? Ou inversement, sont-ce *mes propres* pressions qui interviennent en premier et agissent sur l'état initialement emmêlé pour réduire instantanément l'état de l'atome enfermé dans le dodécaèdre de mon collègue, de sorte que c'est ce dernier qui rencontre l'état démêlé réduit ? Nous l'avons remarqué à la section 6.5, peu importe, au niveau des résultats, la manière dont nous traitons le problème. Cela est d'ailleurs heureux, car l'inverse violerait les principes de la relativité d'Einstein selon lesquels la notion de « simultanéité » n'a aucun effet physiquement observable pour des événements distants (séparés par un intervalle du genre espace). Toutefois, si l'on croit que $|\psi\rangle$ représente la *réalité*, cette réalité a deux interprétations différentes. Certains voient là une raison suffisante pour considérer que $|\psi\rangle$ ne représente pas la réalité. D'autres invoquent les autres raisons d'adopter le point de vue réaliste (*cf.* §6.3) — et sont tout à fait prêts à renoncer à la vision einsteinienne du monde.

Personnellement, je souhaiterais conserver les deux points de vue — réalisme quantique *et* esprit de la vision spatio-temporelle relativiste. Mais cela impose que nous modifiions radicalement notre représentation de la réalité physique. Plutôt que d'exiger que la nouvelle description quantique (voire spatio-temporelle) qui en résultera soit dans le droit fil de celles qui nous sont

déjà familières, nous devrions plutôt rechercher quelque chose qui s'en démarque fortement, tout en leur étant (du moins au début) mathématiquement équivalent.

Cette situation a en fait un précédent célèbre. Avant la découverte de la relativité générale par Einstein, les scientifiques étaient tellement habitués à la théorie merveilleusement précise de Newton dans laquelle les particules, en mouvement dans un espace plat, s'attirent mutuellement selon la loi en inverse du carré de la force gravitationnelle, qu'il leur semblait que l'introduction d'une modification fondamentale dans cette théorie en détruirait inéluctablement la remarquable précision. Pourtant, ce qu'introduisit Einstein fut justement une modification fondamentale. Sa vision de la dynamique gravitationnelle a radicalement transformé le schéma newtonien. L'espace n'est plus plat (il n'est même plus « espace », mais « espace-temps ») ; il n'y a plus de force gravitationnelle ; celle-ci est remplacée par des effets de marée dus à la courbure spatio-temporelle. Et les particules ne se déplacent plus : elles sont représentées par des courbes « statiques » dessinées dans l'espace-temps. Cela a-t-il détruit la remarquable précision de la théorie newtonienne ? Nullement ; elle fut même améliorée à un degré incomparable ! (Voir la section 4.5.)

La même chose va-t-elle arriver à la théorie quantique ? Je crois cela extrêmement probable. Il faudra un changement de point de vue *radical*, ce qui rend délicates les spéculations sur la nature précise de ce changement. En outre, ce changement paraîtra sans doute totalement fou !

Pour conclure cette section, je voudrais mentionner deux idées apparemment folles, dont aucune n'est suffisamment folle, mais qui ont toutes deux leurs mérites. La première est due à Yakir Aharonov et Lev Vaidman (1990), à Olivier Costa de Beauregard (1989) et à Paul Werbos (1989). Selon cette idée, la réalité quantique est décrite par *deux* vecteurs d'état, l'un évoluant dans le sens du temps depuis la dernière occurrence de **R**, l'autre *à rebours*, depuis la prochaine occurrence de **R** dans le futur. Ce second vecteur d'état* a un comportement « téléologique », en ce sens qu'il dépend de ce qui va se passer dans l'avenir et non de ce qui est advenu dans le passé. Certains jugent inacceptable une telle propriété, mais cette théorie ayant des conséquences identiques à celles déduites de la théorie standard, on ne peut la rejeter. Son *avantage* sur la théorie quantique standard est de fournir une description complètement objective de l'état quantique dans les situations EPR admettant, en termes spatio-temporels, une représentation compatible avec l'esprit de la relativité d'Einstein. Cela fournit une (espèce de) solution aux énigmes mentionnées au début de cette section — avec cependant la contrainte d'inclure un état quantique à comportement téléologique, ce que nombre de physiciens estiment troublant. (Pour ma part, je trouve ces aspects téléologiques tout à

* Pour de bonnes raisons mathématiques, le vecteur d'état évoluant à rebours dans le temps est désigné par le *bra* $\langle\phi|$, tandis que celui qui évolue à l'endroit est désigné par le *ket* $|\psi\rangle$. La paire formée par ces deux vecteurs est alors représentée par le produit $|\phi\rangle\langle\psi|$, ce qui est compatible avec la notation utilisée pour les matrices densité (*cf.* §6.4).

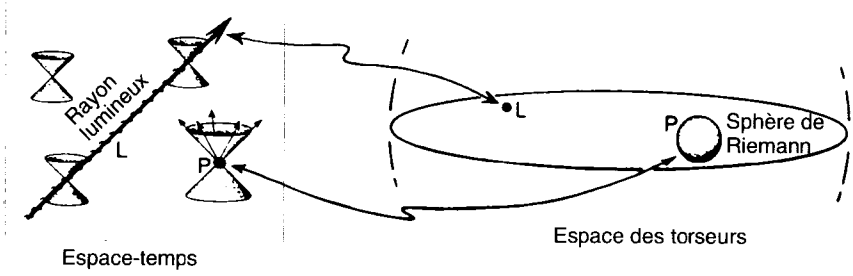


Figure 7.17. La *théorie des torseurs* offre une image physique différente de celle fournie par l'espace-temps. Les rayons lumineux y sont représentés par des points et les événements par des sphères de Riemann.

fait acceptables dans la mesure où ils n'entraînent pas de contradiction avec le comportement physique réel.) Pour les détails, je renvoie le lecteur à la littérature.

L'autre idée que j'aimerais évoquer correspond à ce que l'on désigne par *théorie des torseurs* (Fig. 7.17). Cette théorie fut fortement motivée par ces mêmes énigmes EPR, mais (en tant que telle) elle ne leur a pas *à ce jour* apporté de solution. Sa force réside ailleurs, dans le fait qu'elle fournit, de certaines notions physiques fondamentales (par exemple, les équations électromagnétiques de Maxwell, *cf.* §4.4 et EOLP, p.198-202), des descriptions mathématiques élégantes et imprévues. Elle donne une description non locale de l'espace-temps dans laquelle les rayons lumineux sont représentés par de simples points. C'est cette non-localité spatio-temporelle qui la relie à la non-localité quantique des situations EPR. En outre, elle repose fondamentalement sur les *nombres complexes* et la géométrie qui leur est associée, de sorte qu'elle permet de dégager un lien intime entre les nombres complexes de la théorie quantique gouvernée par **U** et la structure de l'espace-temps. En particulier, la sphère de Riemann (*cf.* §5.10) joue un rôle fondamental au niveau du cône de lumière attaché à un point spatio-temporel (au niveau aussi de la « sphère céleste » d'un observateur attaché à ce point). (Voir David Peat (1988) pour une description non technique des principales idées, ou Stephen Huggett et Paul Tod (1985) pour une description relativement brève mais technique[12].)

Un développement plus approfondi de ces idées sortirait du cadre de ce livre. Je ne les ai mentionnées que pour indiquer qu'il y a plusieurs façons de transformer notre représentation déjà extraordinairement précise du monde physique en quelque chose qui semble très différent des images auxquelles nous sommes aujourd'hui attachés. Afin de respecter une importante exigence de cohérence, ce changement devra être tel que la nouvelle description à laquelle il donnera naissance permette de retrouver tous les résultats corrects de la théorie quantique gouvernée par **U** (ainsi que ceux de la relativité générale). Mais il devra nous permettre d'aller plus loin et modifier la théorie quantique en substituant à **R** un processus physique réel. J'en ai la ferme conviction. Je suis en outre persuadé que cette « modification quantique » devra

s'opérer dans le sens des idées **RO** décrites à la section 6.2. Mentionnons que les théories associant la relativité à une réduction d'état « réaliste » — telle la théorie GRW — se sont jusqu'ici heurtées à de très graves problèmes (particulièrement en ce qui concerne la conservation de l'énergie). Cela tend à me conforter dans l'idée qu'il nous faut changer radicalement notre vision du monde pour pouvoir véritablement progresser au niveau de ces problèmes physiques cruciaux.

Je pense également que tout véritable progrès dans la compréhension physique du phénomène de la *conscience* exigera lui aussi — comme condition préalable — le même changement fondamental de notre vision du monde.

8
Quelles conséquences ?

8.1 Des « dispositifs » artificiellement intelligents

Que pouvons-nous conclure, d'après les discussions précédentes, sur les perspectives ultimes de l'intelligence artificielle ? Les arguments exposés dans la première partie de ce livre ont solidement établi que la technologie des robots contrôlés par ordinateurs *ne permettra pas* de construire une machine *réellement* intelligente — c'est-à-dire capable de comprendre ce qu'elle fait et d'agir sur cette compréhension. Certes les ordinateurs contribuent de manière extrêmement précieuse et efficace aux progrès réalisés dans les domaines scientifique, technologique et social ; certes ils jouent un rôle manifestement important dans la clarification de nombre des problèmes liés aux phénomènes mentaux (peut-être, dans une large mesure, en nous apprenant ce que les phénomènes mentaux authentiques *ne sont pas*). Mais nous sommes obligés de conclure qu'ils font quelque chose de très différent de ce que *nous* faisons lorsque nous appliquons notre réflexion consciente à la résolution d'un problème.

Il devrait toutefois être clair, d'après les dernières discussions de la deuxième partie, que je n'oppose nullement une impossibilité de principe à la construction d'un *dispositif* authentiquement intelligent, pourvu qu'un tel dispositif ne soit pas une « machine » — dans le sens précis où elle serait numériquement contrôlée — mais qu'elle intègre une action physique identique à celle qui suscite notre propre conscience. Ne disposant pas encore d'une théorie physique de cette action, il est certainement prématuré de se demander quand ou si un tel dispositif pourra être construit. Il n'en reste pas moins vrai que l'on peut envisager cette construction dans le cadre du point de vue $\mathcal{C}$ (*cf.* §1.3) auquel

je souscris et selon lequel la conscience sera un jour comprise en termes scientifiques quoique non calculables.

Rien n'oblige, selon moi, un tel dispositif à être de nature biologique. Aucune frontière, à mon avis, ne sépare fondamentalement la biologie et la physique (voire la biologie, la chimie et la physique). Les systèmes biologiques possèdent d'ailleurs généralement une organisation subtile surpassant de beaucoup même la plus sophistiquée de nos créations physiques (pourtant souvent très sophistiquées). Mais il est clair que notre compréhension physique de l'Univers en est encore à ses débuts — particulièrement en ce qui concerne les processus mentaux. Ainsi, on peut penser que nos constructions physiques seront bien plus sophistiquées dans l'avenir et qu'elles intégreront des effets physiques à peine concevables aujourd'hui.

Je ne vois aucune raison de douter que dans un futur plus immédiat, certains des effets déconcertants (les énigmes-**Y**) de la théorie quantique connaîtront de surprenantes applications dans des circonstances appropriées. On entrevoit déjà aujourd'hui la possibilité d'appliquer la théorie quantique à la cryptographie pour obtenir des résultats inaccessibles aux dispositifs classiques. Il existe en particulier plusieurs projets théoriques (*cf.* C. Bennett *et al.* 1983) visant à permettre l'échange d'informations secrètes entre deux personnes à l'aide d'un dispositif, reposant fondamentalement sur des effets quantiques, qui rendrait impossible toute interception du message transmis. Ce type de dispositif a déjà été construit en laboratoire et il ne fait aucun doute qu'il aura des applications commerciales d'ici quelques années. De nombreux autres « projets quantiques » sont actuellement à l'étude en cryptographie. Ils ont donné naissance à une nouvelle discipline, la *cryptographie quantique*, actuellement en plein essor. En outre, peut-être parviendra-t-on un jour à construire un *ordinateur quantique*. Ce concept théorique est toutefois encore très loin de pouvoir se concrétiser, au point qu'il est difficile de prédire quand — voire même si — il se réalisera effectivement (*cf.* Obermayer *et al.* 1988, *a*, *b*).

Il est encore plus difficile de dire si (ou quand) on pourra construire un dispositif dont l'action dépendra d'une théorie physique qui nous est aujourd'hui encore *inconnue*. Je l'ai dit, la découverte de cette théorie est un préalable incontournable à la compréhension de la physique qui régirait un dispositif au comportement non calculable — « non calculable » signifiant, rappelons-le, irréductible à l'action d'une machine de Turing. En vertu des arguments que j'ai développés, on ne pourra envisager la construction d'un tel dispositif qu'après avoir trouvé la théorie physique (**RO**) décrivant correctement la réduction de l'état quantique — et il est très difficile d'évaluer la distance qui nous sépare encore de cette théorie. Il se peut en outre que les particularités de cette théorie **RO** donnent à sa recherche un tour imprévu.

Cette connaissance théorique préalable est en tout état de cause une *supposition* personnelle. Peut-être en ira-t-il autrement. Il est souvent arrivé que des effets physiques n'aient pu être expliqués par la théorie que de nombreuses années après avoir été découverts. La supraconductivité en est un exemple célèbre. Elle fut observée pour la première fois (par Heike Kammerlingh Onnes en 1911) une cinquantaine d'années avant que Bardeen, Cooper et

Schrieffer n'en donnent une interprétation théorique (quantique) complète, en 1957. De même, la supraconductivité à haute température fut découverte en 1986 (*cf.* Sheng *et al.* 1988) alors qu'aucun fondement théorique ne justifiait son existence. (Alors que j'écris ces lignes — début 1994 —, ce phénomène n'a toujours pas reçu d'explication théorique satisfaisante.) Cependant, dans le cas d'un comportement non calculable, on voit difficilement comment on pourrait ne serait-ce que *savoir* quand un objet manifeste un tel comportement. Le concept même de calculabilité ressortit étroitement à la *théorie* et semble difficilement décelable par une observation directe. Peut-être, dans le cadre d'une théorie non calculable, pourra-t-on exhiber un comportement caractéristique de cette non-calculabilité et susceptible alors d'être soumis à l'expérimentation, voire d'être mis en évidence sur un dispositif concret. Je tends personnellement à penser que l'on a peu de chances d'observer ou de mettre en évidence un comportement non calculable sur un dispositif concret si l'on ne dispose pas au préalable d'une théorie.

Afin de développer davantage mes arguments, je vais maintenant supposer que nous *disposons* à la fois de cette théorie — qui, je l'ai dit, doit être une théorie **RO** non calculable de la réduction du vecteur d'état — et d'une confirmation expérimentale de cette théorie. Comment pourrions-nous alors construire un dispositif *intelligent* ? Sur la base de ces seules hypothèses, nous *ne le pourrions pas.* Il nous faudrait d'abord accomplir un autre progrès théorique. Plus précisément, il nous faudrait avoir une théorie expliquant comment la conscience peut émerger au sein d'une structure exploitant de manière adéquate des effets **RO** non calculables. Pour ma part, je n'ai aucune idée de ce que pourrait être une telle théorie. Peut-être, à l'instar de ce qui s'est passé avec la supraconductivité, tombera-t-on presque par hasard sur une structure de ce type avant même de disposer d'une théorie correcte de la conscience. Il va sans dire que cela me paraît très improbable — à moins que se déroule quelque part un processus d'évolution darwinien aux termes duquel l'intelligence résulterait des avantages conférés par la conscience, sans que nous ayons le moins du monde à comprendre comment elle est advenue (c'est d'ailleurs ce qui nous est arrivé !). Mais cela ne pourrait être qu'un processus extrêmement lent : que l'on songe à la lenteur avec laquelle la conscience donne naissance à ses manifestations. Le lecteur en conclura peut-être que la manière la plus satisfaisante de construire un dispositif intelligent consiste à employer les procédures peu méthodiques mais remarquablement efficaces que nous utilisons depuis des millénaires !

Bien entendu, rien de tout cela ne nous empêche de vouloir connaître les mécanismes sous-jacents à l'intelligence et à la conscience. Moi aussi je veux les connaître. La thèse centrale de ce livre est que ces mécanismes ne se réduisent pas à une pure activité de calcul — comme on le pense habituellement de nos jours — et qu'on ne peut guère espérer les comprendre sans avoir une vision plus profonde de la nature même de la matière, du temps, de l'espace et des lois qui les gouvernent. Nous avons également besoin d'une meilleure connaissance de la physiologie du cerveau, en particulier au niveau des processus infimes étudiés depuis seulement quelques années. Il nous faut

mieux comprendre ce qui favorise ou s'oppose à l'intervention de la conscience, les rythmes étranges qui la régissent, les objectifs qui lui sont assignés, et les avantages particuliers procurés par sa possession — en plus de nombreux autres aspects se prêtant à un contrôle expérimental objectif. On le voit, c'est là un immense champ de recherches et il est à prévoir qu'il sera le théâtre de multiples découvertes.

8.2 Ce que les ordinateurs font bien – et moins bien

Même si l'on accepte que le concept actuel d'ordinateur *ne permettra pas* de parvenir à une conscience ou une intelligence authentiques, on ne peut nier l'extraordinaire puissance des ordinateurs modernes, qui va peut-être s'accroître de façon fabuleuse (*cf.* §1.2, §1.10, et Moravec 1988). Certes, ces machines ne *comprendront* pas ce qu'elles feront, mais elles le feront avec une rapidité et une précision incroyables. Une telle activité — qui n'en restera pas moins aveugle — accomplira-t-elle des choses pour lesquelles nous devons recourir à notre esprit ? Les réalisera-t-elle même plus efficacement que nous ? Pouvons-nous déjà entrevoir les domaines dans lesquels ces machines excelleront, ou ceux dans lesquels le cerveau restera toujours supérieur ?

Les ordinateurs jouent déjà extraordinairement bien aux échecs — ils frôlent le niveau des meilleurs grands-maîtres. Au jeu de dames, l'ordinateur Chinook n'a trouvé qu'un adversaire qui lui soit supérieur : la championne Marion Tinsley. Mais à l'antique jeu de go, ils ne semblent pas aller bien loin. Les jeux exigeant de la vitesse les favorisent, tandis que ceux qui autorisent un long temps de réflexion avantagent les joueurs humains. Un ordinateur résout presque instantanément les problèmes d'échecs requérant deux ou trois coups, quelle que soit leur difficulté pour un être humain. En revanche, un problème élémentaire mais dont la résolution exige, par exemple, 50 ou 100 coups peut résister à la machine, tandis qu'il ne pose pas trop de difficultés à un joueur humain expérimenté (*cf.* aussi §1.15, Fig. 1.7).

Ces différences s'expliquent largement par les aptitudes spécifiques des ordinateurs et des hommes. L'ordinateur se contente d'exécuter des calculs sans comprendre le moins du monde ce qu'il fait — même s'il utilise une part du savoir-faire de ses *programmeurs*. Il peut disposer d'une grande quantité d'informations stockées en mémoire, mais cela peut être également le cas pour un joueur humain. L'ordinateur peut appliquer automatiquement et de nombreuses fois de suite le savoir-faire de ses programmeurs, avec une rapidité et une précision extrêmes, et ces aptitudes dépassent de loin celles de tout être humain. Le joueur humain doit, quant à lui, constamment réviser ses jugements et élaborer des stratégies à partir d'une compréhension générale du sens du jeu. L'ordinateur ne sait pas faire cela, mais dans une large

mesure, il peut utiliser sa puissance de calcul pour pallier son manque de compréhension.

Supposons que l'ordinateur ait besoin d'envisager en moyenne p possibilités par coup ; pour anticiper m coups, il doit alors considérer p^m éventualités. Si le calcul de chacune d'elles l'occupe en moyenne durant un temps t, le temps total T mis pour anticiper les m coups est alors :

$$T = t \times p^m.$$

Aux dames, p n'est pas très grand, disons de l'ordre de quatre, de sorte que l'ordinateur peut anticiper un nombre de coups considérable dans le temps imparti — en fait, une vingtaine ($m = 20$) —, tandis qu'au jeu de go, p peut atteindre 200, si bien qu'un ordinateur ne peut probablement guère anticiper au-delà de, disons, 5 coups ($m = 5$). Le cas des échecs est plus ou moins intermédiaire. Maintenant, rappelons-nous que les facultés de jugement et de compréhension d'un être humain sont bien plus lentes que celles d'un ordinateur (grand t pour l'être humain, petit t pour l'ordinateur), mais que ces facultés peuvent réduire très considérablement le nombre *effectif* p (petit p pour l'être humain, grand p pour l'ordinateur), car le joueur humain peut voir que seul un petit nombre de coups possibles mérite d'être pris en considération.

D'une manière générale, il en résulte que les jeux pour lesquels p est grand mais peut être significativement réduit en recourant aux facultés de jugement et de compréhension sont plutôt favorables au joueur humain. Car si l'on se donne un T raisonnablement grand, la réduction par l'homme du « *p effectif* » se traduit par une différence bien plus importante lorsque m est grand dans la formule $T = t \times p^m$ que ne le fait une importante réduction du temps t (ce à quoi les ordinateurs excellent). Mais pour un petit T, c'est en prenant t très petit que l'on peut être plus efficace (car les valeurs de m sont alors généralement petites). Ces faits sont de simples conséquences de la forme « exponentielle » de l'expression $T = t \times p^m$.

Ces considérations sont assez grossières, mais je pense qu'elles suffisent pour dégager le point essentiel. (Si vous n'êtes pas mathématicien, vous pourrez vous faire une idée du comportement de $T = t \times p^m$ en essayant quelques valeurs de t, p et m.) Sans trop entrer dans les détails, il y a toutefois un point qui mérite peut-être d'être éclairci. On pourrait objecter que le joueur humain ne cherche pas réellement à « calculer un grand nombre de coups à l'avance », (*i.e.* ne recherche pas un grand m). Or c'est justement le cas ! Lorsque le joueur humain juge de la valeur d'une position en calculant quelques coups d'avance puis considère qu'il est inutile d'en calculer plus, son calcul porte *en fait* sur un nombre de coups bien plus grand, car son jugement prend en compte l'effet probable sur les coups ultérieurs. Quoi qu'il en soit, ces considérations sommaires permettent de comprendre pourquoi il est bien plus difficile de construire des ordinateurs champions du jeu de go que d'en faire des virtuoses du jeu de dames, pourquoi aux échecs les ordinateurs s'en sortent mieux avec les problèmes requérant peu de coups, et pourquoi ils possèdent un avantage relatif lorsque les temps de réflexion sont brefs.

Sans être particulièrement sophistiqués, ces arguments montrent que le *jugement* humain, qui repose sur la *compréhension*, est un élément essentiel dont ne disposent pas les ordinateurs — les considérations données à la section 1.15 sur la configuration d'échecs de la figure 1.7 vont dans le même sens. Si la compréhension consciente est un processus relativement lent, elle peut cependant considérablement réduire le nombre de possibilités méritant d'être sérieusement étudiées et par là même accroître grandement le nombre de coups *effectivement* envisagés. (Au-delà d'un certain point, on n'a même plus besoin d'essayer.) En fait, il me semble que si l'on veut avoir une idée de ce que les futurs ordinateurs seront capables d'accomplir, il suffit de se demander : « La compréhension est-elle nécessaire à l'accomplissement de telle ou telle tâche ? » Nombre de tâches de la vie quotidienne n'exigent pas une grande compréhension, et il est fort probable qu'elles pourront être excellemment accomplies par des robots numériquement contrôlés. Il existe déjà des machines, pilotées par des réseaux de neurones formels, qui effectuent très honorablement ce genre de tâches. Par exemple, elles s'avèrent raisonnablement efficaces dans la reconnaissance de visages, la prospection minière, la détection des anomalies d'usinage par des moyens acoustiques, la vérification des cartes de crédit, etc.[1] En général, ces machines à stratégie ascendante fonctionnent bien lorsque leurs aptitudes approchent, voire dépassent quelque peu, celles d'experts humains moyens. Mais elles ne témoignent pas d'une « expertise » comparable à celle des systèmes descendants, par exemple ceux destinés aux échecs sur ordinateur ou — plus impressionnant encore — au calcul numérique direct pour lequel les performances des meilleurs calculateurs humains n'approchent jamais celles des calculateurs électroniques. Dans le cas de tâches effectivement confiées aux systèmes à réseaux de neurones formels (ascendants), il est probablement juste de dire que les *humains*, tout comme les ordinateurs, ne font guère appel à la compréhension ; on peut donc penser que pour ces tâches, les ordinateurs auront un certain succès. Lorsque la programmation de l'ordinateur repose sur une bonne dose de stratégie descendante — par exemple, pour le calcul numérique, les échecs sur ordinateur ou le calcul à buts spécifiquement scientifiques —, ils pourront s'avérer extrêmement efficaces. Dans ces cas, la machine, encore une fois, n'a nul besoin de comprendre ce qu'elle fait. La « compréhension » lui a été fournie par les programmeurs humains (*cf.* §1.21).

Remarquons aussi que les systèmes descendants commettent très fréquemment des erreurs — parce que le programmeur en a lui-même commis. Mais c'est le résultat d'une erreur humaine, ce qui est un tout autre problème. Les correcteurs automatiques d'erreurs, s'ils s'avèrent des outils précieux, échouent cependant à détecter les erreurs trop subtiles.

Dans certaines situations, il peut être dangereux d'accorder un trop grand crédit à un système entièrement piloté numériquement. C'est par exemple le cas lorsque le système, après avoir accompli raisonnablement bien sa tâche durant un temps assez long — donnant peut-être le *sentiment* qu'il comprend ce qu'il fait — exécute soudain quelque chose qui paraît complètement fou, révélant ainsi qu'il n'a en fait jamais *réellement* compris ce qu'il faisait (comme

Deep Thought lorsqu'on le confronta au problème d'échecs de la figure 1.7). Ainsi, il faut toujours rester vigilant. On peut certes continuer de travailler avec ces systèmes, mais il faut garder à l'esprit que la « compréhension » n'est pas une qualité purement calculable, autrement dit qu'un robot entièrement contrôlé par ordinateur n'a aucune chance d'en posséder la moindre parcelle.

Bien sûr, les êtres humains n'ont pas tous les mêmes facultés intellectuelles. Et à l'instar des ordinateurs, un être humain peut donner l'impression qu'il comprend ce qu'il fait alors qu'en réalité il n'en est rien. Il semble y avoir une sorte de compromis entre d'une part la compréhension authentique et d'autre part la mémoire et la puissance de calcul. Les ordinateurs ont la mémoire et la puissance de calcul, mais pas la compréhension. Les enseignants le savent bien (mais pas toujours, hélas, les gouvernants), la *compréhension* est la plus précieuse des qualités. C'est elle, et non le simple fait de répéter comme un perroquet des règles et des informations, que l'on cherche à encourager chez les élèves. De fait, les questions intelligentes lors des examens (et particulièrement en mathématiques) sont celles qui permettent de tester la compréhension des candidats en la distinguant de la mémoire ou de l'aptitude au calcul — même si ces deux dernières qualités sont également des outils précieux.

8.3 Esthétique, etc.

La discussion précédente a essentiellement porté sur l'absence fondamentale de « compréhension » dans tout système purement numérique. C'est déjà ce qui faisait l'essentiel de l'argumentation gödelienne exposée à la section 2.5. Cette absence de compréhension dans l'activité algorithmique automatique avait alors mis en lumière les limitations fondamentales du calcul et nous avait incités à rechercher autre chose. Pourtant, la compréhension n'est que l'une des qualités pour lesquelles nous avons besoin de la connaissance consciente. D'une manière plus générale, nous, êtres conscients, tirons bénéfice de toute circonstance dans laquelle nous pouvons « ressentir » les choses ; et j'affirme que c'est justement *cela* qu'un système purement algorithmique ne pourra jamais faire.

On peut se demander dans quelle mesure un robot numériquement contrôlé est *désavantagé* par cette inaptitude à ressentir, de sorte qu'il ne peut apprécier, par exemple, la beauté d'un ciel étoilé ou la splendeur du Tadj Mahal dans le calme d'une soirée, ou encore les complexités magiques d'une fugue de Bach — voire même la froide beauté du théorème de Pythagore. Certes on pourrait se borner à regretter que le robot ne puisse ressentir ce que nous éprouvons lorsque nous sommes en présence de tels spectacles. Pourtant, cette question est plus profonde qu'il n'y paraît. Formulons-la différemment : si l'on reconnaît qu'un robot est incapable de *ressentir* quoi que ce soit, un ordinateur intelligemment programmé pourrait-il toutefois produire de grandes œuvres d'art ?

À mon avis, c'est là une question délicate. Je dirais que la réponse est simplement « non » — ne serait-ce que parce que l'ordinateur ne possède pas les qualités sensuelles nécessaires pour distinguer une œuvre magnifique d'une œuvre médiocre, voire simplement acceptable. On peut cependant se demander pourquoi il est *nécessaire* que l'ordinateur soit capable de « ressentir » pour pouvoir développer ses propres « critères esthétiques » et former ses propres jugements. On pourrait imaginer que ces jugements puissent « émerger » au terme d'une longue période d'éducation de la machine (par stratégie ascendante). Toutefois, en ce qui concerne la compréhension, il me semble bien plus probable que ces critères devraient faire partie des données d'entrée délibérément injectées dans l'ordinateur et être soigneusement distillées à partir d'une analyse descendante détaillée (très probablement épaulée par un ordinateur) effectuée par des êtres humains sensibles à la beauté. C'est en fait là le type de schéma adopté par un certain nombre de chercheurs en intelligence artificielle. Par exemple, lors d'un travail effectué à l'université du Sussex, Christopher Longuet-Higgins a utilisé divers systèmes informatiques composant de la musique selon des critères qu'il leur a lui-même fournis. Déjà au XVIII[e] siècle, Mozart et ses contemporains avaient montré comment construire des « dés musicaux » pouvant être utilisés pour combiner des ingrédients esthétiques connus à des éléments aléatoires afin de produire des compositions vaguement honorables. Des dispositifs analogues ont été adoptés dans les arts plastiques, tel le système « AARON » programmé par Harold Cohen, qui peut produire de nombreux dessins « originaux » en faisant appel à des éléments aléatoires pour combiner selon certaines règles des données d'entrée fixes. (Voir Margaret Boden (1990) pour de nombreux exemples de ce type de « créativité informatique » ; voir aussi Michie et Johnson (1984).)

La plupart des gens, je pense, reconnaît généralement que les créations ainsi obtenues ne soutiennent pas, jusqu'à présent, la comparaison avec les productions d'un artiste humain modérément compétent. Je pense que l'on peut légitimement dire que lorsque les données d'entrée de l'ordinateur atteignent le moindre niveau significatif, le résultat manque d'une certaine « âme » ! Autrement dit, le travail n'*exprime* rien parce que l'ordinateur lui-même ne *ressent* rien.

Bien sûr, il pourrait arriver que, de temps en temps, une telle création informatique générée de manière aléatoire ait, par hasard, une authentique valeur artistique. (C'est l'éternelle question : aurait-il été possible d'écrire *Hamlet* en frappant au hasard sur le clavier d'une machine à écrire ?) De fait, il faut reconnaître que la nature produit elle-même de nombreuses œuvres d'art en recourant à des procédures aléatoires — que l'on songe à la beauté des formations rocheuses ou aux étoiles dans le ciel. Mais sans l'aptitude à *sentir* cette beauté, rien ne permet de distinguer le beau de l'horrible. C'est dans ce processus de *sélection* qu'un système entièrement algorithmique révélerait ses limitations fondamentales.

Ici encore, on pourrait admettre qu'un être humain introduise dans un ordinateur des critères algorithmiques qui fonctionneraient honorablement, mais avec le seul objectif de générer quantité d'exemplaires d'un même type

d'œuvre (comme il arrive souvent pour une mode artistique ordinaire) — jusqu'à ce que les gens se lassent de ces produits et recherchent quelque chose de nouveau.

Ainsi, outre la *compréhension*, un système entièrement algorithmique manquera toujours d'autres qualités, telles les qualités *esthétiques*. Il me semble qu'il faut ajouter à ces dernières d'autres choses qui exigent aussi la conscience, par exemple les jugements *moraux*. Nous l'avons vu dans la première partie, l'appréciation du *vrai* et du *faux* ne se réduit pas à un algorithme. La même chose vaut (et c'est peut-être plus évident encore) pour la *beauté* ou le *bien*. Ces valeurs exigent la conscience et sont donc inaccessibles à des robots numériquement contrôlés. Il doit toujours y avoir injection d'un contrôle continu par une présence consciente, extérieure et sensible — probablement humaine.

Indépendamment de leur nature non algorithmique, on peut se demander si la beauté ou la bonté sont des qualités *absolues*, dans le même sens platonicien où l'on parle de vérité « absolue » — et particulièrement de vérité mathématique. Se pourrait-il que notre connaissance immédiate puisse entrer en contact avec de tels absolus, et que ce soit *cela* qui confère à la conscience sa force fondamentale ? Peut-être y a-t-il là des indices sur la *nature* réelle de notre conscience et sur sa *fonction* ? La conscience est-elle une sorte de « pont » entre nous et les absolus platoniciens ? Je reviendrai sur ces questions à la dernière section de ce livre.

Le problème de la nature absolue de la moralité est lié aux problèmes légaux mentionnés à la section 1.11. Il est lié aussi à la question du « libre arbitre » soulevée la fin de cette même section 1.11 : y aurait-il, irréductible à notre héritage génétique, aux facteurs environnementaux et aux influences du hasard, un « moi » bien défini qui interviendrait à un niveau fondamental dans le contrôle de nos actes ? Je crois que nous sommes très loin de pouvoir répondre à une telle question. Dans le cadre des arguments exposés dans ce livre, le plus que je puisse dire avec certitude est que si ce « moi » existe, il ne peut, d'un point de vue de principe, résulter de l'activité des dispositifs que l'on désigne aujourd'hui sous le nom d'« ordinateurs ».

8.4 Quelques dangers inhérents à l'informatique

Toute technologie à grande échelle présente à la fois des dangers et des avantages. Ainsi, à côté des bénéfices évidents qu'il a apportés, le développement rapide de l'informatique renferme de nombreuses menaces pour notre société. L'un des principaux problèmes semble être l'extraordinaire complexité du réseau reliant les ordinateurs et qui interdit à un individu seul de comprendre l'ensemble de la situation ainsi créée. La difficulté tient non seulement à la technologie informatique elle-même, mais aussi aux communications quasi instantanées qui lient en permanence presque tous les ordinateurs de la

planète. On aura une idée des problèmes qui peuvent survenir en songeant à l'instabilité des marchés financiers sur lesquels s'effectuent pratiquement instantanément des transactions reposant sur des prédictions informatiques qui proviennent de tous les coins du globe. Le problème n'est peut-être pas tant l'impossibilité pour un individu de maîtriser intellectuellement l'ensemble de ce système d'interconnexions, que l'existence d'une instabilité (pour ne pas dire un manque de loyauté) inhérente à un système organisé pour permettre à des individus de faire instantanément fortune en surpassant leurs rivaux en puissance de calcul et en pouvoir de prédiction. Il est cependant très probable que d'autres instabilités et d'autres dangers potentiels résultent de la complexité même des interconnexions de l'ensemble du système.

Certaines personnes pensent peut-être qu'il n'y a rien d'inquiétant à ce qu'un système informatique devienne si complexe qu'il se dérobe à toute compréhension humaine. Peut-être ces gens pensent-ils que les ordinateurs, *eux*, finiront par comprendre ce système. Mais, nous l'avons vu, la compréhension n'est même pas *envisageable* pour un ordinateur, de sorte qu'il n'y a rien à espérer de ce côté.

D'autres problèmes, d'une nature différente, résultent du fait que les progrès technologiques sont si rapides qu'un système informatique peut devenir obsolète peu après son apparition sur le marché. Il y a fort à parier qu'à l'avenir, les utilisateurs seront contraints de renouveler de plus en plus souvent leur matériel et d'utiliser des systèmes qui, à cause des pressions de la concurrence, ne seront pas toujours correctement testés.

Les problèmes profonds posés par les progrès rapides de la technologie informatique sont trop nombreux pour être énumérés ici. Le manque de confidentialité, l'espionnage industriel et le sabotage informatique sont parmi ceux qui viennent immédiatement à l'esprit. Il y aura aussi un jour la possibilité de « forger » l'image d'une personne et de la présenter sur un écran de télévision en lui faisant exprimer des points de vue que cette personne ne veut pas rendre publics[2]. Il y a aussi des problèmes sociaux qui, sans être propres à l'informatique, lui sont quand même liés. Par exemple, la possibilité de reproduire des sons ou des images avec une précision fabuleuse donnera à un très petit nombre d'« artistes » en vogue la possibilité de propager leur production à travers le monde, peut-être aux dépens de ceux qui seront moins favorisés. Les « systèmes experts », qui permettent d'intégrer dans un progiciel les compétences et les connaissances d'un petit nombre d'individus — appartenant par exemple aux professions juridiques ou médicales —, peut-être au détriment des avocats ou des médecins locaux, en sont un autre exemple. Je pense toutefois qu'en l'occurrence, ces systèmes experts, parce qu'ils n'auront pas une *compréhension* entière de telle ou telle situation locale, ne seront qu'un outil auxiliaire et ne pourront se substituer totalement au savoir-faire des médecins ou des avocats humains.

Bien sûr, tous ces progrès n'auront pas que des désavantages pour le commun des mortels. Un public bien plus large pourra disposer bien plus librement de la compétence des experts. En ce qui concerne la confidentialité, il existe aujourd'hui des systèmes à « accès réservé » (voir Gardner 1989) qui

peuvent en principe être utilisés par des individus ou de petites entreprises — avec autant d'efficacité que les grosses entreprises — et *semblent* garantir une sécurité totale. Par leur nature même, ces systèmes exigent des ordinateurs puissants et très rapides (dont l'efficacité dépend de la difficulté qu'il y a à factoriser des grands nombres) et sont aujourd'hui menacés par les progrès que risque de faire l'informatique quantique (*cf.* §7.3 ; voir aussi Obermayer *et al.* (1988*a*, *b*) pour la faisabilité de l'informatique quantique). Nous l'avons mentionné à la section 8.1, on peut utiliser la cryptographie quantique pour se prémunir contre l'interception des messages — l'efficacité de ce système de sécurité dépend elle aussi de la possibilité d'effectuer d'énormes calculs. Il n'y a manifestement aucun moyen systématique d'évaluer les dangers et les avantages de toute nouvelle technologie, qu'elle soit ou non directement liée à l'informatique.

Pour conclure ce survol des problèmes sociaux liés à l'informatique, je vais raconter une histoire fictive mais qui traduit une inquiétude née en moi alors que je méditais sur toute une classe nouvelle de problèmes possibles. Pour autant que je sache, personne n'a encore exprimé cette inquiétude, mais elle me semble correspondre à un nouveau type de danger informatique.

8.5 L'élection truquée

Le jour de l'élection tant attendue approche. Les sondages se succèdent depuis plusieurs semaines. Tout semble indiquer qu'il manque au parti au pouvoir trois ou quatre points pour gagner. Bien sûr, ce chiffre subit des fluctuations dans un sens ou dans l'autre — parce que les sondages reposent sur des échantillons relativement petits, quelques centaines d'électeurs, tandis que sur l'ensemble du pays (plusieurs dizaines de millions de personnes), les opinions varient considérablement d'une région à l'autre. En fait, la marge d'erreur de chacun de ces sondages tourne plus ou moins autour de ces mêmes trois ou quatre points, de sorte qu'aucun n'est réellement fiable. Pourtant, la conclusion d'ensemble est plus crédible. Considérés tous ensemble, ces sondages présentent une marge d'erreur bien plus faible et l'accord entre eux semble correspondre exactement à la légère variation que l'on pourrait prédire d'après un raisonnement statistique. La moyenne des résultats de ces sondages présente une marge d'erreur inférieure à deux points. Certains observateurs prévoient un léger revirement en faveur du parti au pouvoir dans les sondages qui seront réalisés la veille de l'élection et affirment que le jour même, une faible proportion d'électeurs indécis (ou relativement peu concernés) décidera finalement de voter pour lui. Même ainsi, un revirement en faveur de ce parti n'arrangerait ses affaires que s'il l'emportait de quelque huit points sur ses plus proches rivaux, car alors seulement il aurait la majorité relative nécessaire pour stopper la coalition de ses opposants. Mais, tout le monde le sait, les sondages

d'opinion ne sont que des estimations. Seul le *vrai* vote exprime le choix réel du peuple, et pour connaître celui-ci, il faut attendre le dépouillement du scrutin.

Le jour de l'élection arrive, les gens votent, puis vient la fermeture des bureaux. On compte les bulletins. Le résultat est une surprise totale pour presque tout le monde — en particulier pour les instituts de sondage qui avaient consacré tant d'énergie et de savoir-faire, sans parler des réputations mises en jeu, à analyser leurs enquêtes. Le parti au pouvoir se maintient avec une confortable majorité, dépassant de huit points ses plus proches rivaux. Un nombre considérable d'électeurs est stupéfait — même horrifié. D'autres, bien que tout aussi surpris, sont ravis. Pourtant, le résultat est faux. Le trucage a été obtenu à l'aide d'une procédure subtile qui a échappé à tout le monde. Il n'y a pas eu d'urnes bourrées ; aucun bulletin n'a été perdu, remplacé par un autre ou compté deux fois. Les personnes chargées du décompte des voix ont fait leur travail consciencieusement et, pour la plupart, sans se tromper. Pourtant, le résultat est horriblement faux. Comment est-ce arrivé, et qui est le responsable ?

Il se peut que la direction du parti au pouvoir ignore totalement ce qui s'est passé. Elle n'est pas forcément directement responsable, même si elle bénéficie du trucage. En revanche, il y a d'autres personnes, en coulisses, qui ont eu peur pour leur peau en cas de défaite. Elles appartiennent à une organisation plus appréciée (pour d'excellentes raisons !) du parti au pouvoir que de ses opposants — une organisation dont ce parti a eu soin de préserver, voire de renforcer, le secret des activités. Bien que cette organisation soit légale, la majeure partie de ses véritables activités ne l'est pas, et elle ne répugne pas à commettre des actions politiquement perverses. Peut-être ses membres redoutent-ils sincèrement (mais à tort) que les opposants du parti au pouvoir ne ruinent le pays ou ne le livrent aux appétits d'une puissance étrangère. Parmi eux, il y a des experts — des experts extraordinairement doués — en fabrication de virus informatiques !

Rappelons ce que sont les virus informatiques. Le type le plus connu correspond à ceux qui, un jour fixé à l'avance, peuvent détruire toute la mémoire des ordinateurs qu'ils infectent. Par exemple, l'utilisateur horrifié voit sur son écran les lettres quitter leur place, descendre au bas de la fenêtre puis disparaître. Ou encore, un message obscène apparaît à l'écran. Dans tous les cas, les données ont de fortes chances d'être irrémédiablement perdues. En outre, tout disque inséré dans la machine se trouve infecté et transmet son infection à la machine suivante. Il existe bien sûr des programmes antiviraux. Mais ils ne sont efficaces que si l'on peut détecter le virus et si l'on en connaît la nature avant qu'il n'entre en action. Une fois qu'il a frappé, on ne peut plus rien faire.

Les virus de ce type sont habituellement fabriqués par des pirates de l'informatique, souvent des programmeurs mécontents, qui veulent semer le désordre, parfois pour des raisons compréhensibles, parfois gratuitement. Mais les membres de notre organisation ne sont pas, eux, des amateurs ; ce sont des professionnels grassement payés. Leurs virus sont indécelables par les programmes antiviraux standard et sont programmés pour frapper un jour

judicieusement fixé à l'avance — le jour de l'élection. Une fois le travail accompli — travail bien plus subtil que la simple destruction de données —, les virus s'autodétruisent sans laisser, à part l'acte malveillant lui-même, la moindre trace de leur existence.

Pour qu'un tel virus soit efficace dans une élection, il faut qu'existe, lors du décompte des voix, une étape qui n'est pas contrôlée par les êtres humains, soit à la main, soit à l'aide d'une calculette. (Un virus ne peut infecter que les ordinateurs programmables.) Peut-être le contenu de chacune des urnes a-t-il été correctement compté ; mais il faut additionner les résultats. Comment être plus efficace, précis et moderne qu'en effectuant cette addition sur un ordinateur — qui additionnera peut-être une centaine de ces chiffres — plutôt qu'à la main ou sur une calculette ! Cela n'offre certainement aucun risque d'erreur. Car on obtient le même résultat quel que soit le propriétaire de l'ordinateur utilisé. Les membres du parti au pouvoir obtiennent des résultats parfaitement identiques à ceux de leurs opposants, ou de toute autre partie intéressée, ou de tout observateur neutre. Peut-être utilise-t-on des ordinateurs de marques ou de modèles différents, mais cela ne porte pas à conséquence. Les experts de notre organisation connaissent toutes ces machines et ont conçu un virus distinct pour chacune d'elles. Bien que la fabrication de chacun de ces virus diffère légèrement — pour qu'il soit adapté à chaque machine individuelle —, les résultats sont identiques et l'accord entre les diverses machines convainc même les plus sceptiques.

Mais en dépit de cet accord entre les machines, les chiffres sont uniformément faux. Ils ont été intelligemment concoctés selon une formule précise dépendant dans une certaine mesure du comptage réel des votes — d'où l'accord entre les machines et la vague plausibilité du résultat — et visant à donner au parti au pouvoir la majorité qui lui est nécessaire ; et bien qu'il exige un petit effort de crédulité, le résultat semble somme toute incontestable. Il *semble* qu'à la dernière minute, un nombre significatif d'électeurs aient pris peur et voté pour le parti au pouvoir.

Dans la situation imaginaire que je viens de décrire, ce revirement n'a en fait pas eu lieu et le résultat de l'élection est faux. Si ce récit m'a été inspiré par de récentes élections britanniques (celles de 1992), je tiens à dire que le système de comptage officiel adopté en Grande-Bretagne *ne permet pas* ce type de fraude. Toutes les phases du comptage sont faites à la main. Cette méthode peut paraître démodée et inefficace, mais il importe de la conserver — ou du moins de conserver un système contenant des protections bien précises écartant tout soupçon de fraude.

En fait, si l'on regarde le côté positif des choses, les ordinateurs actuels offrent de merveilleuses occasions d'utiliser des procédures de vote dans lesquelles l'opinion de l'électorat serait bien plus équitablement représentée qu'elle ne l'est aujourd'hui. Ce n'est pas le lieu ici d'aborder ces sujets ; je voudrais cependant dire que ces machines permettraient à chaque électeur de transmettre bien plus d'information que ne le fait le simple bulletin de vote. Avec un système contrôlé par ordinateur, cette information pourrait être instantanément analysée et le résultat de l'élection connu immédiatement après

la fermeture des bureaux. Toutefois, ainsi que le montre l'histoire que je viens de raconter, il faut rester extrêmement vigilant avec un tel système et ne l'utiliser que s'il contient des contrôles clairs et efficaces permettant de se prémunir contre toute fraude.

Ce n'est pas seulement durant les élections que l'on doit rester vigilant ; le sabotage des comptes d'un concurrent est un autre exemple de travail qui pourrait être confié à un virus. Et il existe bien d'autres domaines dans lesquels on pourrait envisager l'utilisation de virus à des fins dévastatrices. J'espère que mon récit aura montré de façon suffisamment convaincante que les êtres humains doivent en permanence contester l'autorité et la fiabilité apparentes des ordinateurs. Outre que ces derniers ne comprennent rien, ils se prêtent très facilement aux manipulations que peuvent commettre les êtres humains qui maîtrisent dans le détail la façon dont ils ont été spécifiquement programmés.

8.6 La conscience, un phénomène physique ?

La deuxième partie de ce livre cherchait à déterminer, en termes scientifiques, la localisation physique de l'expérience subjective. Je l'ai dit, cette recherche exige une extension de nos connaissances scientifiques actuelles. Je suis pratiquement convaincu que seule une analyse approfondie de la procédure de réduction de l'état quantique nous permettra de définir les éléments de notre représentation physique actuelle qui nécessitent un changement radical. La physique ne pourra intégrer une entité aussi étrangère à nos concepts physiques actuels que le phénomène de la conscience sans subir une transformation radicale — affectant les fondements mêmes de notre approche philosophique de la nature de la réalité. Je reviendrai sur ce point à la dernière section du livre. Pour l'instant, essayons de répondre à cette question apparemment simple : les arguments présentés dans ce livre permettent-ils de localiser la conscience dans l'Univers connu ?

Je m'empresse de dire que ces arguments s'avèrent assez peu positifs sur ce plan. Ils disent que les ordinateurs actuels ne sont pas conscients, mais ils ne disent pas grand-chose sur les objets qui le *seraient* éventuellement. Par expérience, nous tendons à penser, du moins jusqu'ici, que ce phénomène est plutôt associé aux structures biologiques. À une extrémité de l'échelle, nous avons les êtres humains pour lesquels il est clair que quoi que soit la conscience, nous la supposons normalement présente dans le cerveau humain éveillé (mais aussi rêvant).

Qu'en est-il de l'autre extrémité ? J'ai expliqué que c'est dans les microtubules du cytosquelette, plutôt que dans les neurones, que nous avons le plus de chances de trouver des effets quantiques collectifs (cohérents) — et que sans cette cohérence quantique, la nouvelle physique **RO** ne peut avoir une influence suffisante pour donner naissance à l'activité non calculable imposée

par l'interprétation de la conscience en termes scientifiques. Pourtant, les cytosquelettes sont omniprésents dans les cellules eucaryotes — les cellules constitutives des plantes et des animaux, mais aussi d'animaux unicellulaires tels que les paramécies et les amibes, à l'exclusion des bactéries. Peut-on espérer trouver une trace de conscience dans une paramécie ? Une paramécie *sait*-elle, quel que soit le sens de ce verbe, ce qu'elle fait ? Et qu'en est-il des cellules humaines *individuelles*, qu'elles se trouvent dans le cerveau ou dans le foie ? J'ignore absolument si nous serons obligés d'accepter ces apparentes absurdités lorsque nous en saurons assez sur la nature physique de la conscience pour répondre à ces questions. Ce dont je suis *persuadé*, c'est que ce sont là des *questions scientifiques* auxquelles nous pourrons un jour apporter une réponse, quelle que soit la distance qui nous sépare aujourd'hui de cette réponse.

On affirme parfois, en se fondant sur des considérations philosophiques, qu'il n'existe aucun moyen de savoir si une entité autre que soi-même, *quelle qu'elle soit*, jouit d'une connaissance consciente — sans parler d'une paramécie. À mon avis, une telle attitude est trop étroite et trop pessimiste. Lorsqu'on cherche à établir la présence d'une qualité physique dans un objet, on cherche somme toute à l'établir avec une *certitude absolue*. Je ne vois pas pourquoi nous ne parviendrions pas un jour à répondre à des questions sur la présence de connaissance consciente avec une certitude égale à celle avec laquelle les astronomes fondent leurs affirmations sur des corps célestes situés à de nombreuses années-lumière de la Terre. Il n'y a pas si longtemps, certaines personnes affirmaient qu'on ne pourrait jamais connaître la composition matérielle du Soleil et des étoiles, pas plus qu'on ne pourrait voir la face cachée de la Lune. Pourtant, aujourd'hui, la Lune entière est remarquablement cartographiée (grâce aux sondes spatiales) et l'on connaît avec force détails la composition du Soleil (grâce à l'observation des lignes spectrales présentes dans la lumière solaire et aux simulations détaillées de son activité interne). On connaît également en détail et avec une bonne précision la composition des étoiles. Et l'on connaît même très bien, à certains égards, la composition générale de l'Univers au tout début de son existence (voir la fin de la section 4.5).

Mais en l'absence de tout soubassement théorique, les idées sur la conscience ne sont encore que des conjectures. Pour ma part, je suis fermement convaincu que sur cette planète, la conscience *n'est pas* une exclusivité humaine. Dans l'un de ses programmes télévisés les plus émouvants, David Attenborough[3] a montré une scène face à laquelle on peut difficilement s'empêcher de penser que les éléphants, par exemple, non seulement éprouvent des sentiments violents, mais aussi que ces sentiments ne sont pas très éloignés des croyances religieuses chez les êtres humains. Le chef d'un troupeau — une femelle dont la sœur était morte environ cinq ans plus tôt — fit faire un long détour à ce troupeau pour l'amener à l'endroit même où sa sœur avait disparu. Lorsqu'ils trouvèrent ses ossements, le chef ramassa son crâne avec une grande tendresse et les éléphants se le passèrent les uns aux autres en le caressant de leurs trompes. La présence de compréhension chez les éléphants est également attestée de manière convaincante — et horrible — par un autre programme de télévision[4]. Des images prises depuis un hélicoptère, engagé

dans une action pudiquement qualifiée d'« opération d'élimination des bêtes malades » montrent clairement, à travers les insoutenables cris d'angoisse poussés par les éléphants, que ces animaux avaient pleinement conscience du massacre imminent.

Il existe aussi des preuves tangibles d'une conscience (et d'une conscience de soi) chez les singes, et je ne doute pas que la conscience soit également présente dans des formes de vie animale considérablement moins « évoluées ». Par exemple, dans un autre programme télévisé[5] — sur l'agilité, la détermination et l'ingéniosité extraordinaires de (certains) écureuils —, j'ai été particulièrement frappé par une scène montrant un écureuil s'apercevant qu'en mordant le fil sur lequel il grimpait, il pouvait libérer le contenu d'un sac de noisettes situé à une certaine distance. On imagine difficilement qu'une telle idée ait pu être innée ou résulter d'une expérience antérieure. Pour évaluer les conséquences positives de son geste, l'écureuil avait nécessairement une compréhension rudimentaire de la *topologie* de la situation (*cf.* §1.19). Il me semble qu'on est là en présence d'un acte authentique d'*imagination* de la part de l'animal — qui exige sûrement la présence de conscience !

Il semble ne faire guère de doute que la conscience est une affaire de degré, qu'elle ne se réduit pas à « être là » ou « ne pas être là ». Même chez moi je constate qu'elle est, en fonction du temps, présente à des degrés divers (par exemple, lorsque je rêve, elle semble bien moins présente que lorsque je suis pleinement éveillé).

Jusqu'où peut-on alors descendre ? Sur ce point, les opinions sont très diverses. Pour ma part, j'ai toujours eu du mal à croire que les insectes jouissent de la moindre conscience depuis que j'ai vu un autre documentaire dans lequel un insecte en mangeait voracement un deuxième, en ignorant apparemment qu'il servait lui-même de repas à un troisième. A l'inverse, le comportement d'une fourmi est infiniment complexe et subtil (*cf.* §1.15). Peut-on croire que son système de contrôle, merveilleusement efficace, n'est pas secondé par le principe même, quel qu'il soit, auquel nous devons nos facultés de compréhension ? Ses cellules neuronales possèdent leur propre cytosquelette, et si ce cytosquelette contient des microtubules capables d'entretenir les états quantiques cohérents indispensables, selon moi, à notre propre conscience, ne doit-on pas alors en conclure que la fourmi jouit également de cette qualité insaisissable ? Si les microtubules de notre cerveau possèdent l'énorme sophistication nécessaire pour maintenir une activité collective quantique et cohérente, on comprend difficilement comment la sélection naturelle aurait façonné cette qualité exclusivement pour nous et (certains de) nos cousins multicellulaires. Ces états quantiques cohérents ont probablement constitué des structures précieuses pour les premiers animaux unicellulaires eucaryotes, bien qu'il soit tout à fait possible que ces structures aient eu pour eux un impact très différent de celui qu'elles ont eu pour nous.

Bien entendu, en elle-même, la cohérence quantique *n'implique pas* la conscience — sinon les supraconducteurs seraient conscients ! Pourtant, rien n'interdit qu'une telle cohérence soit un *élément* nécessaire à l'émergence de la conscience. Notre cerveau est une énorme organisation, et puisque la

conscience semble être une caractéristique très *globale* de notre pensée, il semble que nous devions rechercher un type de cohérence qui se manifesterait à une échelle bien plus grande que celle des microtubules, voire des cytosquelettes, individuels. Il y a probablement, entre les états des cytosquelettes d'un grand nombre de neurones distincts, des emmêlements quantiques importants qui englobent de vastes régions du cerveau dans un même état quantique collectif. Mais il nous faut plus. Pour qu'une action *non calculable* adéquate puisse intervenir — action que je considère comme un élément essentiel de la conscience — il faudrait que le système puisse tirer spécifiquement parti des aspects authentiquement *non aléatoires* (non calculables) de **RO**. La proposition que j'ai émise à la section 6.12 nous donne au moins une idée des *échelles* sur lesquelles une action **RO** précise et non calculable pourrait intervenir.

Ainsi, les idées émises dans ce livre donnent au moins un moyen d'*envisager* un niveau auquel la connaissance consciente commencerait à se manifester. Selon moi, les processus admettant une description en termes d'une physique calculable (ou aléatoire) n'entraînent pas la présence de conscience. Mais d'autre part, une action précise non calculable de type **RO**, même intervenant de façon cruciale, n'*impliquerait* pas nécessairement, en elle-même, une telle présence — bien que, selon mon point de vue, elle serait une *condition préalable* à la manifestation de la conscience. Ce n'est certes pas là un critère très précis, mais c'est le meilleur que je puisse trouver pour l'instant. Voyons jusqu'où il peut nous mener.

Quelles conséquences peut-on déduire des suggestions émises à la section 6.12 sur la localisation de la frontière quantique/classique — et des spéculations biologiques des sections 7.5 à 7.7 selon lesquelles cette frontière aurait un rapport avec l'interface interne/externe des systèmes microtubulaires d'une cellule ou d'un ensemble de cellules ? Une idée supplémentaire importante est que si la réduction du vecteur d'état ne survient que lorsqu'une trop grande partie de l'environnement se trouve emmêlée avec le système considéré, alors **RO** se présente effectivement comme un processus *aléatoire* auquel s'appliquent les raisonnements standard EP (résumés à la section 6.6) et se comporte exactement comme **R**. Ce qu'il faut, c'est que cette réduction survienne juste au moment où les aspects non calculables (et inconnus) de **RO** entrent en jeu. Bien que les détails de cette théorie soient encore inconnus, nous pouvons, au moins en principe, nous faire une idée du niveau auquel elle devrait commencer à intervenir. Ainsi, afin que les aspects non calculables de **RO** jouent un rôle important, il faudrait qu'une forme de cohérence quantique se maintienne jusqu'à ce qu'elle déclenche *juste* ce qu'il faut de mouvement matériel pour que **RO** agisse *avant* que l'environnement non emmêlé devienne significativement emmêlé.

Quelles conclusions en tire-t-on au niveau des microtubules ? L'*intérieur* des tubes serait le siège d'« oscillations quantiques cohérentes » faiblement couplées à l'activité calculable de type « automate cellulaire » associée aux sauts de conformation des tubulines situées *sur* les tubes. Tant que ces oscillations quantiques resteraient isolées, le couplage serait trop faible pour déclencher **RO**. Toutefois, ce couplage engloberait progressivement les tubulines et

déclencherait, au-delà d'un certain seuil, la procédure **RO**. Ce qu'il nous faut, c'est que **RO** intervienne *avant* que l'environnement des microtubules ne s'emmêle à l'état quantique, car l'implication des microtubules dans l'emmêlement quantique entraîne immédiatement la perte des aspects non calculables de **RO**, qui s'identifie alors à la procédure aléatoire **R**.

Ainsi, on peut se demander si dans une simple cellule (par exemple une paramécie ou une cellule de foie humain), l'activité conformationnelle des tubulines peut impliquer un déplacement de masse suffisamment important pour que le critère de la section 6.12 soit satisfait et donc que **RO** intervienne précisément à ce stade — comme il le faudrait —, ou si cette activité est insuffisante, de sorte que l'action de **RO** est reportée jusqu'à ce que l'environnement soit *effectivement* perturbé — auquel cas il n'y a pas de phénomène non calculable. À première vue, cette activité semble donner naissance à un déplacement de masse insuffisant, empêchant **RO** d'entrer en scène à ce niveau. Mais si l'on considère de grands ensembles de cellules, la situation devrait apparaître bien plus prometteuse.

Peut-être les éléments dont nous disposons jusqu'ici indiquent-ils effectivement que l'activité non calculable nécessaire à la manifestation de la conscience ne peut survenir qu'au sein d'un vaste ensemble de cellules — par exemple, dans un cerveau de taille appréciable[6]. Il faut bien sûr se garder pour l'instant de tirer des conclusions aussi tranchées. La description de la physique et de la biologie des mécanismes que je propose est encore trop sommaire, et il est clair que l'on doit procéder à des recherches plus approfondies avant de pouvoir se prononcer raisonnablement sur le niveau où intervient la conscience.

D'autres questions méritent également d'être considérées. Par exemple, quelle est la proportion du cerveau mise en jeu dans un état conscient ? Il est très probable que cela ne fait pas intervenir tout le cerveau. Une bonne part de l'activité cérébrale semble en fait inconsciente. Très curieusement, le cervelet (*cf.* §1.14) paraît correspondre à une activité entièrement inconsciente. Il contrôle de manière précise et délicate les gestes que nous effectuons — aux instants où nous les effectuons inconsciemment (*cf.* EOLP, p. 406-412). À cause de cette activité entièrement inconsciente, le cervelet est souvent qualifié de « simple ordinateur ». Il serait certainement très instructif de connaître les différences éventuelles entre l'organisation cellulaire ou cytosquelettique du cervelet et celle du cerveau, puisque c'est avec cette dernière structure que la conscience semble être le plus intimement liée. Il est intéressant de constater qu'au niveau du nombre de neurones, ces deux structures diffèrent peu : il n'y a que deux fois plus de neurones dans le cerveau et, d'une manière générale, bien plus de liaisons synaptiques entre les cellules individuelles du cervelet (*cf.* §1.4, Fig. 1.6). Il existe donc un phénomène plus subtil, dont ne rend pas compte le simple comptage des neurones*.

* Le profane en neuroanatomie que je suis ne peut qu'être frappé par le fait que l'organisation cérébrale présente une bizarrerie (inexpliquée ?) que ne me semble pas exhiber le cervelet. La plupart des nerfs moteurs et sensoriels se croisent et font que le côté gauche du cerveau est

Peut-être serait-il instructif également d'étudier la manière dont s'effectue l'« apprentissage » du contrôle cérébelleux inconscient à partir du contrôle cérébral conscient. Selon la philosophie connexionniste, les procédures d'apprentissage du cervelet seraient très semblables à celles des réseaux de neurones formels. Mais même si c'est le cas, et même s'il est *également* vrai que certaines activités *cérébrales* peuvent s'expliquer (partiellement) ainsi — le modèle connexionniste du fonctionnement du cortex visuel repose implicitement sur cette idée[7] —, rien n'autorise à penser qu'il en aille de même pour les aspects de l'activité cérébrale intervenant au niveau de la conscience. En fait, ainsi que je l'ai vigoureusement affirmé dans la première partie, les fonctions cognitives supérieures dans lesquelles intervient la la conscience sont nécessairement très différentes du modèle connexionniste.

8.7 Trois mondes et trois mystères

Quelle pourrait être la synthèse des idées exposées dans ce livre ? Le problème central auquel j'ai tenté de répondre tout au long de ces pages est celui de l'intégration du phénomène de la conscience dans la vision scientifique que nous avons aujourd'hui du monde. Je le reconnais volontiers, je n'ai pas eu grand-chose à dire sur le problème général de la conscience. Au lieu de cela, je me suis concentré, dans la première partie, sur une seule qualité mentale : la *compréhension consciente*, et en particulier, la compréhension mathématique. L'examen de cette qualité mentale m'a permis d'énoncer cette affirmation maîtresse : il est *impossible* qu'une telle qualité puisse résulter d'une simple activité algorithmique, tout comme il est impossible qu'elle puisse être correctement simulée par un algorithme — et je tiens à souligner que pour moi la compréhension *mathématique* ne jouit sur ce plan d'aucun statut particulier par rapport à toute autre forme de compréhension. La conclusion à laquelle j'ai abouti est que quelle que soit la forme d'activité cérébrale qui est responsable de la conscience (du moins dans cette manifestation particulière), elle dépend d'une physique irréductible à toute simulation algorithmique. Dans la deuxième partie, j'ai tenté de trouver, dans le cadre actuel de la science, une action physique qui pourrait effectivement dépasser les limites imposées par les calculs algorithmiques. Afin de décrire les problèmes profonds auxquels

---

principalement associé à la partie droite du corps — et *vice versa*. En outre, la région liée à la vision se situe tout à l'arrière du cerveau, alors que les yeux sont devant ; la région associée aux pieds se trouve au sommet du cerveau ; celle associée à l'ouïe d'une oreille est diamétralement opposée à cette oreille. Ce n'est pas une caractéristique universelle du cerveau, mais je ne peux m'empêcher de penser que ce n'est pas non plus un hasard. Le cervelet, lui, ne possède pas une telle organisation. Se pourrait-il que la conscience tire un avantage du fait que les signaux neuronaux ont une grande distance à parcourir ?

nous sommes confrontés, je formulerai les choses en termes de trois mondes différents et des trois mystères qui les relient les uns aux autres. Ces mondes ne sont pas sans rappeler ceux de Popper (*cf.* Popper et Eccles 1977), mais ma présentation est très différente.

Si le monde que nous connaissons le plus directement est le *monde de nos perceptions conscientes*, il est aussi celui sur lequel nous avons le moins de connaissances scientifiques précises. Ce monde contient la joie, la douleur et la perception des couleurs. Il contient nos tout premiers souvenirs d'enfance et notre peur de la mort. Il contient l'amour, la compréhension, la connaissance de nombreux faits, l'ignorance, la vengeance. Il contient les images mentales de chaises et de tables, ainsi que les odeurs, les sons et les sensations de toutes sortes qui se mêlent à nos pensées et à nos actes.

Il existe également deux autres mondes, moins directement accessibles que celui de nos perceptions, mais dont nous avons une connaissance bien plus étendue. L'un de ces monde s'appelle le *monde physique*. Il contient les chaises et les tables réelles, les téléviseurs, les automobiles, les êtres humains, les cerveaux humains et l'activité neuronale. Il contient le Soleil, la Lune et les étoiles. Il contient les nuages, les ouragans, les rochers, les fleurs et les papillons. À un niveau plus profond, il contient les molécules, les atomes, les électrons, les photons et l'espace-temps. Il contient également les cytosquelettes, les tubulines et les supraconducteurs. On ne s'explique pas clairement pourquoi le monde de nos perceptions aurait un lien avec le monde physique, mais apparemment, ce lien existe.

Il y a enfin un troisième monde, bien que nombre de gens aient du mal à accepter qu'il existe réellement. C'est le *monde platonicien des formes mathématiques*. C'est dans ce monde que l'on trouve les entiers naturels 0, 1, 2, 3, ... et l'algèbre des nombres complexes. On y trouve aussi le théorème de Lagrange (tout entier naturel est la somme de quatre carrés). On y trouve le théorème de Pythagore de la géométrie euclidienne (sur les carrés des côtés d'un triangle rectangle). On y trouve l'énoncé selon lequel tout couple d'entiers naturels vérifie la relation $a \times b = b \times a$. C'est dans ce même monde platonicien que l'on trouve le fait que ce dernier résultat n'est pas valable pour certains types de « nombres » (tels ceux qui interviennent dans le produit grassmannien de la section 5.15). Ce même monde platonicien contient des géométries autres que la géométrie euclidienne, dans lesquelles le théorème de Pythagore perd sa validité. Il contient les nombres infinis, les nombres non calculables, les ordinaux dénombrables et les ordinaux non dénombrables. Il contient les actions de machines de Turing qui ne s'arrêtent pas, ainsi que les machines-oracles. Il contient de nombreuses familles de problèmes mathématiques insolubles algorithmiquement, tel le problème du pavage par des polyminos. C'est aussi dans ce monde que se trouvent les équations électromagnétiques de Maxwell, les équations gravitationnelles d'Einstein et les innombrables espaces-temps théoriques qui les vérifient — qu'ils soient ou non physiquement réalistes. Il contient les simulations mathématiques de chaises et de tables, telles qu'on les utiliserait dans une « réalité virtuelle », ainsi que les simulations des trous noirs et des ouragans.

De quel droit disons-nous que ce monde platonicien est réellement un « monde » qui « existe » tout autant que les deux autres ? Peut-être le lecteur considère-t-il que ce monde n'est qu'un bric-à-brac de concepts abstraits que les mathématiciens ressortent de temps en temps. Pourtant, son existence repose sur la nature profonde, intemporelle et universelle de ces concepts et sur le fait que leurs lois sont indépendantes de ceux qui les ont découvertes. Ce bric-à-brac — si c'en est réellement un — n'est pas une création humaine. Les entiers naturels existaient bien avant les êtres humains, voire bien avant toute créature terrestre, et continueront d'exister après que toute vie aura disparu. Il a toujours été vrai que chaque entier naturel est la somme de quatre carrés, et ce n'est pas avec Lagrange que cette propriété a commencé d'exister. Les entiers naturels supérieurs à tout nombre pouvant être imprimé par tout ordinateur concevable sont eux aussi la somme de quatre carrés, même si nous n'avons aucune chance de savoir ce que peuvent être ces carrés. Il sera toujours vrai qu'aucune procédure générale calculable ne permet de décider si l'action d'une machine de Turing s'arrête ou non, et cela a été vrai bien avant que Turing ne découvre la notion de calculabilité.

Nombre de personnes objectent toutefois que la nature absolue des vérités mathématiques ne justifie en rien que l'on attribue une « existence » aux concepts et aux vérités mathématiques. (J'ai parfois entendu dire que le platonisme mathématique est « démodé ». Certes Platon est mort il y a environ 2 340 ans, mais ce n'est pas une raison ! Une objection plus sérieuse est celle des philosophes qui trouvent difficile d'admettre qu'un monde entièrement abstrait puisse avoir une influence sur le monde physique. Ce profond problème est en fait l'un des mystères sur lesquels je vais revenir dans un instant.) En vérité, la réalité des concepts mathématiques apparaît bien plus naturelle aux mathématiciens qu'à ceux qui n'ont pas la chance de passer leur temps à explorer les merveilles et les mystères de ce monde. Pour le moment cependant, le lecteur n'aura pas à admettre que les concepts mathématiques forment bien un « monde » doté d'une réalité comparable à celles des mondes physique et mental. Peu importera pour l'instant la manière dont chacun appréhende les concepts mathématiques. Si vous le désirez, vous pouvez considérer l'expression « monde platonicien des formes mathématiques » comme une simple formule que nous emploierons uniquement pour faciliter nos descriptions. Lorsque nous aborderons les trois mystères liés aux trois « mondes », nous verrons qu'elle n'est somme toute pas si anodine.

Quels sont donc ces trois mystères ? Ils sont illustrés à la figure 8.1. Le premier est : Pourquoi des lois très précises et profondément mathématiques jouent-elles un rôle aussi important dans le comportement du monde physique ? On ne sait trop comment, le monde de la réalité physique semble émerger presque mystérieusement du monde platonicien des mathématiques. C'est ce que représente la flèche descendant vers la droite et reliant le monde platonicien au monde physique. Le deuxième mystère est : Comment des êtres doués de perception peuvent-ils naître du monde physique ? Comment des corps matériels subtilement organisés peuvent-ils faire apparaître des entités mentales à partir de leur substance matérielle ? Sur la figure 8.1, cette

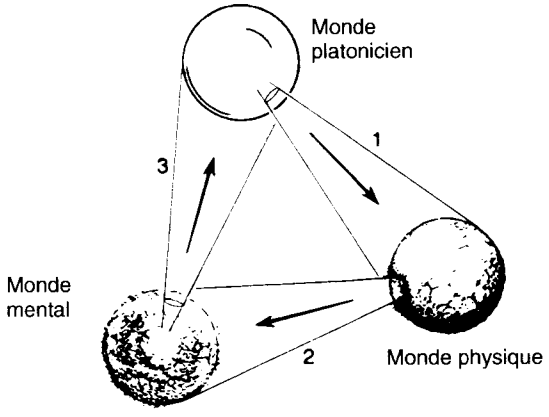


Figure 8.1. Chacun de ces trois mondes — le monde platonicien des formes mathématiques, le monde physique et le monde mental — semble mystérieusement émerger d'une — ou du moins être intimement lié à une — petite région de son prédécesseur.

interrogation est représentée par la flèche reliant en bas le monde physique au monde mental. Le troisième mystère enfin est : Comment l'activité mentale peut-elle apparemment « créer » des concepts mathématiques à partir d'un modèle mental ? Ces outils mentaux apparemment vagues, peu fiables et souvent inadéquats, dont notre monde mental semble être équipé, paraissent en effet capables (du moins pour les meilleurs d'entre eux) de faire naître des formes mathématiques abstraites et par là de permettre à notre esprit de pénétrer, grâce à la compréhension, dans le royaume mathématique platonicien. C'est ce qu'indique la flèche pointant vers le haut et reliant le monde mental au monde platonicien.

Platon s'intéressa énormément à la première de ces flèches (mais aussi, à sa manière, à la troisième), et il eut soin de distinguer entre une forme mathématique parfaite et son « ombre » imparfaite située dans le monde physique. Ainsi, un triangle mathématique (ou euclidien, ainsi que nous le désignons plus précisément aujourd'hui) avait la somme de ses angles exactement égale à deux angles droits, tandis qu'un triangle physique, par exemple construit en bois avec la plus grande précision possible, avait des angles dont la somme approchait certes deux angles droits, mais ne leur était pas exactement égale. Platon exposa ses idées à l'aide d'une parabole. Il imagina des hommes confinés dans une caverne et enchaînés de sorte qu'ils ne pouvaient voir les formes parfaites évoluant derrière eux et étaient réduits à contempler les ombres de ces formes que la lumière d'un feu projetait sur la paroi de la caverne. Tout ce qu'ils pouvaient directement observer de ces formes, c'étaient leurs ombres imparfaites passablement déformées par la danse des flammes. Les formes parfaites représentaient les formes mathématiques, et les ombres, le monde de la « réalité physique ».

Depuis l'époque de Platon, le rôle sous-jacent des mathématiques dans notre description de la structure et du comportement du monde physique s'est considérablement accru. En 1960, l'éminent physicien Eugene Wigner donna une conférence restée fameuse, intitulée « L'extravagante efficacité des mathématiques dans les sciences physiques » dans laquelle il évoquait la précision stupéfiante et l'adéquation subtile que les physiciens ne cessent de constater dans les mathématiques sophistiquées qui leur servent à décrire la réalité.

Selon moi, l'exemple le plus impressionnant est la relativité générale d'Einstein. Certains disent parfois que les physiciens ne font que remarquer, de temps en temps, des régularités dont le comportement se trouve très bien expliqué par des concepts mathématiques. On ajoute alors qu'ils ont tendance à orienter leurs intérêts vers les domaines où leurs descriptions mathématiques marchent bien, de sorte qu'il n'est pas surprenant que les mathématiques soient si bien adaptées à ces descriptions. Un tel point de vue m'apparaît bien éloigné de la vérité, pour la simple raison qu'il n'explique en rien la profonde unité existant entre mathématiques et mécanismes de l'Univers, unité qu'atteste, par exemple, la théorie d'Einstein. Lorsque Einstein publia sa théorie, il ne visait pas à résoudre des énigmes observationnelles. La théorie de la gravitation newtonienne fonctionnait depuis 250 ans et avait fait preuve d'une précision extraordinaire — de l'ordre de un dix-millionième (ce qui déjà justifiait plus qu'amplement que l'on prît au sérieux l'existence d'un fondement mathématique sous-jacent à la réalité physique). On avait certes observé une anomalie dans le mouvement de Mercure, mais ce n'était pas une raison suffisante pour renoncer à la théorie de Newton. S'interrogeant sur les fondements de la physique, Einstein pressentit néanmoins que l'on pouvait faire mieux en changeant le cadre même de cette théorie gravitationnelle. Dans les toutes premières années qui suivirent sa parution, peu d'effets confortaient la théorie d'Einstein et l'accroissement en précision qu'elle offrait était insignifiant par rapport à ce que donnait la théorie de Newton. Aujourd'hui toutefois, presque 80 ans plus tard, cette précision a pratiquement été multipliée par dix millions. Einstein ne fit pas que « remarquer des régularités » dans le comportement des corps physiques. Il découvrit une sous-structure mathématique profonde préexistante dans les mécanismes de l'Univers. D'autre part, il n'était pas à la recherche d'un phénomène physique qui permettrait de confirmer une bonne théorie. Il découvrit cette relation mathématique précise dans la structure même de l'espace-temps — le plus fondamental des concepts physiques.

Les autres théories sur les processus physiques fondamentaux ont toujours renfermé une structure mathématique sous-jacente qui s'est avérée non seulement extraordinairement précise, mais aussi mathématiquement sophistiquée. (Et de peur que le lecteur ne pense que le « renversement » de théories physiques plus anciennes, telle celle de Newton, ne les invalide définitivement, je m'empresse de préciser que *ce n'est pas le cas.* Les anciennes théories, lorsqu'elles sont suffisamment bonnes — telles celles de Galilée et de Newton —, survivent et conservent leur place au sein des nouvelles théories.) En outre, les mathématiques elles-mêmes puisent une grande inspiration dans

les données subtiles et imprévues que l'on recueille sur le comportement détaillé de la nature. Que ce soit la théorie quantique — dont le lien intime avec des mathématiques subtiles (*e.g.* les nombres complexes) est, je l'espère, apparu suffisamment dans les quelques pages que nous lui avons consacrées dans ce livre —, la relativité générale ou les équations électromagnétiques de Maxwell, toutes ces théories ont grandement stimulé le progrès des mathématiques. Mais cela n'est pas seulement vrai pour les théories récentes. Ce le fut au moins autant pour les théories plus anciennes telles que la mécanique newtonienne (qui a donné naissance au calcul différentiel) et l'analyse de la structure de l'espace faite pas les anciens Grecs (qui nous a donné le concept même de géométrie). On a souvent souligné l'extraordinaire précision des mathématiques lorsqu'on les applique à la description du comportement des corps physiques (que l'on songe à la précision à 11 ou 12 chiffres de l'électrodynamique quantique). Mais le mystère va plus loin que cela. Les concepts qui sous-tendent les processus physiques ont une profondeur, une subtilité et une *fécondité mathématique* remarquables. C'est là un point généralement peu connu — sauf des physiciens-mathématiciens.

Qu'il soit bien clair que cette fécondité mathématique — qui constitue une motivation précieuse pour les mathématiciens — ne correspond pas à une simple question de mode (bien que le rôle de cette dernière ne soit pas totalement négligeable). Les idées développées dans le seul but d'élargir notre compréhension des mécanismes du monde physique ont souvent éclairé de manière profonde et imprévue des problèmes mathématiques qui étaient *déjà* l'objet d'un intérêt considérable. L'un des exemples récents les plus frappants fut l'utilisation, par Simon Donaldson, d'Oxford, des théories de type Yang-Mills (qui avaient été élaborées par les physiciens pour expliquer mathématiquement les interactions entre des particules subatomiques). Ces idées lui permirent de démontrer des propriétés concernant les variétés à quatre dimensions[8] — propriétés qui attendaient d'être expliquées depuis de nombreuses années. En outre, de telles propriétés mathématiques, bien qu'elles ne soient pas souvent anticipées par les êtres humains avant qu'ils ne disposent d'intuitions appropriées, ont existé, intemporelles, dans le monde platonicien, sous forme de vérités immuables attendant d'être découvertes — par ceux qui avaient les bonnes idées et faisaient les bons raisonnements.

J'espère que le lecteur est convaincu de la relation intime et authentique — bien que profondément mystérieuse — unissant le monde mathématique platonicien à celui des objets physiques. J'espère également que l'existence même de cette relation encouragera les sceptiques à considérer ce monde comme un monde à part entière. Qui sait d'ailleurs si certains d'entre eux n'iront pas plus loin que moi sur cette voie. Peut-être doit-on attribuer une réalité platonicienne à d'autres concepts abstraits, qui ne sont pas forcément de nature mathématique. Platon lui-même attribuait aux concepts de « bien » ou de « beauté » (*cf.* §8.3) la même réalité que l'on reconnaît aux concepts mathématiques. Personnellement, je ne m'oppose pas à une telle extension, mais elle n'a pas jusqu'ici joué un rôle important dans les considérations que j'expose dans ce livre. Les problèmes d'éthique, de morale et d'esthétique n'intervien-

nent en rien dans mes discussions — mais ce n'est pas une raison pour ne pas les considérer comme aussi « réels » que ceux que j'ai examinés. À l'évidence, ce sont là des problèmes dignes d'intérêt, mais je n'ai pas eu à les aborder dans ce livre[9].

De même, je ne me suis guère attardé sur le mystère (flèche 1, descendant vers la droite sur la figure 8.1) du rôle sous-jacent que le monde mathématique platonicien joue dans le monde physique — j'ai été plus prolixe sur les deux autres mystères, encore bien moins compris. Les questions abordées dans la première partie portaient essentiellement sur la troisième flèche, à savoir le mystère de notre perception de la vérité mathématique, et plus précisément comment, grâce à la contemplation mathématique, nous semblons pouvoir « faire apparaître » ces formes mathématiques platoniciennes. Ces formes parfaites seraient alors de simples ombres de nos pensées imparfaites. Une telle vision du monde platonicien — monde uniquement produit par notre propre activité mentale — contredit totalement les conceptions de Platon. Pour lui, le monde des formes parfaites est un monde premier, intemporel et indépendant de notre existence. Dans la conception platonicienne, ma troisième flèche pointerait plutôt vers le bas que vers le haut et irait du monde des formes parfaites vers celui de notre esprit. Considérer le monde mathématique comme le produit de notre pensée revient à adopter un point de vue *kantien* et non celui, platonicien, auquel j'adhère ici.

De même, certains affirmeront peut-être qu'il faut inverser le sens de l'une de mes autres flèches — voire des deux. Peut-être l'évêque Berkeley aurait-il préféré que ma *deuxième* flèche aille du monde mental vers le monde physique, la « réalité physique » devenant alors une simple ombre de notre existence mentale. D'autres (les « nominalistes ») diront que je dois inverser ma *première* flèche, le monde des mathématiques n'étant qu'un simple reflet de la réalité physique. Mes propres convictions (cela devrait être évident à la lecture de ce livre) s'opposent fermement à l'inversion de ces deux premières flèches, et il est certainement tout aussi évident que je suis quelque peu embarrassé par l'orientation apparemment « kantienne » prise par la *troisième* flèche de la figure 8.1 ! Selon moi, le monde des formes parfaites est un monde premier (comme le pensait Platon) — son existence étant presque une nécessité logique — dont les *deux* autres mondes ne sont que les ombres.

Pour mettre un peu d'ordre dans la hiérarchie entre les trois mondes de la figure 8.1, je suggère de considérer les flèches sous un éclairage différent. Le point essentiel concernant ces flèches n'est pas tant leurs orientations que le fait que chacune d'elles établit une correspondance entre une *petite* région d'un monde et la *totalité* du monde suivant. En ce qui concerne ma première flèche, on m'a souvent fait remarquer que la plus grande partie du monde des mathématiques (telles qu'elles se définissent à travers l'activité des mathématiciens) semble n'avoir que peu de liens — si tant est qu'elle en ait — avec le comportement des corps physiques. Ainsi, seule une infime partie du monde platonicien sous-tend la structure de notre Univers physique. De même, ma seconde flèche exprime le fait que notre existence mentale n'émerge que d'une partie infime du monde physique — partie où sont réunies les conditions

nécessaires à l'apparition de la conscience, comme dans le cerveau humain. De même enfin, ma troisième flèche ne se rapporte qu'à une partie infime de notre activité mentale, celle qui s'occupe des problèmes de l'absolu et de l'intemporalité — et plus particulièrement de la vérité mathématique. La plupart du temps, notre existence mentale est requise par de tout autres sujets !

Ces correspondances semblent paradoxales dans la mesure où chaque monde paraît « émerger » d'une *infime* partie seulement du monde précédent. Toutefois, je considère que ces flèches se contentent de traduire les diverses correspondances, sans affirmer la moindre émergence « réelle » ; autrement dit, je ne préjuge en rien de la hiérarchie, si elle existe, entre ces trois mondes.

Pourtant, même ainsi, la figure 8.1 traduit un autre aspect de mes opinions (ou de mes préjugés). Ma description semble supposer que la *totalité* de chaque monde se réfléchit effectivement à l'intérieur d'une (petite) région de son prédécesseur. Peut-être mes préjugés sont-ils erronés. Peut-être existe-t-il des aspects du comportement du monde physique qui *échappent* à toute description en termes mathématiques ; peut-être existe-t-il une vie mentale qui *ne repose pas* sur des structures physiques (telles que le cerveau) ; peut-être existe-t-il des vérités mathématiques qui sont *fondamentalement inaccessibles* à la raison et à l'intuition humaines. Pour tenir compte de ces possibilités, il faudrait redessiner la figure 8.1 de manière que certains de ces mondes — voire tous — s'étendent au-delà de la correspondance donnée par la flèche.

Dans la première partie, je me suis particulièrement intéressé à certaines conséquences du célèbre théorème d'incomplétude de Gödel. Peut-être certains lecteurs croyaient-ils que ce théorème affirme l'existence, dans le monde platonicien des vérités mathématiques, de parties qui échappent fondamentalement à la compréhension et à l'intuition humaines. J'espère que mes arguments ont clairement démontré que *ce n'est pas* le cas[10]. Les propositions mathématiques particulières fournies par l'ingénieuse argumentation gödelienne sont accessibles au cerveau humain — à condition d'être construites à l'aide de systèmes mathématiques (formels) préalablement reconnus comme des moyens valides pour déterminer la vérité mathématique. L'argumentation gödelienne ne dit pas qu'il existe des vérités mathématiques inaccessibles, mais que l'intuition humaine n'est réductible ni à un raisonnement formel ni à des procédures calculables. En outre, elle affirme clairement l'existence du monde mathématique platonicien. La vérité mathématique n'est pas déterminée arbitrairement par les règles d'un système formel d'origine humaine ; elle a un caractère absolu, irréductible à tout système de règles algorithmiques. Cet a priori platonicien (en tant qu'il s'oppose au point de vue formaliste) fut à la base des motivations de Gödel. En revanche, les déductions faites à partir de son théorème illustrent la nature profondément mystérieuse de nos perceptions mathématiques. Ces perceptions ne résultent pas uniquement d'un « calcul » ; elles mettent fondamentalement en jeu un autre élément, qui n'existerait pas en l'absence de ce phénomène appelé conscience et qui est somme toute au centre du monde des perceptions.

La deuxième partie a surtout porté sur des questions liées à la deuxième flèche (bien que celles-ci ne puissent être véritablement abordées sans faire

référence à la première). Comment le monde physique donne-t-il naissance à la conscience ? Comment la conscience peut-elle émerger d'éléments en apparence aussi inertes que la matière, l'espace et le temps ? Nous n'avons pu répondre à ces questions, mais j'espère au moins que le lecteur se rend compte que la matière *elle-même* est mystérieuse, comme l'est l'espace-temps dans le cadre duquel les théories physiques évoluent aujourd'hui. Le fait est que nous sommes encore trop ignorants de la nature de la matière et des lois qui la gouvernent pour pouvoir comprendre l'organisation qui, dans le monde physique, donne naissance à la conscience. En outre, plus nous examinons en profondeur la nature de la matière, plus cette dernière nous apparaît insaisissable, mystérieuse et mathématique. Si l'on demande aux meilleures théories scientifiques actuelles ce qu'*est* la matière, on obtient une réponse formulée dans la langue des mathématiques, pas tant en termes d'un système d'équations (bien que celles-ci soient également importantes), qu'en termes de concepts mathématiques subtils exigeant de longues méditations avant d'être correctement compris.

Si la relativité générale d'Einstein a montré comment nos concepts mêmes d'espace et de temps se sont transformés pour devenir plus mystérieux et plus mathématiques, la mécanique quantique, elle, nous a montré, et à un degré encore plus marqué, que notre concept de *matière* avait subi un sort similaire. En outre, cette théorie a profondément modifié notre conception de la réalité. Comment la simple *possibilité* contrafactuelle d'un événement — comment un événement qui *ne se produit pas* réellement — peut-elle avoir une influence sur les événements qui se produisent *effectivement* ? Il y a, dans le mystère du fonctionnement de la mécanique quantique, quelque chose qui *semble* bien plus proche que ne l'est la physique classique du mystère de l'existence de l'esprit. Personnellement, je ne doute pas que lorsque nous disposerons de théories plus profondes, la place de l'esprit dans la théorie physique paraîtra moins incongrue qu'elle ne l'est aujourd'hui.

Aux sections 7.7 et 8.6, j'ai tenté de définir les circonstances physiques qui pourraient favoriser l'apparition de la conscience. Qu'il soit bien clair toutefois que je *ne considère pas* l'apparition de la conscience comme résultant simplement du mouvement d'une quantité précise de matière quantiquement cohérente — mouvement défini par une théorie **RO** de la frontière quantique/classique. Le lecteur l'aura probablement compris, un tel mouvement n'est qu'une condition préalable à l'apparition d'une action non calculable dans le cadre de notre représentation physique actuelle. La conscience authentique implique la connaissance immédiate d'une variété infinie de choses qualitativement différentes — telles la couleur verte d'une feuille, le parfum d'une rose, le chant d'un oiseau ou le contact caressant d'un chat ; mais aussi l'écoulement du temps, les émotions, les soucis, l'émerveillement et l'appréciation d'une idée. Elle implique l'espérance, les idéaux, les intentions, et la volonté réelle de l'accomplissement d'innombrables mouvements corporels concrétisant ces intentions. L'étude de la neuroanatomie, des troubles neurologiques, de la psychiatrie et de la psychologie nous en a beaucoup appris sur le détail des relations unissant la nature physique de notre cerveau à notre état mental.

Il ne fait aucun doute que nous pouvons interpréter ces relations en termes de comportement physique de quantités critiques de masse animées d'un mouvement quantique cohérent. Pourtant, si l'on ne recherche pas une nouvelle physique, on restera prisonnier d'une physique entièrement calculable, ou encore d'une physique calculable et aléatoire. Cette calculabilité ne laisse aucune place à une description physique de l'intentionnalité et de l'expérience subjective. Ce n'est qu'en s'en affranchissant que l'on peut espérer parvenir à une telle description.

Nombre de ceux qui partagent ce point de vue objecteront cependant que cette description est irrémédiablement hors de notre portée. À ceux qui pensent ainsi, je demande un peu de patience. La science n'est pas une discipline figée. Les mystérieux développements de la mécanique quantique contiennent déjà, selon moi, des indices montrant que le concept d'esprit est un peu plus proche de notre compréhension de l'Univers physique qu'il ne l'était auparavant — certes, pour l'instant, seulement *un petit peu* plus proche. Je pense qu'une fois que nous entreverrons la *nouvelle* théorie physique qui nous manque actuellement, ces indices deviendront bien plus convaincants. Mais la science a encore beaucoup de chemin à parcourir : *cela* est pour moi une certitude !

En outre, la possibilité même d'une compréhension de ces sujets par l'homme nous apprend quelque chose sur les aptitudes résultant de la possession de la conscience. À l'évidence, certains parmi nous, tels Newton ou Einstein, ou Archimède, Galilée, Maxwell, ou Dirac — ou Darwin, Leonard de Vinci, Rembrandt, Picasso, Bach, Mozart ou Platon, ou ces grands esprits qui purent concevoir *l'Iliade* ou *Hamlet* — semblent avoir particulièrement bénéficié de cette faculté de « sentir » la vérité ou la beauté. Mais chacun de nous est potentiellement en unisson avec les mécanismes de la nature, ce qui transparaît dans nos facultés mêmes de compréhension et de sensibilité conscientes, quel que soit le niveau auquel elles s'expriment. Chaque cerveau conscient humain est construit à partir de composants physiques subtils qui nous permettent de tirer avantage de l'organisation de notre Univers fondamentalement mathématique — de sorte qu'à notre tour, grâce à cette qualité platonicienne qu'est la « compréhension », nous pouvons percevoir directement, à de multiples niveaux, le comportement de notre Univers.

Ce sont là des questions profondes, et nous sommes très loin de pouvoir y répondre. Je pense qu'on ne pourra pas leur trouver de réponse claire si l'on refuse de prendre en compte les caractéristiques mutuellement dépendantes de *tous* ces mondes. Aucun de ces problèmes ne trouvera indépendamment sa solution. J'ai parlé de trois mondes et des mystères qui les unissent. Nul doute que ces trois mondes n'en sont en fait qu'*un seul*, monde dont nous ne pouvons aujourd'hui même pas entrevoir la véritable nature.

Épilogue

Jessica et son père émergèrent de la grotte. Il faisait maintenant tout à fait nuit, et le silence régnait. Plusieurs étoiles brillaient déjà dans le ciel. Jessica dit à son père :

« Tu sais, Papa, quand je regarde le ciel, j'ai de la peine à croire que la Terre soit *vraiment* en mouvement — qu'elle tourne sur elle-même à ces milliers de kilomètres à l'heure —, même si je *sais* que c'est vrai. »

Elle s'arrêta et fixa le ciel un instant.

« Papa, parle-moi des étoiles... »

Notes

Chapitre 1

1. Voir en particulier Good (1965), Minsky (1986), Moravec (1988).
2. Moravec (1988) fonde son argumentation en faveur de ce type d'échelle temporelle sur la proportion de cortex dont il considère qu'elle est déjà correctement modélisée (essentiellement celle située dans la rétine), ainsi que sur une estimation du rythme auquel la technologie informatique progressera dans l'avenir. Ses estimations du début 1994 sont inchangées par rapport à 1988 ; *cf.* Moravec (1994).
3. Ces quatre points de vue sont explicitement décrits dans (par exemple) Johnson-Laird (1987), p. 252 (bien que ce qu'il appelle « thèse de Church-Turing » soit essentiellement ce que je dénomme à la section 1.6 « thèse de Turing » plutôt que « thèse de Church »).
4. Voir par exemple Searle (1980), Lockwood (1989).
5. Voir Moravec (1988).
6. Turing (1950) ; voir EOLP, p. 6-11.
7. Voir Searle (1980, 1992).
8. La complexité de ce problème réside dans le fait que la physique actuelle dépend de l'utilisation d'une action *continue* plutôt que discrète (numérique). Même la *signification* du terme « calculabilité » est, dans ce contexte, sujette à diverses interprétations. Pour des discussions sur ce sujet, voir Pour-El (1974), Smith et Stephenson (1975), Pour-El et Richards (1979, 1981, 1982, 1989), Blum, Schub et Smale (1989), Rubel (1988, 1989). Je reviendrai sur ce sujet à la section 1.8.
9. Je dois cette jolie formule au présentateur de l'émission « Thought for the Day », sur BBC Radio 4.
10. L'IA vit effectivement le jour dans les années cinquante en utilisant des procédures descendantes relativement élémentaires (*e.g.* Grey Walter

1953). Le « perceptron » introduit par Frank Rosenblatt (1959, 1962) en reconnaissance des formes fut le premier dispositif « connexionniste » (à réseaux de neurones formels) réussi ; il déclencha un vif intérêt pour les procédures ascendantes. Toutefois, en 1969, Marvin Minsky et Seymour Papert (*cf.* Minsky et Papert 1972) mirent en évidence certaines limitations fondamentales associées à ces procédures. Elles devaient être surmontées par Hopfield (1982), et la recherche sur les dispositifs artificiels de type réseau neuronal fait aujourd'hui dans le monde entier l'objet d'une activité considérable. (Voir par exemple Beks et Hamker 1992 et Gernoth *et al.* 1993 pour quelques applications à la physique des hautes énergies.) Les articles de John McCarthy (1979) et de Alan Newell et Herbert Simon (1976) sont des jalons importants en recherche sur la stratégie descendante en IA. Voir Freedman (1994) pour un récit passionnant de toute cette histoire. Pour d'autres discussions récentes sur les procédures et les perpectives de l'IA, voir Grossberg (1987), Baars (1988) ; pour une approche classique du sujet, voir Dreyfus (1972) ; et pour un point de vue récent par un pionnier de l'IA, voir Gelernter (1994) ; *cf.* aussi divers articles dans Broadbent (1993) et Khalfa (1994).

11. Pour un exposé sur le λ-calcul, voir Church (1941) et Kleene (1952).
12. Pour diverses publications concernant ces problèmes, voir par exemple Pour-El (1974), Smith et Stephenson (1975), Pour-El et Richards (1989), Blumb, Schub et Smale (1989). Le problème du lien entre l'activité cérébrale et ces sujets a été notamment examiné par Rubel(1985).
13. En ce qui concerne le problème du pavage, ce que Robert Berger a en fait démontré est que le pavage pour les pavés de *Wang* n'admet aucune solution algorithmique générale. Les pavés de Wang (ainsi nommés en l'honneur du logicien Hao Wang) sont de simples pavés carrés aux bords colorés ; les couleurs doivent être identiques d'un pavé à l'autre, les pavés ne pouvant subir ni rotation ni réflexion. Il est toutefois facile de concevoir, pour tout ensemble de pavés de Wang, un ensemble correspondant de polyminos pavant le plan si et seulement si l'ensemble de pavés de Wang considéré pave lui aussi le plan. Ainsi, l'insolubilité algorithmique du pavage par les polyminos découle immédiatement de celle des pavés de Wang.

    Il faut souligner, en relation avec le problème du pavage par des polyminos, que si un ensemble de polyminos donné *ne pave pas* le plan, ce fait *peut* être algorithmiquement vérifié (comme lorsque l'action d'une machine de Turing s'arrête, ou qu'un ensemble d'équations diophantiennes possède une solution), car lorsqu'on tente de couvrir un domaine carré $n \times n$ avec les pavés, pour des valeurs de plus en plus grandes de n, l'impossibilité du pavage du plan entier se manifeste pour une valeur *finie* de n. Ce que l'on ne peut vérifier algorithmiquement, ce sont les situations pour lesquelles les pavés *pavent* le plan.
14. Voir Freedman (1994) pour un compte rendu sur certaines des premières aspirations outrancièrement optimistes de l'IA.

15. Je suis reconnaissant à diverses personnes, en particulier à Lee Loevinger, de m'avoir familiarisé avec ces problèmes. Voir Hodgson (1991) pour une remarquable discussion sur la pertinence de la physique moderne et des algorithmes par rapport à notre comportement.
16. Voir par exemple Smithers (1990).
17. Par exemple, Sloman (1992) me reproche de faire trop de cas, dans EOLP, du terme mal défini de « conscience », alors que lui-même utilise allègrement le terme « esprit » (selon moi) encore moins bien défini !
18. Searle (1980, 1982).
19. Voir l'article de Searle (1980) dans Hofstadter et Dennett (1981), p. 372. Je me demande cependant si Searle ne privilégie pas davantage $\mathcal{B}$ que $\mathcal{C}$.
20. Voir Hofstadter (1981) pour une présentation divertissante d'une suggestion de cette nature ; *cf.* aussi EOLP, p. 23-24.
21. Pour un compte rendu accessible de la notion de « complexité algorithmique », voir Chaitin (1975).
22. Voir Hsu *et al.* (1990).
23. Voir Freedman (1994).
24. Voir par exemple Moravec (1994).
25. Voir Putnam (1960), Smart (1961), Benacerraf (1967), Good (1967, 1969), Lewis (1969, 1989), Hofstadter (1981), Bowie (1982) en relation avec les arguments de Lucas ; voir aussi Lucas (1970). Ma propre version, telle qu'elle est brièvement exposée dans EOLP, p. 453-455, a été attaquée dans diverses revues ; *cf.* en particulier Sloman (1992) et de nombreux commentateurs dans *Behavioral and Brain Sciences* : Boolos (1990), Butterfield (1990), Chalmers (1990), Davis (1990, 1993), Dennett (1990), Doyle (1990), Glymore et Kelly (1990), Hodgkin et Houston (1990), Kentridge (1990), MacLennan (1990), McDermott (1990), Manaster-Ramer *et al.* (1990), Mortensen (1990), Perlis (1990), Roskies (1990), Tsotsos (1990), Wilensky (1990) ; voir également mes propres réponses : Penrose (1990, 1993*d*), et aussi Guccione (1993) ; voir aussi Dodd (1991), Penrose (1991*b*).
26. D'une émission de la télévision anglaise — probablement *The Dream Machine* (décembre 1991), quatrième partie de la série BBC intitulée *The Thinking Machine.* Voir aussi Freedman (1994) pour une discussion des progrès récents accomplis en « compréhension » IA, concernant particulièrement l'intrigant projet « Cyc » de Douglas Lenat.
27. Pour un compte rendu populaire et vivant, voir Woolley (1992).
28. Par exemple, une telle suggestion a été avancée par Richard Dawkins dans ses conférences de Noël 1992 sur la BBC.
29. Voir par exemple le compte rendu de Freedman (1994) sur les travaux accomplis dans cette direction par Lenat et d'autres chercheurs.

Chapitre 2

1. Cela pourrait sembler parfaitement « évident » — et ne devrait certainement pas faire l'objet de désaccords entre mathématiciens ! Pourtant, le problème surgit avec la notion d'« existence » appliquée aux grands ensembles infinis. (Voir par exemple Smorynski 1975, Rucker 1984, Moore 1990.) Nous avons vu, à travers l'exemple du paradoxe de Russell, que l'on doit être particulièrement prudent sur ces sujets. Selon un certain point de vue, un ensemble n'existe réellement que si l'on dispose d'au moins une *règle* précise (pas nécessairement calculable) spécifiant les éléments qui lui appartiennent et ceux qui ne lui appartiennent pas. C'est précisément ce que *ne fournit pas* l'axiome du choix, car il n'existe aucune règle spécifiant l'élément que l'on doit prendre dans chaque ensemble de la collection. (Certaines conséquences de l'axiome du choix sont très peu intuitives — et presque paradoxales. Peut-être est-ce là une des raisons qui expliquent le désaccord. Je ne suis même pas totalement sûr d'avoir une position claire sur ce sujet !)
2. Dans le dernier chapitre de son livre (1966), Cohen souligne que bien qu'il ait montré que l'hypothèse du continu est INDÉCIDABLE dans le cadre des procédures ZF, il n'a pas abordé la question de savoir si elle est ou non réellement *vraie* — puis examine comment on pourrait effectivement répondre à cette question ! Cela me semble clairement démontrer qu'il *ne considère pas* que l'acceptation ou le refus de l'hypothèse du continu soit une affaire personnelle. Cette attitude s'oppose aux opinions souvent exprimées sur les conséquences des résultats de Gödel-Cohen, à savoir qu'il existe de nombreuses « théories des ensembles » possibles également « valides » d'un point de vue mathématique. À travers ces remarques, Cohen se révèle être, comme Gödel, un vrai platonicien, pour qui les vérités mathématiques sont des entités *absolues* et non pas arbitraires. Cette position est totalement en accord avec mes propres vues (*cf.* §8.7).
3. Voir par exemple Hofstadter (1981), Bowie (1982).
4. Par exemple, voir divers commentaires dans *Behavioral and Brain Sciences*, **13**, 1990, p. 643-705.
5. Cette terminologie fut suggérée par Hofstadter (1981). C'est l'« autre » théorème de Gödel — son théorème de *complétude* — qui nous dit que de tels modèles non standard existent toujours.
6. En fait, cela dépend des énoncés que l'on considère comme appartenant à ce qui est appelé ici « géométrie euclidienne ». Dans la terminologie habituelle des logiciens, la « géométrie euclidienne » n'inclut que certains types particuliers d'énoncés, et il s'avère que la vérité ou la fausseté de ces énoncés peut s'obtenir à l'aide d'une procédure algorithmique — d'où l'affirmation selon laquelle la géométrie euclidienne peut se formuler dans le cadre d'un système formel. Toutefois, selon d'*autres* interprétations, l'« arithmétique » ordinaire ferait également partie de la

« géométrie euclidienne », de sorte que certaines classes d'énoncés deviennent *insolubles* algorithmiquement. On rencontrerait la même situation si l'on considérait que le problème du pavage par les polyminos fait partie de la géométrie euclidienne — ce qui paraîtrait naturel. En ce sens, pas plus que l'arithmétique, la géométrie euclidienne n'est formulable dans un système formel !

7. Voir le commentaire de Davis (1993).
8. Voir aussi Kreisel (1960, 1967), Good (1967).
9. Voir Freedman (1994) pour certains des problèmes rencontrés par les ordinateurs lorsqu'ils tentèrent d'élaborer leurs « propres » mathématiques. En général, ces machines ne vont pas très loin. Elles ont besoin d'être guidées par l'homme !

Chapitre 3

1. Cette citation est extraite de Rucker (1984) et Wang (1987). Elle apparaît dans la Conférence Gibbs donnée par Gödel en 1951, dont le texte intégral paraîtra dans les œuvres complètes de Gödel, volume 3 (1995). Voir aussi Wang (1993), p. 118.
2. Voir Hodges (1983), p. 361. Cette citation provient de la conférence donnée en 1947 par Turing devant la London Mathematical Society et reproduite dans Turing (1986).
3. La procédure consiste à intégrer $\mathbb{ZF}$ dans le système de Gödel-Bernays ; voir Cohen (1966), chapitre 2.
4. Voir Hallett (1984), p. 74.
5. Ce nombre d'états d'univers — de l'ordre de $10^{10^{123}}$ — est le volume de l'espace des phases — mesuré dans les unités absolues définies à la section 6.11 — accessible à un univers qui contiendrait la quantité de matière présente dans notre Univers observable. On peut évaluer ce volume à l'aide de la formule de Bekenstein-Hawking pour l'entropie d'un trou noir de masse totale égale à celle observable pour l'univers et en prenant l'exponentielle de cette entropie — en unités absolues. Voir EOLP, p. 371-377.
6. Voir Moravec (1988, 1994).
7. Voir par exemple Eccles (1973) et EOLP, chapitre 9.
8. Voir Gleick (1987) et Schroeder (1991) pour une approche grand public de cette activité.
9. C'est là un ingrédient de la théorie classique de von Neumann et Morgenstern (1944).
10. Voir Gleick (1987), Schroeder (1991).
11. Voir Smorynski (1975, 1983) et Rucker (1984) pour une présentation grand public.

12. C'est là un théorème tout à fait fascinant (et pas trop compliqué) de géométrie euclidienne plane qui est remarquablement difficile à démontrer de manière directe. Il s'avère que l'une de ses démonstrations consiste à en trouver une généralisation appropriée bien plus facile à traiter, puis à déduire le résultat original en tant que cas particulier. Ce type de procédure est très courant en mathématiques, mais ce n'est pas du tout ainsi que procéderait normalement un ordinateur, car la recherche de la généralisation exige une ingéniosité et une intuition considérables. Pour une démonstration algorithmique, en revanche, il faudrait pourvoir l'ordinateur d'un système bien défini de règles descendantes qu'il appliquerait alors implacablement à une vitesse incroyable. Toutefois, la conception de ces règles exigerait une ingéniosité humaine considérable.
13. Voir Freedman (1994) pour un historique de certaines de ces tentatives.
14. Cette affirmation devrait être compatible avec la discussion de la section 1.8 ; elle est en accord avec l'hypothèse habituelle selon laquelle les systèmes analogiques admettent un traitement par des méthodes analogiques. Voir les références à la note 12 du chapitre 1.
15. L'idée que les neurones ne seraient pas les simples commutateurs binaires que l'on croyait autrefois qu'ils étaient semble faire l'objet d'un consensus de plus en plus large parmi les spécialistes. Voir par exemple les livres de Scott (1977), Hameroff (1987), Edelman (1989) et Pribram (1991). Nous le verrons au chapitre 7, certaines idées de Hameroff vont revêtir pour nous une importance cruciale.
16. Fröhlich (1968, 1970, 1975, 1984, 1986) ; ces idées ont été exploitées notamment par Marshall (1989), Lockwood (1989), Zohar (1990). Elles auront également une certaine importance pour nous (*cf.* §7.5). Voir aussi Beck et Eccles (1992).
17. Voir par exemple Smith et Stephenson (1975), Pour-El et Richards (1989), Blum *et al.* (1989) et Rubel (1989).
18. On trouvera dans Gardner (1970), Poundstone (1985) et Young (1990) de bonnes présentations du « jeu de la vie » de Conway.
19. Voir par exemple Johnson-Laird (1983), Broadbent (1993).
20. Discuté dans Broadbent (1993).

Chapitre 4

1. Voir par exemple Dennett (1991), p. 49.
2. Une des ces importantes équations est le « premier principe de la thermodynamique » : $\mathrm{d}E = T\mathrm{d}S - p\mathrm{d}V$, où E, T, S, p et V sont respectivement l'énergie, la température, l'entropie, la pression et le volume d'un gaz.
3. Voir par exemple Dennett (1991).

4. Sakharov (1967) ; voir Misner *et al.* (1973), p. 428.
5. Pour une description vivante mais pas très détaillée de la deuxième loi, voir EOLP, chapitre 6. Pour une présentation plus sophistiquée, voir, par ordre de difficulté, Davies (1974) et O. Penrose (1970).

Chapitre 5

1. Penrose (1993*b*, 1994*a*), Zimba et Penrose (1993).
2. La première idée d'une expérience précise est due à Clauser et Horne (1974) et Clauser, Horne et Shomony (1978).
3. Les premiers indices expérimentaux d'une confirmation positive de la non-localité des prédictions quantiques sont dus à Freedman et Clauser (1972) ; ils furent suivis quelques années plus tard par les résultats bien plus définitifs d'Aspect, Grangier et Roger (1982) ; voir aussi Aspect et Grangier (1986).
4. Il existe un autre type d'explication « classique » des effets EPR observés jusqu'ici par Aspect et d'autres. Cette explication — *l'effondrement retardé* — est due à Euan Squires (1992*a*) et repose sur le fait qu'il y aurait un retard significatif dans l'accomplissement réel d'une mesure par les détecteurs situés aux deux endroits concernés. Cette suggestion s'inscrit dans le cadre d'une théorie — nécessairement non conventionnelle, telles celles que nous rencontrerons aux sections 6.9 ou 6.12 — qui donne des prédictions précises sur les instants probables auxquels chacune des deux mesures quantiques a *objectivement* lieu. À cause des influences aléatoires contrôlant ces deux instants, l'un des détecteurs effectuerait sa mesure significativement plus tôt que l'autre — si tôt, en fait, que (dans les expériences réalisées jusqu'ici) un signal aurait suffisamment de temps pour quitter le premier détecteur à la vitesse de la lumière et aller informer le second du résultat de la première mesure.
   Ainsi, la réalisation de toute mesure s'accompagnerait d'une « onde d'information » voyageant à la vitesse de la lumière dans toutes les directions à partir de l'occurrence de la mesure. Cette situation est parfaitement compatible avec la théorie de la relativité classique (*cf.* §4.4), mais serait en désaccord avec les explications de la théorie quantique pour des distances suffisamment grandes. En particulier, les « dodécaèdres magiques » de la section 5.3 ne s'expliquent pas en termes d'effondrement retardé. Bien sûr, aucune « expérience » de ce type n'a encore été effectuée et on pourrait penser qu'elle contredirait les prédictions de la théorie quantique. Il existe toutefois une objection plus sérieuse : l'effondrement retardé rencontrerait de graves difficultés avec d'autres types de mesures quantiques et conduirait à une violation de toutes les lois de conservation standard. Par exemple, lorsqu'un atome radioactif émet une

particule chargée — disons une particule α —, deux détecteurs largement séparés pourraient recevoir la même particule α, violant ainsi simultanément les lois de conservation de l'énergie, de la charge électrique et du nombre baryonique ! (Si les détecteurs sont suffisamment distants, l'« onde d'information » émise par le premier détecteur n'aurait pas assez de temps pour prier le second détecteur de ne pas observer cette même particule α !) Toutefois, ces lois de conservation seraient malgré tout vérifiées « en moyenne », et j'ignore s'il existe une observation qui contredit cette idée. Pour une évaluation récente du statut de l'effondrement retardé, voir Home (1994).

5. Abner Shimony m'a appris que Kochen et Specker connaissaient déjà une formulation EPR de leur propre exemple.
6. Pour des exemples de configurations géométriques différentes, voir Peres (1990, 1991), Mermin (1990), Penrose (1994*a*).
7. Le « miroir semi-argenté » le plus efficace ne serait en fait pas du tout argenté ; ce serait un mince éclat de matériau transparent ayant juste la bonne épaisseur par rapport à la longueur d'onde de la lumière. Il remplirait sa fonction en faisant subir à la lumière une succession complexe de réflexions internes et de transmissions, de sorte que les rayons finalement transmis et réfléchis auraient même intensité. En vertu de l'« unitarité » de la transformation résultante, il s'ensuivrait, entre les rayons finalement transmis et réfléchis, un décalage de phase effectif d'un quart de longueur d'onde, ce qui donnerait le facteur « i » requis. Voir Klein et Furtak (1986) pour une discussion plus complète.
8. Par exemple, Dirac (1947), Davies (1984).
9. Il existe un certain arbitraire sur le choix du facteur de phase que j'ai adopté ici pour l'état réfléchi. Il dépend partiellement du type de miroir utilisé. En fait, contrairement au miroir « semi-argenté » mentionné à la note 7 (qui n'était probablement pas du tout argenté), on peut considérer que ces deux miroirs sont en fait totalement argentés. Le facteur « i » que j'ai adopté ici est une sorte de compromis permettant de réaliser un accord superficiel avec le facteur obtenu dans le cas de la réflexion par des miroirs « semi-argentés ». Peu importe en vérité le facteur adopté pour la réflexion par les miroirs pleinement argentés pourvu que l'on reste cohérent au niveau des choix concernant les deux miroirs en question.
10. Par exemple, Kochen et Specker (1967) et les références données à la note 6.

Chapitre 6

1. On peut, en un certain sens, considérer la propriété « bosonique » des photons mentionnée à la section 5.16 comme un exemple d'emmêlement quantique pour lequel les observations de Hanbury Brown et

Twiss (1954, 1956) fournissent effectivement une confirmation sur de grandes distances (voir la note en bas de page **278**).

2. Everett (1957), Wheeler (1957), DeWitt et Graham (1973), Geroch (1984).
3. Squires (1990, 1992*b*).
4. Bell (1992).
5. Pour différents arguments en faveur de la réalité objective de la fonction d'onde, voir Aharonov, Anandan et Vaidman (1993).
6. Voir par exemple d'Espagnat (1989).
7. Voir d'Espagnat (1989), Zurek (1991, 1993), Paz, Habib et Zurek (1993).
8. Cela semble être la conclusion du programme SETI de F. Drake.
9. Mes propres suggestions, bien que se rangeant fermement dans le camp « gravitationnel », ont récemment trouvé la formulation précise qui leur faisait défaut ; voir Penrose (1993*a*, 1994*b*). Cette proposition partage, avec celle initiale de Ghirardi-Rimini-Weber, l'idée que la réduction serait un processus soudain et discontinu. Toutefois, la plupart des recherches actuelles se concentrent sur un processus de réduction *continu* (stochastique), tel le processus original de Pearle (1976). Voir Diósi (1992), Ghirardi *et al.* (1990*b*), Percival (1994). Pour des recherches visant à rendre ce processus compatible avec la relativité, voir Ghirardi *et al.* (1992), Gisin (1989), Gisin et Percival (1993).
10. Schrödinger (1935*a*), voir aussi EOLP, p. 316-319.
11. Voir aussi Diósi (1989), Ghirardi *et al.* (1990*a*), Penrose (1993*a*).
12. Zeilinger *et al.* (1988).
13. Weber (1960), Braginski (1977).
14. Toutefois, les motivations générales données dans EOLP, chapitre 7, sembleraient conforter davantage la proposition défendue ici (et avancée dans Penrose 1993*a*) que le critère « un graviton » donné dans EOLP. D'autres recherches sont nécessaires pour préciser les liens entre ces deux propositions.
15. Voir Penrose (1991*a*) ; voir aussi EOLP, p. 239.

Chapitre 7

1. Voir par exemple Lisboa (1992).
2. French (1940), Gelber (1958), Applewhite (1979), Fukui et Asai (1976).
3. Dryl (1974).
4. Hameroff et Watt (1982), Hameroff (1987), Hameroff *et al.* (1988).
5. Voir Koruga *et al.* (1993) pour un exposé accessible sur les clathrines, et Curl et Smalley (1991) pour une présentation grand public des fullerènes.

6. Voir Stretton *et al.* (1987).
7. Par exemple, le temps de basculement trouvé par Hameroff pour les tubulines semble concorder avec la fréquence de 5×10^{10}Hz trouvée par Fröhlich pour les cellules.
8. Voir par exemple Isham (1989, 1994), Smolin (1993, 1994).
9. Bien que figurant dans un premier jet de l'article de Deutsch (1991), cette idée ne parut pas dans la version définitive. Deutsch m'a assuré que la raison pour laquelle il l'avait alors supprimée est qu'elle n'avait pas de rapport avec le thème particulier de l'article. Quoi qu'il en soit, eu égard aux objectifs que je cherche à atteindre, la valeur de cette idée ne réside pas tant dans le fait qu'elle serait l'idée « correcte » dans le cadre actuel de la gravitation quantique — puisqu'un tel cadre n'existe pas pour l'instant —, que dans son pouvoir de suggérer de futurs développements — ce qui est effectivement le cas !
   Pour une autre approche de la non-calculabilité en « informatique quantique », voir Castagnoli *et al.* (1992).
10. Quoi qu'il en soit, nos représentations physiques normales du temps ne distinguent pas entre un « écoulement » vers le futur et un « écoulement » vers le passé. (Toutefois, en vertu du deuxième principe de la thermodynamique, le « retour vers le passé » est concrètement irréalisable à l'aide de l'évolution temporelle des équations dynamiques.)
11. Voir aussi Dennett (1991).
    Certains spectateurs du film *Une brève histoire du temps* sur Stephen Hawking et son œuvre ont peut-être trouvé très étranges mes propres opinions sur la relation entre la conscience et l'écoulement du temps. Je saisis cette occasion pour dire que c'est dû à une à coupure malheureuse et hautement trompeuse dans la séquence filmée.
12. Pour d'autres informations sur les torseurs, voir aussi Penrose et Rindler (1986), Ward et Wells (1990), Bailey et Baston (1990).

Chapitre 8

1. Voir par exemple Lisboa (1992).
2. C'est Joël de Rosnay qui m'a fait découvrir cette idée.
3. *Echo of the elephants* (BBC, janvier 1993).
4. *If the rains don't come* (BBC, septembre 1992).
5. *Daylight robbery* (BBC, août 1993).
6. On peut s'interroger sur l'absence de centrioles dans les neurones (*cf.* p. 353). Les cytosquelettes d'autres types de cellules semblent exiger la présence de centrioles jouant le rôle de « centres de contrôle » ; peut-être les cytosquelettes des neurones dépendent-ils d'une autorité plus globale !

7. Marr (1982) et, par exemple, Brady (1993).
8. Donaldson (1983) ; voir Delvin (1988), chapitre 10, pour une présentation non technique.
9. Le « Monde 3 » de Popper contient des constructions mentales présentant une certaine similitude avec celles qui résideraient dans ce monde platonicien élargi ; voir Popper et Eccles (1977). Toutefois, son Monde 3 n'est pas doté d'une existence intemporelle indépendante de nous ; ce n'est pas non plus un monde qui sous-tend la structure même de la réalité physique. Ainsi, son statut est très différent du « monde platonicien » envisagé ici.
10. Dans l'introduction à son livre, Mostowski (1957) montre clairement que des arguments de type gödelien n'ont aucun lien avec le problème de l'existence ou de la non-existence d'énoncés mathématiques *absolument* indécidables. Il faut donc considérer qu'à partir de maintenant, le problème du démontrable et du réfutable reste entièrement ouvert. Cette question est, comme les deux autres, une pure affaire de convictions personnelles !

Remerciements

En rédigeant ce livre, j'ai bénéficié du concours de nombreuses personnes, trop nombreuses pour être toutes individuellement remerciées, même si je me rappelais tous leurs noms. Toutefois, je suis particulièrement redevable à Guido Bacciagaluppi et Jeremy Butterfield pour leur lecture critique de certaines parties du premier jet de cet ouvrage : ils ont découvert une grave erreur de raisonnement dans ce qui depuis est devenu le chapitre 3. Je remercie également Abhay Ashtekar, Mary Bell, Bryan Birch, Geoff Brooker, David Chalmers, Francis Crick, David Deutsch, Solomon Feferman, Robin Gandy, Susan Greenfield, Andrew Hodges, Dipankar Home, Ezio Insinna, Dan Isaacson, Roger James, Richard Jozsa, John Lucas, Bill McColl, Claus Moser, Graeme Michison, Ted Newman, Oliver Penrose, Johnathan Penrose, Stanley Rosen, Ray Sachs, Graeme Segal, Aaron Sloman, Lee Smolin, Ray Streater, Valerie Willoughby, Anton Zeilinger, et surtout Artur Ekert pour leur concours et les diverses informations qu'ils m'ont fournies. Que ce soit de vive voix ou par écrit, nombre de personnes m'ont fait part de leurs commentaires sur mon précédent livre, *l'Esprit, l'ordinateur et les lois de la physique.* Je les remercie ici — même si la plupart d'entre elles attendent encore une réponse à leur lettre ! Si je n'avais bénéficié de leurs divers points de vue sur cet ouvrage, je n'aurais probablement pas eu le courage d'en écrire un second.

Je remercie les organisateurs des Messenger Lectures à Cornell University (ma conférence en ce lieu a donné son titre à la dernière section de ce livre), des Gifford Lectures à l'université de St Andrews, des Forder Lectures en Nouvelle-Zélande, des Gregynogg Lectures à l'université d'Aberystwyth, et de la série de conférences que j'ai données à Five Colleges, à Amherst, Massachusetts, sans compter ceux des innombrables conférences « impromptues » données dans diverses parties du monde. Ces conférences m'ont fourni l'occasion de faire connaître mes idées et de recueillir de précieuses réactions lors des discussions qui s'ensuivaient. Je remercie l'Isaac Newton Institute, à Cambridge, ainsi que Syracuse University et Penn State University pour m'avoir octroyé,

respectivement, une chaire de Professeur extraordinaire de mathématiques et de physique, et la chaire Francis R. Pentz et Helen M. Pentz de Professeur extraordinaire de physique et de mathématiques. Je remercie également pour son aide la National Science Foundation (contrats PHY 86-12424 et PHY 43-96246).

Trois personnes enfin méritent une mention spéciale. Angus MacIntyre s'est dépensé sans compter pour vérifier les raisonnements de logique mathématique des chapitres 2 et 3 et me fournir de nombreuses références bibliographiques. Je lui adresse mes remerciements les plus chaleureux. Stuart Hameroff m'a enseigné la biologie du cytosquelette et de ses microtubules — structures dont, deux années auparavant, j'ignorais jusqu'à l'existence ! Je le remercie vivement pour ses précieuses informations et pour l'aide qu'il m'a apportée en vérifiant la majeure partie du chapitre 7. Je lui suis éternellement redevable de m'avoir ouvert les yeux sur les merveilles d'un monde nouveau. Comme tous ceux que je remercie, il n'est bien sûr en rien responsable des erreurs qui peuvent subsister dans ce livre. Mais par-dessus tout, c'est à ma Vanessa bien-aimée que je dois des remerciements particuliers, et pour plusieurs raisons : pour m'avoir expliqué pourquoi certaines parties du livre exigeaient d'être réécrites ; pour son aide salvatrice lors de l'établissement des références ; et pour son amour, sa patience et sa profonde compréhension, notamment lorsque je sous-estimais constamment le temps que me prendrait la rédaction ! Et je la remercie aussi pour m'avoir — sans le savoir — fourni, dans mon *imagination*, le modèle de la petite Jessica qui apparaît dans ma petite histoire. Je regrette de ne pas l'avoir connue lorsqu'elle avait cet âge !

Crédits des illustrations

Les éditeurs remercient les personnes ou institutions suivantes pour les avoir autorisés, parfois implicitement, à reproduire les illustrations figurant dans ce livre.

Fig. 1.1 A. Nieman/Science Photo Library.

Fig. 4.12 J. C. Mather *et al.*, *Astrophys. J.*, **354**, L37, (1990).

Fig. 5.7 A. Aspect et P. Grangier, in *Quantum concepts in space and time* (R. Penrose et C. J. Isham éd.) Oxford University Press, 1986, p. 1-27.

Fig. 5.8 Ashmolean Museum, Oxford.

Fig. 7.2 R. Wichterman, *The biology of paramecium*, New York, Plenum Press, 2e édition, 1986.

Fig. 7.6 Eric Grave/Science Photo Library.

Fig. 7.7 H. Weyl (1943), *Symmetry*, © 1952 Princeton University Press.

Fig. 7.10 N. Hirokawa, in *The neuronal cytoskeleton*, (R. D. Burgoyne éd.), New York, Wiley-Liss, 1991, p. 5-74.

Bibliographie

Aharonov, Y. et Albert, D. Z. (1981). Can we make sense out of the measurement process in relativistic quantum mechanics ? *Phys. Rev.*, **D24**, 359-370.

Aharonov, Y. et Vaidman, L. (1990). Properties of a quantum system during the time interval between two measurements. *Phys. Rev.*, **A41**, 11.

Aharonov, Y., Anandan, J. et Vaidman, L. (1993), Meaning of the wave function. *Phys. Rev.*, **A47**, 4616-4626.

Aharonov, Y., Bergmann, P. G. et Leibowitz, J. L. (1964). Time symmetry in the quantum process of measurement. In *Quantum theory and measurement* (J. A. Wheeler et W. H. Zurek éd.). Princeton University Press, 1983 ; originellement in *Phys. Rev.*, **B134**, 1410-1416.

Aharonov, Y., Albert, D. Z. et Vaidman, L. (1986). Measurement process in relativistic quantum mechanics. *Phys. Rev.*, **D34**, 1805-1813.

Albert, D. Z. (1983). On quantum-mechanical automata. *Phys. Lett.*, **98A(5, 6)**, 249-252.

Albrecht-Buehler, G. (1985). Is the cytoplasm intelligent too ? *Cell and Muscle Motility*, **6**, 1-21.

Anthony, M. et Biggs, N. (1992). *Computational learning theory, an introduction.* Cambridge University Press.

Applewhite, P. B. (1979). Learning in protozoa. In *Biochemistry and physiology of protozoa*, vol. 1 (M. Levandowsky et S. H. Hunter éd.), Academic Press, New York, p. 341-355.

Arhem, P. et Lindahl, B. I. B. éd. (1983). Neuroscience and the problem of consciousness : theoretical and empirical approaches. In *Theoretical medicine*, **14**, n° 2, Kluwer Academic Publishers.

Aspect, A. et Grangier P. (1986). Experiments on Einstein-Podolsky-Rosen-type correlations with pairs of visible photons. In *Quantum concepts in space and time* (R. Penrose et C. J. Isham éd.). Oxford University Press.

Aspect, A., Grangier P. et Roger, G. (1982). Experimental realization of Einstein-Podolsky-Rosen-Bohm *Gedankenexperiment* : a new violation of Bell's inequalities. *Phys. Rev. Lett.*, **48**, 91-94.

Baars, B. J. (1988). *A cognitive theory of consciousness.* Cambridge University Press.

Bailey, T. N. et Baston, R. J. éd. (1990). *Twistors in mathematics and physics.* London Mathematical Society Lecture Notes Series, **156**, Cambridge University Press.

Baylor, D. A., Lamb , T. D. et Yau, K.-W. (1979). Responses of retinal rods to single photons. *J. Physiol.*, **288**, 613-634.

Beck, F. et Eccles, J. C. (1992). Quantum aspects of consciousness and the role of consciousness. *Proc. Nat. Acad. Sci.*, **89**, 11357-11361.

Becks, K.-H. et Hemker, A. (1992). An artificial intelligence approach to data analysis. In *Proceedings of 1991 CERN School of Computing* (C. Verkerk éd.). CERN, Suisse.

Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. *Physics*, **1**, 195-200.

Bell, J. S. (1966). On the problem of hidden variables in quantum theory. *Rev. Mod. Phys.*, **38**, 447-452.

Bell, J. S. (1987). *Speakable and unspeakable in quantum mechanics.* Cambridge University Press.

Bell, J. S. (1990). Against measurement. *Physics World*, **3**, 33-40.

Benacerraf, P. (1967). God, the Devil and Gödel. *The Monist*, **51**, 9-32.

Benioff, P. (1982). Quantum mechanical Hamiltonian models of Turing Machines. *J. Stat. Phys.*, **29**, 515-546.

Bennett, C. H., Brassard, G., Breidbart, S. et Wiesner, S. (1983). Quantum cryptography, or unforgetable subway tokens. In *Advances in cryptography.* Plenum, New York.

Bernard, C. (1875). *Leçons sur les anesthésiques et sur l'asphyxie.* J. B. Baillière, Paris.

Blakemore, C. et Greenfield, S. éd. (1987). *Mindwaves : thoughts on intelligence, identity, and consciousness.* Blackwell, Oxford.

Blum, L., Shub, M. et Smale, S. (1989). On a theory of computation and complexity over the real numbers : NP completeness, recursive functions and universal machines. *Bull. Amer. Math. Soc.*, **21**, 1-46.

Bock, G. R. et Marsh, J. (1993). *Experimental and theoretical studies of consciousness.* Wiley.

Boden, M. (1977). *Artificial intelligence and natural man.* The Harvester Press, Hassocks.

Boden, M. A. (1990). *The creative mind : myths and mecanisms.* Wiedenfeld and Nicolson, Londres.

Bohm, D. (1952). A suggested interpretation of the quantum theory in terms of « hidden » variables, I and II. In *Quantum theory and measurement* (J. A. Wheeler et W. H. Zurek éd.). Princeton University Press, 1983 ; originellement in *Phys. Rev.*, **85**, 166-193.

Bohm, D. et Hiley, B. (1994). *The undivided universe.* Routledge, Londres.

Boole, G. (1854). *An investigation of the laws of thought.* Dover, New York, 1958. Trad. fr.(Souleymane Bachir Diagne), *Les Lois de la pensée*, Paris, Vrin, 1992.

Boolos, G. (1990). On seeing the truth if the Gödel sentence. *Behavioural and Brain Sciences*, **13(4)**, 655.

Bowie, G. L. (1982). Lucas' number is finally up. *J. of Philosophical Logic*, **11**, 279-285.

Brady, M. (1993). Computational vision. In *The simulation of human intelligence* (D. Broadbent éd.). Blackwell, Oxford.

Braginsky, V. B. (1977). The detection of gravitational waves and quantum non-disturbtive measurements. In *Topics in theoretical and experimental gravitation physics* (V. de Sabbata et J. Weber éd.), Plenum, Londres, p. 105.

Broadbent, D. (1993). Comparison with human experiments. In *The simulation of human intelligence* (D. Broadbent éd.). Blackwell, Oxford.

Brown, H. R. (1993). Bell's other theorem and its connection with nonlocality. Part I. In *Bell's theorem and the foundations of physics* (A. van der Merwe et F. Selleri éd.). World Scientific, Singapour.

Butterfield, J. (1990). Lucas revived ? An undefended flank. *Behavioural and Brain Sciences*, **13(4)**, 658.

Castagnoli, G., Rasetti, M. et Vincenti, A. (1992). Steady, simultaneous quantum computation : a paradigm for the investigation of nondeterministic and non-recursive computation. *Int. J. Mod. Phys. C*, **3**, 661-689.

Caudill, M. (1992). *In our own image. Building an artificial person.* Oxford University Press.

Chaitin, G. J. (1975). Randomness and mathematical proof. *Scientific American*, (mai 1975), 47.

Chalmers, D. J. (1990). Computing the thinkable. *Behavioural and Brain Sciences*, **13(4)**, 658.

Chandrasekhar, S. (1987). *Truth and beauty. Aesthetics and motivations in science.* The University of Chicago Press.

Chang, C.-L. et Lee, R. C.-T. (1987). Symbolic logic and mechanical theorem proving. Academic Press, New York, 2ème éd. (1ère éd. 1973).

Chou, S.-C. (1988). *Mechanical geometry theorem proving.* Ridel.

Christian, J. J. (1994). On definite events in a generally covariant quantum world. Préprint.

Church, A. (1936). An unsolvable problem of elementary number theory. *Am. Jour. of Math.*, **58**, 345-363.

Church, A. (1941). *The calculi of lambda-conversion.* Annals of Mathematics Studies, n° 6, Princeton University Press.

Churchland, P. M. (1984). *Matter and consciousness.* Bradford Books, MIT Press, Cambridge, Massachusetts.

Clauser, J. F. et Horne, M. A. (1974). Experimental consequences of objective local theories. *Phys. Rev.*, **D10**, 526-535.

Clauser, J. F., Horne, M. A. et Shimony, A. (1978). Bell's Theorem : experimental tests and implications. *Rpts. on Prog. in Phys.*, **41**, 1881-1927.

Cohen, P. C. (1966). *Set theory and the continuum hypothesis.* Benjamin, Menlo Park, Californie.

Conrad, M. (1990). Molecular computing. In *Advances in computers* (M. C. Yovits éd.). Vol. 31, Academic Press, Londres.

Conrad, M. (1992). Molecular computing : the lock-key paradigm. *Computer* (novembre 1992), 11-20.

Conrad, M. (1993). The fluctuon model of Force, Life, and computation : a constructive analysis. *Appl. Math. and Comp.*, **56**, 203-259.

Costa de Beauregard, O. (1989). In *Bell's theorem, quantum theory, and conceptions of the universe* (M. Kafatos éd.). Kluwer, Dordrecht.

Craik, K. (1943). *The nature of explanation.* Cambridge University Press.

Crick, F. (1994). *The astonishing hypothesis. The scientific search for the soul.* Charles Scribner's Sons, New York, et Maxwell Macmillan International. Trad. fr. (M. Prouteau), *L'Hypothèse stupéfiante : à la recherche scientifique de l'âme*, Paris, Plon, 1995.

Crick, F. et Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, **2**, 263-275.

Crick, F. et Koch, C. (1992). The problem of consciousness. *Sci. Amer.*, **267**, 110.

Curl, R. F. et Smalley, R. E. (1991). Fullerenes. *Scientific American*, **265**, n° 4, 32-41.

Cutland, N. J. (1980). *Computability. An introduction to recursive function theory.* Cambridge University Press.

Davenport, H. (1952). *The higher arithmetic.* Hutchinson's University Library.

Davies, P. C. W. (1974). *The physics of time asymmetry.* Surrey University Press, Belfast.

Davies, P. C. W. (1984). *Quantum mechanics.* Routledge, Londres.

Davis, M. éd. (1965). *The undecidable — basic papers on undecidable propositions, unsolvable problems and computable functions.* Raven Press, Hewlett, New York.

Davis, M. (1978). What is a computation ? In *Mathematics today ; twelve informal essays* (L. A. Steen éd.). Springer-Verlag, New York.

Davis, M. (1990). Is mathematical insight algorithmic ? *Behavioural and Brain Sciences*, **13(4)**, 659.

Davis, M. (1993). How subtle is Gödel's theorem ? *Behavioural and Brain Sciences*, **16**, 611-612.

Davis, M. et Hersch, R. (1965). Hibert's tenth problem. *Scientific American*, (novembre 1973), 84.

Davis, M. et Hersch, R. (1982). *The mathematical experience.* Harvester Press, Hassocks. Trad. fr. (L. Chambadal), *L'Univers mathématique*, Paris, Dunod, 1985.

de Broglie, L. (1956). *Tentative d'interprétation causale et non linéaire de la méanique ondulatoire.* Gauthier-Villars, Paris.

Deeke, L., Grötzinger, B. et Kornhuber, H. H. (1976). Voluntary finger movements in man : cerebral potentials and theory. *Biol. Cybernetics*, **23**, 99.

Dennett, D. C. (1990). Betting your life on an algorithm. *Behavioural and Brain Sciences*, **13(4)**, 660.

Dennett, D. C. (1991). *Consciousness explained.* Little, Brown and Company. Trad. fr., *La Conscience expliquée*, Paris, Odile Jacob, 1993.

Deutsch, D. (1985). Quantum theory, the Church-Turing principle and the universal quantum computer. *Proc. Roy. Soc. Lond.*, **A400**, 97-117.

Deutsch, D. (1989). Quantum computational networks. *Proc. Roy. Soc. Lond.*, **A425**, 73-90.

Deutsch, D. (1991). Quantum mechanics near closed time-like lines. *Phys. Rev.*, **D44**, 3197-3217.

Deutsch, D. (1992). Quantum computation. *Phys. World*, **5**, 57-61.

Deutsch, D. et Josza, R. (1992). Rapid solution of problems by quantum computation. *Proc. Roy. Soc. Lond.*, **A439**, 553-558.

Deutsch, D. et Ekert, A. (1993). Quantum communication moves into the unknown. *Phys. World*, **6**, 22-23.

Devlin, K. (1988). *Mathematics : the new golden age.* Penguin Books, Londres. Trad. fr. (G. Kreweras), *Mathématiques contemporaines, un nouvel âge d'or*, Paris, Masson, 1992.

DeWitt, B. S. et Graham, R. D. éd. (1973). *The many-worlds interpretation of quantum mechanics.* Princeton University Press.

Dicke, R. H. (1981). Interaction-free quantum measurements : a paradox ? *Am. J. Phys.*, **49**, 925-930.

Diósi, L. (1989). Models for universal reduction of macroscopic quantum fluctuations. *Phys. Rev.*, **A40**, 1165-1174.

Diósi, L. (1992). Quantum measurement and gravity for each other. In *Quantum chaos, quantum measurement.* NATO AS1 Series C. Math. Phys. Sci. 357 (P. Cvitanovic, I. C. Percival et A. Wirzba éd.). Kluwer, Dordrecht.

Dirac, P. A. M. (1947). *The principles of quantum mechanics.* 3ème éd., Oxford University Press. Trad. fr. (A. Proca et J. Ullmo), *Les Principes de la mécanique quantique*, Paris, J. Gabay, 1990.

Dodd, A. (1991). Gödel, Penrose, and the possibility of AI. *Artificial Intelligence Review*, **5**.

Donaldson, S. K. (1983). An application of gauge theory to four dimensional topology. *J. Diff. Geom.*, **18**, 279-315.

Doyle, J. (1990). Perceptive questions about computation and cognition. *Behavioural and Brain Sciences*, **13(4)**, 661.

Dreyfus, H. L. (1972). *What computers can't do.* Harper and Row, New York. Trad. fr. (R.-M. Vassalo), *L'Intelligence artificielle : mythes et limites*, Paris, Flammarion, 1984.

Dryl, S. (1974). Behaviour and motor response in paramecium. In *Paramecium — a current survey* (W. J. van Wagtendonk éd.). Elsevier, Amsterdam, p. 165-218.

Dummet, M. (1973). *Frege : philosophy of language.* Duckworth, Londres.

Dustin, P. (1984). *Microtubules.* Springer-Verlag, Berlin, 2ème édition, révisée et corrigée.

Eccles, J. C. (1973). *The understanding of the brain.* McGraw-Hill, New York.

Eccles, J. C. (1989). *Evolution of the brain : creation of the self.* Routledge, Londres. Trad. fr. (J.-M. Luccioni), *Évolution du cerveau et création de la conscience : à la recherche de la vraie nature de l'homme*, Paris, Flammarion, 1994.

Eccles, J. C. (1992). Evolution of consciousness. *Proc. Nat. Acad. Sci.*, **89**, 7320-7324.

Eccles, J. C. (1994). *How the self controls its brain.* Springer-Verlag, Berlin.

Eckert, R., Randall, D. et Augustine, G. (1988). *Animal physiology. Mechanism and adaptation.* Freeman, New York, chapitre 11.

Eckhorn, R., Bauer, R. Jordan, W., Brosch, M., Kruse, W., Munk, M. et Reitboeck, H. J. (1988). Coherent oscillations : a mechanism of feature linking in the visual cortex ? *Biol. Cybernetics*, **60**, 121-130.

Edelman, G. M. (1976). Surface modulation and cell recognition on cell growth. *Science*, **192**, 218-226.

Edelman, G. M. (1987). *Neural Darwinism, the theory of neuronal group selection.* Basic Books, New York.

Edelman, G. M. (1988). *Topobiology, an introduction to molecular embryology.* Basic Books, New York.

Edelman, G. M. (1989). *The remembered present. A biological theory of consciousness.* Basic Books, New York. Trad. fr. (A. Gerschenfeld), *Biologie de la conscience*, Paris, Le Seuil, 1994.

Edelman, G. M. (1992). *Bright air, brilliant fire : on the matter of the mind.* Allen Lane, The Penguin Press, Londres.

Einstein, A, Podolsky, P. et Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete ? In *Quantum theory and measurement* (J. A. Wheeler et W. H. Zurek éd.). Princeton University Press, 1983 ; originellement in *Phys. Rev.*, **47**, 777-780.

Elitzur, A. C. et Vaidman, L. (1993). Quantum-mechanical interaction-free measurements. *Found. of Phys.*, **23**, 987-997.

Elkies, N. G. (1988). On $A^4 + B^4 + C^4 = D^4$. *Maths. of Computation*, **51**, n° 184, 825-835.

d'Espagnat, B. (1989). *Conceptual foundations of quantum mechanics.* Addison-Wesley, Reading, Massachusetts, 2ème éd. Traduction de *Fondements conceptuels de la mécanique quantique.* Paris, Hermann, 1972.

Everett, H. (1957). « Relative State » formulation of quantum mechanics. In *Quantum theory and measurement* (J. A. Wheeler et W. H. Zurek éd.). Princeton University Press, 1983 ; originellement in *Rev. of Modern Physics*, **29**, 454-462.

Feferman, S. (1988). Turing in the Land of O(z). In *The universal Turing machine : a half-century survey* (R. Herken éd.). Kammerer und Unverzagt, Hambourg.

Feynman, R. P. (1948). Space-time approach to non-relativistic quantum mechanics *Rev. Mod. Phys.*, **20**, 367-387.

Feynman, R. P. (1982). Simulating physics with computers. *Int. J. Theor. Phys.*, **21(6/7)**, 467-488.

Feynman, R. P. (1985). Quantum mechanical computers. *Optics News* (février), 11-20.

Feynman, R. P. (1986). Quantum mechanical computers. *Foundations of Physics*, **16(6)**, 507-531.

Fodor, J. A. (1983). *The modularity of mind.* MIT Press, Cambridge, Massachusetts. Trad. fr. (A. Gerschenfeld), *La Modularité de l'esprit : essai sur la psychologie des facultés*, Paris, Minuit, 1986.

Franks, N. P. et Lieb, W. R. (1982). Molecular mechanics of general anaesthesia. *Nature*, **300**, 487-493.

Freedman, D. H. (1994). *Brainmakers.* Simon and Schuster, New York.

Freedman, S. J. et Clauser, J. F. (1972). Experimental test of local hidden-variable theories. In *Quantum theory and measurement* (J. A. Wheeler et W. H. Zurek éd.). Princeton University Press, 1983 ; originellement in *Phys. Rev. Lett.*, **28**, 938-941.

Frege, G. (1893). *Grundgesetze der Arithmetik, begriffsschriftlich abgeleitet.* Vol. 1, H. Pohle, Iéna. Trad. fr., *Les Fondements de l'arithmétique*, Paris, Le Seuil, 1970.

French, J. W. (1940). Trial and error learning in paramecium. *J. Exp. Psychol.*, **26**, 609-613.

Fröhlich, H. (1968). Long-range coherence and energy storage in biological systems. *Int. Jour. of Quantum Chem.*, **II**, 641-649.

Fröhlich, H. (1970). Long-range coherence and the actions of enzymes. *Nature*, **228**, 1093.

Fröhlich, H. (1975). The extraordinary dielectric properties of biological materials and the actions of enzymes. *Proc. Nat. Acad. Sci.*, **72(11)**, 4211-4215.

Fröhlich, H. (1984). General theory of coherent excitations on biological systems. In *Nonlinear electrodynamics in biological systems* (W. R. Adey et A. F. Lawrence éd.). Plenum Press, New York.

Fröhlich, H. (1986). Coherent excitations in active biological systems. In *Modern Bioelectrochemistry* (F. Gutmann et H. Keyzer éd.). Plenum Press, New York.

Fukui, K et Asai, H. (1976). Spiral motion of paramecium caudatum in small capillary glass tube. *J. Protozool.*, **23**, 559-563.

Gandy, R. (1988). The confluence of ideas in 1936. In *The universal Turing machine : a half-century survey* (R. Herken éd.). Kammerer und Unverzagt, Hambourg.

Gardner, M. (1965). *Mathematical magic show.* Alfred Knopf, New York et Random House, Toronto.

Gardner, M. (1970). Mathematical games : the fantastic combinations of John Conway's new solitaire game « Life ». *Scientific American*, **223**, 120-123.

Gardner, M. (1989). *Penrose tiles to trapdoor cipher.* Freeman, New York.

Gelber, B. (1958). Retention in paramecium aurelia. *J. Comp. Physiol. Psych.*, **51**, 110-115.

Gelernter, D. (1994). *The muse in the machine.* The Free Press, Macmillan Inc., New York et Collier Macmillan, Londres.

Gell-Mann, M. et Hartle, J. B. (1993). Classical equations for quantum systems. *Phys. Rev.*, **D47**, 3345-3382.

Gernoth, K. A., Clark, J. W., Prater, J. S. et Bohr, H. (1993). Neural network models of nuclear systematics. *Phys. Lett.*, **B300**, 1-7.

Geroch, R. (1984). The Everett interpretation. *Nous*, **4** (numéro spécial sur les fondements de la mécanique quantique), 617-633.

Geroch, R. et Hartle, J. B. (1986). Computability and physical theories. *Found. Phys.*, **16**. 533.

Ghirardi, G. C., Rimini, A. et Weber, T. (1980). A general argument against superluminal transmission through the quantum mechanical measurement process. *Lett. Nuovo Cim.*, **27**, 293-298.

Ghirardi, G. C., Rimini, A. et Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Phys. Rev.*, **D34**, 470.

Ghirardi, G. C., Grassi, R. et Rimini, A. (1990*a*). Continuous-spontaneous reduction model involving gravity. *Phys. Rev.*, **A42**, 1057-1064.

Ghirardi, G. C., Grassi, R. et Pearle, P. (1990*b*). Relativistic dynamical reduction models : general framework and examples. *Foundations of Physics*, **20**, 1271-1316.

Ghirardi, G. C., Grassi, R. et Pearle, P. (1992). Comment on « Explicit collapse and superluminal signals ». *Phys. Lett.*, **A166**, 435-438.

Ghirardi, G. C., Grassi, R. et Pearle, P. (1993). Negotiating the tricky border between quantum and classical. *Physics Today*, **46**, 13.

Gisin, N. (1989). Stochastic quantum dynamics and relativity. *Helv. Phys. Acta*, **62**, 363-371.

Gisin, N. and Percival, I. C. (1993). Stochastic wave equations versus parallel world components. *Phys. Lett.*, **A175**, 144-145.

del Giudice, E., Doglia, S. et Milani, M. (1983). Self-focusing and ponderomotive forces of coherent electric waves — a mechanism for cytoskeleton formation and dynamics. In *Coherent excitations in biological systems* (H. Fröhlich et F. Kremer éd.). Springer-Verlag, Berlin.

Gleick, J. (1987). *Chaos. Making a new science.* Penguin Books, Londres. Trad. fr. (Ch. Jeanmougin), *La Théorie du chaos : vers une nouvelle science*, Paris, Flammarion, 1991.

Glymour, C. et Kelly, K. (1990). Why you'll never know wether Roger Penrose is a computer. *Behavioural and Brain Sciences*, **13(4)**, 666.

Gödel, K. (1931). Über formale unentscheidbare Sätze per Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, **38**, 173-198.

Gödel, K. (1940). *The consistency of the axiom of choice and of the generalized continuum-hypothesis with the axioms of set theory.* Princeton University Press et Oxford University Press.

Gödel, K. (1949). An example of a new type of cosmological solution of Einstein's field equations of gravitation. *Rev. of Mod. Phys.*, **21**, 447.
Gödel, K. (1986). *Kurt Gödel, collected works*, vol. I (1929-1936) (S. Feferman *et al.* éd.). Oxford University Press.
Gödel, K. (1990). *Kurt Gödel, collected works*, vol. II (1938-1974) (S. Feferman *et al.* éd.). Oxford University Press.
Gödel, K. (1995). *Kurt Gödel, collected works*, vol. III (S. Feferman *et al.* éd.). Oxford University Press.
Golomb, S. W. (1965). *Polyominoes.* Scribner and Sons, New York.
Good, I. J. (1965). Speculations concerning the first ultra intelligent machine. *Advance Computers*, **6**, 31-88.
Good, I. J. (1967). Human and machine logic. *Brit. J. Philos. Sci.*, **18**, 144-147.
Good, I. J. (1969). Gödel's theorem is a red herring. *Brit. J. Philos. Sci.*, **18**, 359-373.
Graham, R. L. et Rothschild, B. L. (1971). Ramsey's theorem for *n*-parameter sets. *Trans. Am. Math. Soc.*, **59**, 290.
Grangier, P., Roger, G. et Aspect, A. (1986). Experimental evidence for a photon anticorrelation effet on a beam splitter : a new light on single-photon interferences. *Europhysics Letters*, **1**, 173-179.
Grant, P. M. (1994). Another December revolution ? *Nature*, **367**, 16.
Gray, C. M. et Singer, W. (1989). Stimulus-specific neuronal oscillations in orientations columns of cat visual cortex. *Proc. Nat. Acad. Sci. USA*, **86**, 1689-1702.
Green, D. G. et Bossomaier, T. éd. (1993). *Complex systems : from biology to computation.* IOS Press.
Greenberger, D. M., Horne, M. A. et Zeilinger, A. (1989). Going beyond Bell's theorem. In *Bell's theorem, quantum theory, and conceptions of the universe* (M. Kafatos éd.). Kluwer, Dordrecht, p. 73-76.
Greenberger, D. M., Horne, M. A., Shimony, A et Zeilinger, A. (1990). Bell's theorem without inequalities. *Am. J. Phys.*, **58**, 1131-1143.
Gregory, R. L. (1981). *Mind in science : a history of explanations in psychology and physics.* Weidenfeld et Nicholson Ltd (réédition, Penguin, 1984).
Grey Walter, W. (1953). *The living brain.* Gerald Duckworth and Co. Ltd, Londres.
Griffiths, R. (1984). Consistent histories and the interpretation of quantum mechanics. *J. Stat. Phys.*, **36**. 219.
Grossberg, S. éd. (1987). *The adaptive brain I : Cognition, learning, reinforcement and rythm* et *The adaptive brain II : Vision, speech, language and motor control.* North-Holland, Amsterdam.
Grünbaum, B. et Shephard, G. C. (1987). *Tilings and patterns.* Freeman, New York.
Grundler, W et Keilmann, F. (1983). Sharp resonances in yeast growth proved nonthermal sensitivity to microwaves. *Phys. Rev. Lett.*, **51**, 1214-1216.
Guccione, S. (1993). Mind the truth : Penrose's new step in the Gödelian argument. *Behavioural and Brain Sciences*, **16**, 612-613.

Haag, R. (1992). *Local quantum physics : fields, particles, algebras.* Springer-Verlag, Berlin.

Hadamard, J. (1945). *The psychology of invention in the mathematical field.* Princeton University Press. Trad. fr. (J. Hadamard), *Essai sur la psychologie de l'invention dans le domaine des mathématiques*, Paris, Gauthier-Villars, 1975.

Hallett, M. (1984). *Cantorian set theory and limitation of size.* Clarendon Press, Oxford.

Hameroff, S. R. (1974). Chi : a neural hologram ? *Am. J. Clin. Med.*, **2(2)**, 163-170.

Hameroff, S. R. (1987). *Ultimate computing. Biomolecular consciousness and nano-technology.* North-Holland, Amsterdam.

Hameroff, S. R. et Watt, R. C. (1982). Information processing in microtubules. *J. Theor. Biol.*, **98**, 549-561.

Hameroff, S. R. et Watt, R. C. (1983). Do anesthetics act by altering electron mobility ? *Anesth. Analg.*, **62**, 936-940.

Hameroff, S. R., Rasmussen, S. et Mansson, B. (1988). Molecular automata in microtubules : basic computational logic of the living state ? In *Artificial Life, SFI studies in the siences of complexity* (C. Langton éd.). Addison-Wesley, New York.

Hanbury Brown, R. et Twiss, R. Q. (1954). A new type of interferometer for use in radioastronomy. *Phil. Mag.*, **45**, 663-682.

Hanbury Brown, R. et Twiss, R. Q. (1956). The question of correlation between photons in coherent beams of light. *Nature*, **177**, 27-29.

Harel, D. (1987). *Algorithmics. The spirit of computing.* Addison-Wesley, New York.

Hawking S. W. (1975). Particle creation by Black Holes. *Commun. Math. Phys.*, **43**, 199-220.

Hawking S. W. (1982). Unpredictability of quantum gravity. *Commun. Math. Phys.*, **87**, 395-415.

Hawking S. W. et Israel, W. éd. (1987). *300 years of gravitation.* Cambridge University Press.

Hebb, D. O. (1949). *The organization of behaviour.* Wiley, New York.

Hecht, S., Shlaer, S. et Pirenne, M. H. (1941). Energy, quanta and vision. *Journal of General Physiology*, **25**, 891-940.

Herbert, N. (1993). *Elemental mind. Human consciousness and the new physics.* Dutton Books, Penguin Publishing.

Heyting, A. (1956). *Intuitionism : an introduction.* North-Holland, Amsterdam.

Heywood, P. et Redhead, M. L. G. (1983). Nonlocality and the Kochen-Specker Paradox. *Found. Phys.*, **13**, 481-499.

Hodges, A. P. (1983). *Alan Turing : the enigma.* Burnett Books and Hutchinson, Londres ; Simon and Schuster, New York. Trad. fr. (N. Zimmermann), *Alan Turing ou l'énigme de l'intelligence.* Paris, Payot, 1988.

Hodgking, D. et Houston, A. I. (1990). Selecting for the con in consciousness. *Behavioural and Brain Sciences*, **13(4)**, 668.

Hodgson, D. (1991). *Mind matters : consciousness and choice in a quantum world.* Clarendon Press, Oxford.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach : an eternal golden braid.* Harvester Press, Hassocks. Trad. fr. (J. Henry et R. French), *Gödel, Escher, Bach : les brins d'une guirlande éternelle.* Paris, InterÉditions, 1985.

Hofstadter, D. R. (1981). A conversation with Einstein's brain. In *The Mind's I* (D. R. Hofstadter et D. C. Dennett éd.). Basic Books, Penguin, Harmondsworth, Middlesex. Trad. fr. (J. Henry), *Vues de l'esprit*, Paris, InterÉditions, 1987.

Hofstadter D. R. et Dennett D. C éd. (1981). *The Mind's I.* Basic Books, Penguin, Harmondsworth, Middlesex. Trad. fr. (J. Henry), *Vues de l'esprit*, Paris, InterÉditions, 1987.

Home, D. (1994). A proposed new test of collapse-induced quatum nonlocality. Préprint.

Home, D. et Nair, R. (1994). Wave function collapse as a nonlocal quantum effet. *Phys. Lett.*, **A187**, 224-226.

Home, D. et Selleri, F. (1991). Bell's theorem and the EPR paradox. *Rivista del Nuovo Cimento*, **14**, N.9.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci.*, **79**, 2554-2558.

Hsu, F.-H, Anantharaman, T., Campbell, M. et Nowatzyk, A. (1990). A grandmaster chess machine. *Scientific American*, **263**.

Huggett, S. A. et Tod, K. P. (1985). *An introduction to twistor theory.* London Math. Soc. student texts. Cambridge University Press.

Hughston, L. P., Jozsa, R. et Wootters, W. K. (1993). A complete classification of quantum ensembles having a given density matrix. *Phys. Lett.*, **A183**, 14-18.

Isham, C. J. (1989). Quantum gravity. In *The new physics* (P. C. W. Davies éd.). Cambridge University Press, p. 70-93.

Isham, C. J. (1994). Prima facie questions in quantum gravity. In *Canonical relativity : classical and quantum* (J. Ehlers et H. Friedrich éd.). Springer-Verlag, Berlin.

Jibu, M., Hagan, S. Pribram, K. Hameroff, S. R. et Yasue, K. (1994). Quantum optical coherence in cytoskeletal microtubules : implications for brain function. *Bio. Systems* (sous presse).

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge University Press.

Johnson-Laird, P. N. (1987). How could consciousness arise from the computations of the brain ? In *Mindwaves : thoughts on intelligence, identity, and consciousness* (C. Blakemore et S. Greenfield, éd.). Blackwell, Oxford.

Károlyházy, F. (1966). Gravitation and quantum mechanics of macroscopic bodies. *Nuo. Cim. A*, **42**, 390-402.

Károlyházy, F. (1974). Gravitation and quantum mechanics of macroscopic bodies. *Magyar Fizikai Polyoirat*, **12**, 24.

Károlyházy, F., Frenkel, A. et Lukács, B. (1986). On the possible role of gravity on the reduction of the wave function. In *Quantum concepts in space and time* (R. Penrose et C. J. Isham éd.). Oxford University Press.

Kasumov, A. Y., Kislov, N. A. et Khodos, I. I. (1993). Can the observed vibration of a cantilever of supersmall mass be explained by quantum theory ? *Microsc. Microanal. Microstruct.*, **4**, 401-406.

Kentridge, R. W. (1990). Parallelism and patterns of thought. *Behavioural and Brain Sciences*, **13(4)**, 670.

Khalfa, J. éd. (1994). *What is intelligence ? The Darwin College lectures.* Cambridge University Press.

Klarner, D. A. (1981). My life among the Polyominoes. In *The mathematical gardner* (D. A. Klarner éd.). Prindle, Weber and Schmidt, Boston (Massachusetts), et Wadsworth Int., Bellmont (Californie).

Kleene, S. C. (1952). *Introduction to metamathematics.* North-Holland, Amsterdam et Van Nostrand, New York.

Klein, M. V. et Furtak, T. E. (1986). *Optics.* Wiley, New York, 2ème éd.

Kochen, S. et Specker, E. P. (1967). The problem of hidden variables in quantum mechanics. *J. Math. Mech.*, **17**, 59-88.

Kohonen, T. (1984). *Self-organisation and associative memory.* Springer-Verlag, Berlin.

Komar, A. B. (1969). Qualitative features of quantized gravitation. *Int. J. Theor. Phys.*, **2**, 157-160.

Koruga, D. (1974). Microtubule screw symmetry : packing of spheres as a latent bioinformation code. *Ann. NY Acd. Sci.*, **466**, 953-955.

Koruga, D., Hameroff, S., Withers, J. Loutfy, R. et Sundareshan, M. (1993). *Fullerene C_{60}. History, physics, nanobiology, nanotechnology.* North-Holland, Amsterdam.

Kosko, B. (1994). *Fuzzy thinking : the new science of fuzzy logic.* Harper Collins, Londres.

Kreisel, G. (1960). Ordinal logic and the characterization of informal concepts of proof. *Proc. of the Internat. Cong. of Mathematics, Aug. 1958.* Cambridge University Press.

Kreisel, G. (1967). Informal rigour and completeness proofs. In *Problems in the philosophy of mahematics* (I. Lakatos éd.). North-Holland, Amsterdam, p. 138-186.

Laguës, M., Xiao Ming Xie, Tebbji, H., Xiang Zhen Xu, Mairet, V., Hatterer, C. *et al.* (1993). Evidence suggesting superconductivity at 250 K in a sequentially deposited cuprate film. *Science*, **262**, 1850-1851.

Lander, L. J. et Parkin, T. R. (1966). Counterexample to Euler's conjecture on sums of like powers. *Bull. Amer. Math. Soc.*, **72**, 1079.

Leggett, A. J. (1984). Schrödinger's cat and her laboratory cousins. *Contemp. Phys.*, **25(6)**, 583.

Lewis, D. (1969). Lucas against mechanism. *Philosophy*, **44**, 231-233.

Lewis, D. (1989). Lucas against mechanism II. *Can. J. Philos.*, **9**, 373-376.

Libet, B. (1990). Cerebral processes that distinguish conscious experience from inconscious mental functions. In *The principles of design and*

operation of the brain (J. C. Eccles et O. D. Creutzfeldt éd.), Experimental Brain research series 21, Springer-Verlag, Berlin, p. 185-205.

Libet, B. (1992). The neural time-factor in perception, volition and free will. *Revue de métaphysique et de morale*, **2**, 255-272.

Libet, B., Wright, E. W. Jr, Feinstein, B. et Pearl, D. K. (1979). Subjective referral of the timing for a conscious sensory experience. *Brain*, **102**, 193-224.

Linden, E. (1993). Can animals think ? *Time Magazine* (mars), 13.

Lisboa, P. G. J. éd. (1992). *Neural networks : current applications.* Chapman Hall, Londres.

Lockwood, M. (1989). *Mind, brain and the quantum.* Blackwell, Oxford.

Longair, M. S. (1993). Modern cosmology — a critical assessment. *Q. J. R. Astr. Soc.*, **34**, 157-199.

Longuet-Higgins, H. C. (1987). Mental processes : studies in cognitive science, Part II. MIT Press, Cambridge (Massachusetts).

Lucas, J. R. (1961). Minds, machines and Gödel. *Philosophy*, **36**, 120-124. Reproduit in Alan Ross Anderson éd. (1964), *Minds and Machines.* Englewood Cliffs.

Lucas, J. R. (1970). *The freedom of the will.* Oxford University Press.

McCarthy, J. (1979). Ascribing mental qualities to machines. In *Philosophical perspectives in artificial intelligence* (M. Ringle éd.). Humanities Press, New York.

McCullogh, W. S. et Pitts, W. H. (1943). A logical calculus of the idea immanent in nervous activity. *Bull. Math. Biophys.*, **5**, 115-133. Reproduit in McCullogh, W. S., *Embodiments of mind*, MIT Press, 1965.

McDermott, D. (1990). Computation and consciousness. *Behavioural and Brain Sciences*, **13(4)**, 676.

MacLennan, B. (1990). The discomforts of dualism. *Behavioural and Brain Sciences*, **13(4)**, 673.

Majorana, E. (1932). Atomi orientati in campo magnetico variabile. *Nuovo Cimento*, **9**, 43-50.

Manaster-Ramer, A., Savitch, W. J. et Zadrozny, W. (1990). Gödel redux. *Behavioural and Brain Sciences*, **13(4)**, 675.

Margulis, L. (1975). *Origins of eukaryotic cells.* Yale University Press, New Haven (Connecticut).

Markov, A. A. (1958). The insolubility of the problem of homeomorphy. *Dokl. Akad. Nauk. URSS*, **121**, 218-220.

Marr, D. E. (1982). *Vision : a computational investigation into the human representation and processing of visual information.* Freeman, San Francisco.

Marshall, I. N. (1989). Consciousness and Bose-Einstein condensates. *New Ideas in Psychology*, **7**.

Mermin, D. (1985). Is the moon there when nobody looks ? Reality and the quantum theory. *Physics Today*, **38**, 38-47.

Mermin, D. (1990). Simple unified form of the major no-hidden-variables theorems. *Phys. Rev. Lett.*, **65**, 3373-3376.

Michie, D. et Johnston, R. (1984). *The creative computer. Machine intelligence and human knowledge.* Viking Penguin.

Minsky, M. (1968). Matter, mind and models. In *Semantic information processing* (M. Minsky éd.). MIT Press, Cambridge (Massachusetts).

Minsky, M. (1986). *The society of mind.* Simon and Schuster, New York. Trad. fr. (J. Henry), *La Société de l'esprit*, Paris, InterÉditions, 1988.

Minsky, M. et Papert, S. (1972). *Perceptrons : an introduction to computational geometry.* MIT Press, Cambridge (Massachusetts).

Misner, C. W., Thorne, K. et Wheeler, J. A. (1973). *Gravitation.* Freeman, New York.

Moore, A. W. (1990). *The infinite.* Routledge, Londres.

Moravec, H. (1988). *Mind children : the future of robot and human intelligence.* Harvard University Press, Cambridge (Massachusetts). Trad. fr., *Une vie après la vie : les robots, avenir de l'intelligence*, Paris, Odile Jacob, 1992.

Moravec, H. (1994). *The age of mind : transcending the human condition through robots.* Sous presse.

Mortensen, C. (1990). The powers of machines and minds. *Behavioural and Brain Sciences*, **13(4)**, 678.

Mostowski, A. (1957). *Sentences undecidable in formalized arithmetic : an exposition of the theory of Kurt Gödel.* North-Holland, Amsterdam.

Nagel, E. et Newman, J. R. (1958). *Gödel's proof.* Routledge and Keagan Paul.

von Neumann, J. (1932). *Mathematische Grundlagen des Quantenmechanik.* Springer-Verlag, Berlin. Trad. fr. *Les Fondements de la mécanique quantique*, Paris, Jacques Gabay, 1988.

von Neumann, J. et Morgenstern, O. (1944). *Theory of games and economic behaviour.* Princeton University Press.

Newell, A. et Simon, H. A. (1976). Computer science as empirical enquiry : symbols and search. *Communications of the ACM*, **19**, 113-126.

Newell, A., Young, R. et Polk, T. (1993). The approach through symbols. In *The simulation of human intelligence* (D. Broadbent éd.). Blackwell, Oxford.

Newton, I. (1687). *Philosophiae Naturalis Principia Mathematica.* Réédition, Cambridge University Press. Trad. fr. (E. du Châtelet), *Principes mathématiques de la philosophie naturelle*, Paris, Jacques Gabay, 1990.

Newton, I. (1730). *Opticks.* Dover, New York, 1952. Trad. fr. (J.-P. Marat), *Optique*, Paris, Bourgois, 1985.

Oakley, D. A. éd. (1985). *Brain and mind.* Methuen, Londres.

Obermayer, K., Teich, W. G. et Mahler, G. (1988*a*). Structural basis of multistationary quantum systems. I. Effective single-particle dynamics. *Phys. Rev.*, **B37**, 8096-8110.

Obermayer, K., Teich, W. G. et Mahler, G. (1988*b*). Structural basis of multistationary quantum systems. II. Effective few-particle dynamics. *Phys. Rev.*, **B37**, 8111-8121.

Omnès, R. (1992). Consistent interpretations of quantum mechanics. *Rev. Mod. Phys.*, **64**, 339-382.

Pais, A. (1991). *Niels Bohr's times.* Clarendon Press, Oxford.

Pauling, L. (1964). The hydrate microcrystal theory of general anesthesia. *Anesth. Analg.*, **43**, 1.

Paz, J. P. et Zurek, W. H. (1993). Environment induced-decoherence, classicality and consistency in quantum histories. *Phys. Rev.*, **D48(6)**, 2728-2738.

Paz, J. P., Habib, S. et Zurek, W. H. (1993). Reduction of the wave packet : preferred observable and decoherence time scale. *Phys. Rev.*, **D47(2)**, 3ème série, 488-501.

Pearle, P. (1976). Reduction of the state-vector by a nonlinear Schrödinger equation. *Phys. Rev.*, **D13**, 857-868.

Pearle, P. (1989). Combining stochastic dynamical state-vector reduction with spontaneous localization. *Phys. Rev.*, **A39**, 2277-2289.

Pearle, P. (1992). Relativistic model for state-vector reduction. In *Quantum chaos, quantum measurement.* NATO AS1 Series C. Math. Phys. Sci. 358 (Copenhague 1991). Kluwer, Dordrecht.

Peat, F. D. (1988). *Superstrings and the search for the theory of everything.* Contemporary Books, Chicago.

Penrose, O. (1970). *Foundations of statistical mechanics : a deductive treatment.* Pergamon, Oxford.

Penrose, O. et Onsager, L. (1956). Bose-Einstein condensation and liquid helium. *Phys. Rev.*, **104**, 576-584.

Penrose, R. (1980). On Schwarzschild causality — a problem for « Lorentz covariant » general relativity. In *Essays in general relativity* (en hommage à A. Taub.) (F. J. Tipler éd.). Academic Press, New York, p. 1-12.

Penrose, R. (1987). Newton, quantum theory and reality. In *300 years of gravitation.* (S. W. Hawking et W. Israel éd.). Cambridge University Press.

Penrose, R. (1990). Author's response. *Behavioural and Brain Sciences*, **13(4)**, 692.

Penrose, R. (1991*a*). The mass of the classical vacuum. In *The philosophy of vacuum* (S. Saunders et H. R. Brown éd.). Clarendon Press, Oxford.

Penrose, R. (1991*b*). Response to Tony Dodd's « Gödel, Penrose, and the possibility of AI ». *Artificial Intelligence Review*, **5**, 235.

Penrose, R. (1993*a*). Gravity and quantum mechanics. In *General relativity and gravitation 1992. Proceedings of the Thirteenth International Conference on General Relativiy and Gravitation held at Cordoba, Argentina, 28 June-4 July 1992. Part 1 : Plenary lectures*, (R. J. Gleiser, C. N. Kozameh et O. M. Moreschi éd.). Institute of Physics Publications, Bristol.

Penrose, R. (1993*b*). Quantum non-locality and complex reality. In *The Renaissance of general relativity* (en l'honneur de D. W. Sciama) (G. Ellis, A. Lanza et J. Miller éd.). Cambridge University Press.

Penrose, R. (1993*c*). Setting the scene : the claim and the issues. In *The simulation of human intelligence* (D. Broadbent éd.). Blackwell, Oxford.

Penrose, R. (1993*d*). An emperor still without mind. *Behavioural and Brain Sciences*, **16**, 616-622.

Penrose, R. (1994*a*). On Bell non-locality without probabilities : some curious geometry. In *Quantum reflections* (en l'honneur de J. S. Bell) (J. Ellis et A. Amati éd.). Cambridge University Press.

Penrose, R. (1994*b*). Non-locality in and objectivity in quantum state reduction. In *Fudamental aspects of quantum theory* (J. Anandan et J. L. Safko éd.). World Scientific, Singapour.

Penrose, R. et Rindler, W. (1984). *Spinors and space-time*, vol. 1 : *Two-spinor calculus and relative fields.* Cambridge University Press.

Penrose, R. et Rindler, W. (1986). *Spinors and space-time*, vol. 1 : *Spinor and twistor methods in space-time geometry.* Cambridge University Press.

Percival, I. C. (1994). Primary state diffusion. *Proc. R. Soc. Lond.*, **A**.

Peres, A. (1985). Reversible logic and quantum computers. *Phys. Rev.*, **A32(6)**, 3266-3276.

Peres, A. (1990). Incompatible results of quantum measurements. *Phys. Lett.*, **A151**, 107-108.

Peres, A. (1991). Two simple proofs of the Kochen-Specker theorem. *J. Phys. A : Math. Gen*, **24**, L175-L178.

Perlis, D. (1990). The emperor's old hat. *Behavioural and Brain Sciences*, **13(4)**, 680.

Planck, M. (1906). *The theory of heat radiation* (d'après des conférences données à Berlin en 1906-1907 et traduites de l'allemand par M. Masius). Dover, New York, 1959.

Popper, K. et Eccles, J. R. (1977). *The self and its brain.* Springer International.

Post, E. L. (1936). Finite combinatory processes-formulation I. *Jour. Symbolic Logic*, **1**, 103-105.

Poundstone, W. (1985). *The recursive universe : cosmic complexity and the limits of scientific knowledge.* Oxford University Press.

Pour-EL, M. B. (1974). Abstrat computability and its relation to the general purpose analog computer. (Some connections between logic, differential equations and analog computers.) *Trans. Amer. Math. Soc.*, **119**, 1-28.

Pour-EL, M. B. et Richards, I. (1979). A computable ordinary dfferential equation which possesses no computable solution. *Ann. Math. Logic*, **17**, 61-90.

Pour-EL, M. B. et Richards, I. (1981). The wave equation with computable initial data such that its unique solution is not computable. *Adv. in Math.*, **39**, 215-239.

Pour-EL, M. B. et Richards, I. (1982). Noncomputability in models of physical phenomena. *Int. J. Theor. Phys.*, **21**, 553-555.

Pour-EL, M. B. et Richards, I. (1989). *Computability in analysis and physics.* Springer-Verlag, Berlin.

Pribram, K. H. (1966). Some dimensions of remembering : steps toward a neuropsychological model of memory. In *Macromolecules and behaviour* (J. Gaito éd.). Academic Press, New York, p. 165-187.

Pribram, K. H. (1975). Toward a holonomic theory of perception. In *Gestalttheorie in der modernen Psychologie* (S. Ertel éd.). Erich Wengenroth, Cologne, p. 161-184.

Pribram, K. H. (1991). *Brain and perception : holonomy and structure in figural processing.* Lawrence Erlbaum Assoc., New Jersey.

Putnam, H. (1960). Minds and machines. In *Dimensions of mind* (S. Hook éd.), New York Symposium. Reproduit in *Minds and machines* (A. R. Anderson éd.), Prentice-Hall, 1964, p. 43-59 ; reproduit également in *Dimensions of mind : a symposium* (*Proceedings of the third annual NYU Institute of Philosophy*), NYU Press, 1964, p. 148-179.

Ramon y Cajal, S. (1955). *Studies on the cerebral cortex* (traduction de L. M. Kroft). Lloyd-Luke, Londres.

Redhead, M. L. G. (1987). *Incompleteness, nonlocality, and realism.* Clarendon Press, Oxford.

Rosenblatt, F. (1962). *Principles of neurodynamics.* Spartan Books, New York.

Roskies, A. (1990). Seeing truth or just seeming true ? *Behavioural and Brain Sciences*, **13(4)**, 682.

Rosser, J. B. (1936). Extensions of some theorems of Gödel and Church. *Jour. Symbolic Logic*, **1**, 87-91.

Rubel, L. A. (1985). The brain as an analog computer. *J. Theoret. Neurobiol.*, **4**, 73-81.

Rubel, L. A. (1988). Some mathematical limitations of the general purpose analog computer. *Adv. in Appl. Math.*, **9**, 22-34.

Rubel, L. A. (1989). Digital simulation of analog computation and Church's thesis. *Jour. Symbolic Logic*, **54(3)**, 1011-1017.

Rucker, R. (1984). *Infinity and the mind : the science and philosophy of the infinite.* Paladin Books, Granada Publishing Ltd, Londres. (Première publication par Harvester Press Ltd, 1982.)

Sacks, O. (1973). *Awakenings.* Duckworth, Londres. Trad. fr. (Ch. Cler), *L'Éveil*, Paris, Le Seuil, 1993.

Sacks, O. (1985). *The man who mistook his wife for a hat.* Duckworth, Londres. Trad. fr. (É. de la Héronnière), *L'Homme qui prenait sa femme pour un chapeau*, Paris, Le Seuil, 1992.

Sagan, L. (1967). On the origin of mitosing cells. *J. Theor. Biol.*, **14**, 225-274.

Sakharov, A. D. (1967). Vacuum quantum fluctuations in curved space and the theory of gravitation. *Doklady Akad. Nauk. URSS*, **177**, 70-71. Trad. angl. *Sov. Phys. Doklady*, **12**, 1040-1041 (1968).

Schrödinger, E. (1935*a*). Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften*, **23**, 807-812, 823-828, 844-849. Trad. angl. (J. T. Trimmer) in *Proc. Amer. Phil. Soc.*, **124**, 323-338. In *Quantum theory and measurement* (J. A. Wheeler et W. H. Zurek éd.). Princeton University Press, 1983.

Schrödinger, E. (1935*b*). Probability relations between separated systems. *Proc. Camb. Phil. Soc.*, **31**, 555-563.

Schrödinger, E. (1967). *What is Life ?* et *Mind and Matter.* Cambridge University Press. *Qu'est-ce que la vie ?*, L. Keffer trad., Paris, Le Seuil, 1993. *L'Esprit et la matière*, M. Bitbol trad., Paris, Le Seuil, 1990.

Schroeder, M. (1991). *Fractals, chaos, power laws. Minutes from an infinite paradise.* Freeman, New York.

Scott, A. C. (1973). Information processing in dendritic trees. *Math. Bio. Sci.*, **18**, 153-160.

Scott, A. C. (1977). *Neurophysics.* Wiley Interscience, New York.

Searle, J. R. (1980). Minds, brains and programs. In *The behavioural and brain sciences*, vol. 3. Cambridge University Press. (Reproduit in *The Mind's I* (D. R. Hofstadter et D. C. Dennett éd.). Basic Books, Penguin, Harmondsworth, (Middlesex), 1981.)

Searle, J. R. (1992). *The rediscovery of the mind.* MIT Press, Cambridge (Massachusetts).

Seymore, J. et Norwood, D. (1993). A game for life. *New Scientist*, **139**, n° 1889, 23-26.

Sheng, D., Yang, J., Gong, C. et Holz, A. (1988). A new mechanism of high Tc superconductivity. *Phys. Lett.*, **A133**, 193-196.

Sloman, A. (1992). The emperor's real mind : review of Roger Penrose's *The Emperor's New Mind. Artificial Intelligence*, **56**, 355-396.

Smart, J. J. C. (1961). Gödel's theorem, Church's theorem and mechanism. *Synthèse*, **13**, 105-110.

Smith, R. J. O. et Stephenson, J. (1975). *Computer simulation of continuous systems.* Cambridge University Press.

Smith, S., Watt, R. C. et Hameroff, S. R. (1984). Cellular automata in cytoskeletal lattice proteins. *Physica D*, **10**, 168-174.

Smolin, L. (1993). What have we learned from non-perturbative quantum gravity ? In *General relativity and gravitation 1992. Proceedings of the Thirteenth International Conference on General Relativiy and Gravitation held at Cordoba, Argentina, 28 June-4 July 1992. Part 1 : Plenary lectures*, (R. J. Gleiser, C. N. Kozameh et O. M. Moreschi éd.). Institute of Physics Publications, Bristol.

Smolin, L. (1994). Time, structure and evolution in cosmology. In *Temponelle scienziae filosofia* (E. Agazzi éd.). World Scientific, Singapour.

Smorynski, C. (1975). *Handbook of mathematical logic.* North-Holland, Amsterdam.

Smorynski, C. (1983). « Big » news from Archimedes to Friedman. *Notices Amer. Math. Soc.*, **30**, 251-256.

Smullyan, R. (1961). *Theory of formal systems.* Princeton University Press.

Smullyan, R. (1992). *Gödel incompleteness theorem.* Oxford Logic Guide n° 19. Oxford University Press. Trad. fr.(M. Margenstern), *Les Theorèmes d'incomplétude de Gödel*, Paris, Masson, 1993.

Squires, E. J. (1986). *The mystery of the quantum world.* Adam Hilger Ltd, Bristol.

Squires, E. J. (1990). On an alleged proof of the quantum probability law. *Phys. Lett.*, **A145**, 67-68.

Squires, E. J. (1992*a*). Explicit collapse and superluminal signals. *Phys. Lett.*, **A163**, 356-368.

Squires, E. J. (1992*b*). History and many-worlds quantum theory. *Found. Phys. Lett.*, **5**, 279-290.

Stairs, A. (1983). Quantum logic, realism and value-definiteness. *Phil. Sci.*, **50(4)**, 578-602.

Stapp, H. P. (1979). Whiteheadian approach to quantum theory and the generalized Bell's theorem. *Found. Phys.*, **9**, 1-25.

Stapp, H. P. (1993). *Mind, matter, and quantum mechanics.* Springer-Verlag, Berlin.

Steen, L. A. éd. (1978). *Mathematics today : twelve informal essays.* Springer-Verlag, Berlin.

Stoney, G. J. (1881). On the physical units of nature. *Phil. Mag.*, (Series 5) **11**, 381.

Stretton, A. O. W., Davis, R. E., Angstadt, J. D., Donmoyer, J. E., Johnson, C. D. et Meade, J. A. (1987). Nematode neurobiology using Ascaris as a model system. *J. Cellular Biochem.*, **511A**, 144.

Thorne, K. S. (1994). *Black holes & time warps : Einstein's outrageous legacy.* W. W. Norton and Company, New York.

Torrence, J. (1992). *The concept of nature. The Herbert Spencer lectures.* Clarendon Press, Oxford.

Tsotsos, J. K. (1990). Exactly which emperor is Penrose talking about ? *Behavioural and Brain Sciences*, **13(4)**, 686.

Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc.*, (series 2), **42**, 230-265 ; correction, **43**, 544-546.

Turing, A. M. (1939). Systems of logic based on ordinals. *Proc. Lond. Math. Soc.*, **45**, 161-228.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, **59**, n° 236 ; reproduit in *The Mind's I* (D. R. Hofstadter et D. C. Dennett éd.). Basic Books, Penguin, Harmondsworth (Middlesex), 1981.

Turing, A. M. (1986). Lecture to the London Mathematical Society on 20 February 1947. In *A. M. Turing's ACE report of 1946 and other papers* (B. E. Carpenter et R. W. Doran éd.). The Charles Babbage Institute, vol. 10, MIT Press, Cambridge (Massachusetts).

Waltz, D. L. (1982). Artificial intelligence. *Scientific American*, **247**(4), 101-122.

Wang, Hao (1974). *From mathematics to philosophy.* Routledge, Londres.

Wang, Hao (1987). *Reflections on Kurt Gödel.* MIT Press, Cambridge (Massachusetts).

Wang, Hao (1993). On physicalism and algorithmism : can machines think ? *Philosophia mathematica* (series III), 97-138.

Ward, R. S. et Wells, R. O. Jr (1990). *Twistor geometry and field theory.* Cambridge University Press.

Weber, J. (1960). Detection and generation of gravitational waves. *Phys. Rev.*, **117**, 306.

Weinberg, S. (1960). *The first three minutes : a modern view of the origin of the uiverse.* André Deutsch, Londres. Trad. fr., *Les trois premières minutes de l'Univers*, Paris, Le Seuil, 1978.

Werbos, P. (1989). Bell's theorem ; the forgotten loophole and how to exploit it. In *Bell's theorem, quantum theory, and conceptions of the universe* (M. Kafatos éd.). Kluwer, Dordrecht.

Wheeler, J. A. (1957). Assessment of Everett's « relative state » formulation of quantum theory. *Rev. Mod. Phys.*, **29**, 463-465.

Wheeler, J. A. (1975). On the nature of quantum geometrodynamics. *Annals of Phys.*, **2**, 604-614.

Wigner, E. P. (1960). The unreasonable effectiveness of mathematics. *Commun. Pure Appl. Math.*, **13**, 1-14.

Wigner, E. P. (1961). Remarks on the mind-body question. In *The scientist speculates* (I. J. Good éd.). Heinemann, Londres. (Reproduit in E. Wigner (1967), *Symmetries and reflections.* Indiana University Press, Bloomington ; et in *Quantum theory and measurement* (J. A. Wheeler et W. H. Zurek éd.). Princeton University Press, 1983).

Wilensky, R. (1990). Computability, consciousness and algorithms. *Behavioural and Brain Sciences*, **13(4)**, 690.

Will, C. (1988). *Was Einstein right ? Putting general relativity to the test.* Oxford University Press. Trad. fr. (F. Balibar et M. Biezunski), *Les Enfants d'Einstein : la relativité générale à l'épreuve de l'observation.* Paris, InterÉditions, 1988.

Wolpert, L. (1992). *The unnatural nature of science.* Faber and Faber, Londres.

Woolley, B. (1992). *Virtual worlds.* Blackwell, Oxford.

Wykes, A. (1969). *Doctor Cardano. Physician extraordinary.* Frederick Muller.

Young, A. M. (1990). *Mathematics, physics and reality.* Robert Briggs Associates, Portland (Orégon).

Zeilinger, A., Gaehler, R., Shull, C. G. et Mampe, W. (1988). Single and double slit diffraction of neutrons. *Rev. Mod. Phys.*, **60**, 1067.

Zeilinger, A., Horne, M. A. et Greenberger, D. M. (1992). Higher-order quantum entanglement. In *Squeezed states and quantum uncertainty*, (D. Han, Y. S. Kim et W. W. Zachary éd.). NASA Conf. Publ. 3135. NASA, Washington (DC).

Zeilinger, A., Zukowski, M., Horne, M. A., Bernstein, H. J. et Greenberger, D. M. (1994). Einstein-Podolsky-Rosen correlations in higher dimensions. In *Fudamental aspects of quantum theory* (J. Anandan et J. L. Safko éd.). World Scientific, Singapour.

Zimba, J. (1993). Finitary proofs of contextuality and nonlocality using Majorana representation of spin-3/2 states, M.Sc. thesis, Oxford.

Zimba, J. et Penrose, R. (1993). On Bell non-locality without probabilities : more curious geometry. *Stud. Hist. Phil. Sci.*, **24(5)**, 697-720.

Zohar, D. (1990). *The quantum self. Human nature and consciousness defined by the New Physics.* William Morrow and Company, Inc., New York.

Trad. fr. (P. Couturiau), *Conscience et science contemporaine : le moi quantique.* Monaco, Le Rocher, 1992.

Zohar, D. et Marshall, I. (1994). *The quantum society. Mind, Physics and a new social vision.* Bloomsbury, Londres.

Zurek, W. H. (1991). Decoherence and the transition from quantum to classical. *Physics Today*, **44**, n° 10, 36-44.

Zurek, W. H. (1993). Preferred states, predictability, classicality and the environment-induced decoherence. *Prog. of Theor. Phys.*, **89(2)**, 281-302.

Zurek, W. H., Habib, S. and Paz, J. P. (1993). Coherent states via decoherence. *Phys. Rev. Lett.*, **70(9)**, 1187-1190.

Index

Achevé d'imprimer en octobre 1995
dans les ateliers de Normandie Roto Impression s.a.
61250 Lonrai
N° d'imprimeur : I5-2025
Dépôt légal : octobre 1995

ROGER PENROSE

LES OMBRES DE L'ESPRIT

A la recherche d'une science de la conscience

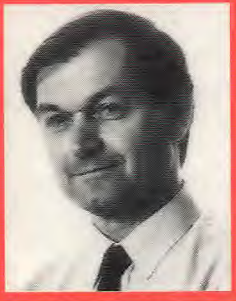
Traduit de l'anglais par Christian Jeanmougin

La compréhension de la conscience appartient au champ d'action de la science. L'auteur en est persuadé. Mais faut-il pour cela penser que nous disposons déjà de tous les éléments nécessaires ? Le cadre scientifique actuel suffit-il à l'explication des mécanismes de l'esprit ?

Roger Penrose articule sa pensée selon deux axes. Il montre d'abord, en se fondant sur des arguments mathématiques liés au théorème de Gödel, que la pensée consciente est irréductible au calcul. Aucun ordinateur ne pourra donc jamais émuler la conscience humaine.

Il explique ensuite la nécessité d'une nouvelle physique pour comprendre l'esprit. Sa thèse repose sur une réexposition des principes fondamentaux de la mécanique quantique et l'affirmation que l'activité cérébrale n'est pas uniquement d'ordre déterministe. C'est en se dotant de nouveaux outils conceptuels que la science percera un jour le mystère de la conscience.

Outre l'intérêt extrême de son sujet, ce livre constitue une superbe occasion de se familiariser avec les théorèmes fondateurs des mathématiques du XX^{e} siècle ainsi qu'avec les principaux concepts de la physique et les derniers développements de la neurobiologie.



Professeur à l'université d'Oxford, Roger Penrose, mathématicien et physicien de réputation mondiale – lauréat avec Stephen Hawking du prix Wolf en 1988 – poursuit ici la quête commencée avec son précédent ouvrage L'Esprit, l'ordinateur et les lois de la physique.

POPL Image I. Bourland

9 782729 605582

INTEREDITIONS

ISBN 2 7296 0558 4