Université de Montréal

Les limitations imposées par le théorème de Gödel aux machines pensantes

par
Alexandre Brunet
Département de philosophie
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès Arts (M. A.)
en philosophie

août 2001

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :
Les limitations imposées par le théorème de Gödel
aux machines pensantes

présenté par :
Alexandre Brunet

a été évalué par un jury composé des personnes suivantes :

Mémoire accepté le : _____

SOMMAIRE

Ce mémoire a pour sujet la possibilité que le théorème de Gödel implique une réfutation du Mécanisme, à savoir la thèse qui soutient qu'une machine numérique puisse reproduire le comportement, voire le fonctionnement de l'esprit humain. Le théorème de Gödel en lui-même n'affirme bien sûr rien de tel. Le premier chapitre est consacré justement à un rappel de la démonstration de Gödel prouvant l'incomplétude de tout système formel suffisamment fort pour contenir l'arithmétique de Peano. De plus, l'équivalence qui existe entre un système formel et une machine de Turing sera présentée. Sous certaines réserves, on arrive ainsi à montrer que pour toute machine numérique, tel un ordinateur, il existe une proposition mathématique que nous, humains, savons vraie mais que la machine ne peut démontrer.

L'argument de Lucas, présenté dans le chapitre 4, a l'ambition de réfuter le Mécanisme en montrant que la précédente limitation imposée par le théorème de Gödel s'applique à toute machine alors qu'elle ne s'applique pas à l'esprit humain. Or il s'avère que le Mécanisme se décline en deux versions, qui doivent chacune être contestée si l'on veut adéquatement réfuter le Mécanisme. Cette distinction, qui fera l'objet du second chapitre, influence grandement la portée de tout argument s'appuyant sur le théorème de Gödel. Nous verrons que la version forte du Mécanisme soutient que l'esprit *est* une machine, alors que la version faible ne se compromet pas sur la nature de l'esprit, mais soutient tout de même que son comportement peut être reproduit.

La distinction entre les deux versions du Mécanisme limite la portée de l'argument de Lucas à l'interprétation faible. Qui plus est, les objections auxquelles il doit faire face sont sérieuses. Principalement, l'argument de Lucas semble négliger que la trop grande complexité du système formel équivalent à une machine pensante pourrait très bien empêcher l'esprit humain de connaître la proposition gödelienne vraie mais indémontrable par la machine. En fait, il faut comprendre que l'argument de Lucas s'appuie sur ce qu'*en principe* tout être humain peut faire, alors qu'une catégorie d'objections se concentre sur ce qui peut être fait en pratique. D'autres objections pertinentes peuvent s'ajouter, tant et si bien que l'argument de Lucas ne s'en sort pas indemne.

Paul Benacerraf s'intéresse pour sa part à la contrepartie négligée par l'argument de Lucas, c'est-à-dire la version forte. Comme nous le verrons dans le chapitre 5, il présente un raisonnement formalisé qui conclut que la limitation du théorème de Gödel oblige tant l'esprit que les machines pensantes à ignorer leur propre algorithme précis. Ce résultat s'accorde bien avec les objections de l'argument de Lucas qui considèrent que la complexité de l'esprit humain prévient toute spécification complète de notre algorithme. La conclusion de Benacerraf est contestée par Lucas qui soutient qu'on ne peut garantir l'ignorance de notre algorithme qu'en niant tout bonnement que nous sommes des machines. Mais cette riposte est déboutée par un autre raisonnement formel qui montre que l'ignorance peut être garantie par l'impossibilité d'une méthode algorithmique permettant d'énumérer les nombres ordinaux transfinis.

Le chapitre 6 montre comment Roger Penrose prend la relève de Lucas mais sans vraiment l'outrepasser. Les mêmes objections propres à l'interprétation faible sont au rendez-vous, et Penrose y répond comme Lucas en soutenant la primauté de l'argument de principe sur les considérations pratiques. Selon lui, aucun type de « construction » de machines pensantes ne permettra d'échapper au fait que leurs algorithmes demeurent en principe connaissables et sûrs. Pour ce qui est de l'interprétation forte, Penrose reprend la riposte de Lucas à Benacerraf. Cette reprise est aussi discutable que l'originale.

Au bout du compte, s'il existe véritablement une supériorité de l'esprit humain sur les machines pensantes reposant sur le théorème de Gödel, ce ne peut être que dans le contexte de l'IA faible. Or la réfutation du Mécanisme s'avère peu convaincante dans ce contexte, les objections à l'argument de Lucas sont nombreuses et sérieuses. Pour ce qui est de l'IA forte, la position de Benacerraf se révèle tout à fait plausible. Le théorème de Gödel nous préviendrait d'avoir une connaissance complète de l'algorithme de notre esprit. Cette limitation de principe correspond somme toute assez bien à une limitation de pratique engendrée par notre incapacité actuelle à venir à bout de la complexité du cerveau humain.

TABLE DES MATIÈRES

Remerciements

Je désire remercier mon directeur de recherche François Lepage. Son soutien, aussi bien académique que financier, sa grande disponibilité et ses commentaires pertinents ont très certainement contribué à l'aboutissement de ce projet. La confiance qu'il m'a accordée tout au long de mes études m'a assurément encouragé à poursuivre une carrière universitaire.

Je remercie également le Fonds FCAR pour son aide financière sous forme de bourse de maîtrise.

À mes parents

Donnez-moi un levier et un point d'appui
et je soulèverai la Terre

Archimède

0. Présentation générale

Le « robo sapiens sapiens » est-il réellement une espèce en voie d'apparition? L'idée qu'un robot puisse un jour dépasser, autant physiquement qu'intellectuellement, le commun des mortels nous fait toujours un peu frémir. Alors que certains y voient l'évolution normale de l'intelligence artificielle (IA) et la promesse d'un soulagement de notre fardeau quotidien, plusieurs redoutent que ce dépassement ne se mue en un asservissement de l'espèce humaine. Plusieurs œuvres de science-fiction se sont chargées de dépeindre la vengeance des robots contre leurs créateurs, une vengeance engendrée par des années, voire des siècles d'esclavage au service des humains. Plus rarement imagine-t-on des robots pourvus d'une moralité irréprochable, d'une compassion exemplaire et d'une patience à toute épreuve envers notre race.

Le commun des mortels reste toutefois incrédule devant une telle utopie, qu'elle soit bienheureuse ou calamiteuse. C'est l'existence même de robots aussi perfectionnés qui est sérieusement mise en doute. Pour bien des sceptiques, le véritable talon d'Achille des machines serait leur incapacité d'avoir de véritables émotions ou d'être animées par un véritable libre-arbitre. Malgré toutes nos percées scientifiques concernant le fonctionnement de notre cerveau, nous n'avons toujours pas épuisé tout le mystère entourant l'esprit humain, si bien qu'ils sont encore nombreux ceux qui croient que le corps biologique et l'esprit sont deux. Quel type de rapport existe-t-il entre l'esprit, à savoir le principe pensant qui anime toute personne humaine, et le cerveau, que l'on conçoit communément comme le siège biologique de l'esprit? La question est loin d'être résolue, mais son incidence est pourtant cruciale en ce qui concerne la possible existence des robots. Imaginons un instant que seul un support essentiellement biologique permette l'existence de l'esprit. La possibilité de cerveau électronique ne serait pas pour autant réfutée, mais de tels cerveaux ne pourraient prétendre à une véritable intelligence.

Le mystère de l'existence de l'esprit humain nous prévient à bien des égards de déterminer avec certitude si nous allons un jour reproduire son fonctionnement grâce à l'intelligence artificielle. Mais si nous pouvions dégager une caractéristique ou une fonction de l'esprit qui soit suffisamment précise, peut-être pourrions-nous dès aujourd'hui établir que l'esprit n'est pas reproductible par des moyens mécaniques? Considérons donc les mathématiques, une discipline dont nous croyons tout être humain

capable de maîtriser les rudiments. Qui plus est, il s'agit d'un domaine de connaissances réputé pour sa rigueur et son exactitude. Si une machine intelligente venait à être construite, elle se devrait, elle aussi, de comprendre adéquatement les mathématiques. C'est donc au caractère décisif de la compréhension des mathématiques que nous nous en remettrons pour déterminer s'il est possible qu'une machine soit aussi intelligente qu'un être humain.

C'est ici qu'entre en scène le célèbre théorème de Gödel, publié pour la première fois en 1931. Le théorème de Gödel n'est pas seulement le théorème le plus marquant de la logique mathématique contemporaine, il est sans doute celui dont la véritable signification est le plus discutée. En gros, il démontre que tout système formel consistant et contenant l'arithmétique de Peano est incapable de prouver un certain théorème vrai de même que sa propre consistance. Un tel système formel est donc considéré incomplet. Or toute machine programmée pouvant prouver des théorèmes sera équivalente à un système formel particulier. Donc toute machine sera incomplète. Certains y voient la preuve ultime que les prouesses de l'esprit humain ne sont pas le résultat d'une sorte de programme informatique exécuté par les neurones de notre cerveau, car, contrairement aux machines, l'esprit humain ne peut être incomplet, il peut percevoir la vérité de tout théorème sans nécessairement passer par une preuve formelle.

J. R. Lucas fut sans doute le premier à avoir défendu cette thèse avec ardeur et de manière systématique. Comment peut-il espérer réfuter le Mécanisme, à savoir la thèse soutenant que le fonctionnement ou le comportement de l'esprit humain peut être reproduit par une machine numérique? Il s'agit de montrer qu'une limitation telle que celle que nous venons de voir ne s'applique qu'aux machines. Par contre, si on montre que la limitation ne s'applique pas aux machines ou qu'elle s'applique aussi bien aux machines qu'aux humains, alors la réfutation échoue. Le présent mémoire examinera dans ce contexte les différents arguments contre le Mécanisme et les objections qu'ils n'ont pas manqué de susciter.

Parmi les principaux concernés par la possible réfutation de la thèse mécaniste se trouvent les spécialistes de l'intelligence artificielle (IA). Ceux-ci sont au jour le jour impliqués dans le projet à long terme de construire une machine qui soit l'égale de l'humain, une machine qui pense, une machine pensante. Ce projet se fonde

philosophiquement sur le Mécanisme, et c'est pourquoi la réfutation de l'un entraîne l'échec de l'autre et inversement. Si le Mécanisme semble à première vue univoque, nous allons réaliser dans le deuxième chapitre qu'il existe deux versions du projet de l'intelligence artificielle. En fait, ce projet pourrait pratiquement se décliner en autant d'interprétations et de versions qu'il existe de spécialistes! Nous allons quant à nous retenir un critère qui permettra de bien les distinguer. Dans l'interprétation forte, l'esprit *est* une machine pensante, alors que pour l'IA faible, l'esprit n'en est pas une, il est d'une autre nature qui ne se réduit pas au fonctionnement d'une machine. Cela n'empêche pas cependant une machine de *se comporter* extérieurement comme l'esprit le ferait.

Cette distinction s'avérera fort pertinente dans l'évaluation des différents arguments et de leurs objections. Les arguments ne réfutent pas nécessairement les deux versions de l'IA, et les objections s'appliquent à certains arguments selon la version de l'IA qui est en jeu. Et pourtant, la plupart des auteurs défendent leurs thèses sans prendre les précautions qui s'imposent. Il en résulte une confusion généralisée qui laisse plutôt perplexe même le lecteur aguerri quant à la validité et la portée de tous ces arguments. Ce mémoire visera donc à rendre limpide ce qui a toutes les apparences d'un grand fouillis pour bien juger de la valeur des arguments présentés.

Tout au long du mémoire, nous serons en quelque sorte obligés de faire certaines concessions à l'un ou l'autre des partis afin d'éviter de trop nous éloigner du sujet principal et de permettre aux arguments de se développer dans toute leur ampleur. Par exemple, la question se posera, dans la perspective de l'IA forte, de savoir si nous pouvons connaître de manière empirique l'algorithme de notre cerveau. Nous imaginons sans peine la très grande difficulté d'une telle tâche. Mais si nous nous gardons de concéder qu'une telle tâche est possible, à tout le moins en principe, alors une véritable comparaison entre les prouesses humaines et leurs équivalents artificiels est impossible. Par contre, rien ne sert de trop nous avancer dans l'examen d'une telle possibilité, elle ne ferait que retarder ce qui est pour nous d'un véritable intérêt, à savoir l'incidence du théorème de Gödel sur les machines pensantes.

En ce qui concerne la progression de ce travail, signalons que la présentation des arguments de Lucas et de ceux qui les ont revisités ne pouvait se faire sans un exposé

préalable de certaines notions. Le premier chapitre est donc consacré à une explication du théorème de Gödel et de ses liens avec les machines pensantes. Au second chapitre, ce sera la distinction entre les deux versions du Mécanisme ainsi que leurs implications et justifications respectives qui seront exposées. Un troisième chapitre sera consacré aux enjeux philosophiques que soulève la nature de la vérité mathématique. Dans ce chapitre, nous constaterons qu'un premier argument purement philosophique pourra être développé contre le Mécanisme. Mais plus décisifs se voudront les trois chapitres suivants consacrés respectivement aux argument de J. R. Lucas, Paul Benacerraf et Roger Penrose. L'argument de Lucas est un véritable classique et bien qu'il soit un peu « naïf » et vague, il contient l'essentiel de la charge que l'on peut faire contre la version faible de l'IA. Nous évaluerons la force de cet argument à la lumière de quatre catégories d'objections. De son côté, croyant reprendre l'argument de Lucas en l'améliorant, Benacerraf montre plutôt comment le théorème de Gödel limite les connaissances des machines pensantes dans la version forte de l'IA. Enfin, l'argument de Penrose se veut une réfutation complète du Mécanisme en reprenant à son compte les deux précédents arguments. Cependant, notre analyse révèlera que son argumentation n'est pas aussi décisive que Penrose peut le croire.

En somme, nous conclurons que le théorème de Gödel n'est pas une véritable entrave à l'apparition de « robo sapiens sapiens ». Si le théorème de Gödel impose une quelconque limitation aux machines artificielles pensantes, elle sera de toute évidence partagée par l'esprit humain dans le contexte de d'IA forte.

CHAPITRE 1.
LE THÉORÈME DE GÖDEL

1.1 Les deux théorèmes d'incomplétude de Gödel

Ce qu'on appelle couramment le théorème de Gödel réfère en fait à deux résultats de Gödel de 1931[1], les bien nommés premier et deuxième théorèmes d'incomplétude, qui concernent certains types de systèmes formels.[2] Bien que le deuxième théorème soit parfois cité sans mention du premier par l'un ou l'autre des participants au débat sur le Mécanisme, il faut bien réaliser qu'il ne peut être obtenu sans la démonstration du premier théorème. Dans ce qui suit, je présenterai une esquisse de leur preuve respective afin de mieux saisir leurs conditions et limites propres, sans toutefois mesurer pour l'instant leurs pleines conséquences sur le Mécanisme.

La démonstration du premier théorème de Gödel requiert un système formel F contenant les axiomes et les opérations de l'arithmétique de Peano en plus des opérations usuelles des fonctions de vérité, de quantification et d'identité. F est un système *formel* seulement s'il existe une procédure de décision qui permet de déterminer si une suite de formules $A_1,\ldots,A_n$ est véritablement une preuve de A_n, à savoir si A_i (où $1 \leq i \leq n$) est un axiome de F ou une formule obtenue à partir de ceux qui le précèdent dans la suite à l'aide de règles d'inférence. Par convention, on dira que A_n, le dernier terme de la suite, est un *théorème* de F.

Pour un tel système formel F, le premier théorème établit que si F est ω-consistant, alors il existe une formule indécidable $\mathcal{M}$ pour F. F est dit *ω-consistant* si pour toute formule $A(x)$ de F contenant une seule variable libre x, si F peut dériver $A(n)$ pour tout nombre naturel n, alors F ne peut dériver $(\exists x)\neg A(x)$. S'il n'existe aucune formule A tel que A et $\neg A$ sont à la fois des théorèmes de F, alors F est dit *consistant*. Par l'entremise de la tautologie $(\neg A \supset (A \supset B))$, on déduit aisément que F est consistant si et seulement s'il existe au moins une formule A qui n'est pas un théorème de F. Or la

---

[1] Pour une version française de l'article original allemand, voir Nagel et Newman 1989, pp. 105-143.

[2] Dorénavant, je parlerai *du* théorème de Gödel quand la distinction entre le premier et le second théorème n'est pas cruciale, dans la mesure où les deux démontrent *l'incomplétude* du type de système formel en question.

définition de l'ω-consistance implique justement qu'une certaine formule n'est pas un théorème de F, donc F est consistant. Autrement dit, l'ω-consistance de F implique sa consistance (mais la converse n'est pas valide). Une formule est dite *indécidable* s'il existe pour le système formel F en question une formule close A telle que ni A ni $\neg A$ n'est un théorème de F.[3] Peut-être puis-je ajouter qu'un tel résultat caractérise l'*incomplétude* de F qui, grossièrement, fait référence au fait qu'il aurait été souhaitable, pour des raisons qui seront plus claires bientôt, que la formule A ne soit pas indécidable.

Comment démontre-t-on le premier théorème? Étant donné que F contient l'arithmétique de Peano, il permet d'exprimer les concepts et opérations standards de la théorie des nombres naturels. Un des coups de génie de Gödel fut alors de coupler chaque symbole du vocabulaire de F à un entier naturel particulier, ce qui permit d'exprimer les formules de F à l'aide d'un nombre, appelé à juste titre nombre de Gödel. Ainsi, pour toute formule A de F, il existe un nombre de Gödel $g(A)$. Par la suite, on peut montrer qu'il existe une relation R_F parmi les nombres telle que $R_F[g(A_1,\ldots,A_n), g(A_n)]$ si et seulement si une suite $A_1,\ldots,A_n$ est une preuve de A_n dans F. La prochaine étape consiste à montrer qu'il existe un prédicat $B(x,y)$ définissable dans F tel que pour toute paire de nombre $\underline{m}$ et $\underline{n}$, $R_F(\underline{m},\underline{n})$ tient si et seulement si $B(m,n)$ est un théorème de F, où m et n sont les numéraux représentant les nombres $\underline{m}$ et $\underline{n}$ dans F. Cette dernière étape est à proprement dit le point culminant de l'*arithmétisation* de F, c'est-à-dire que la syntaxe (les règles de manipulation des symboles) de F est maintenant arithmétisée (ou formalisée) à l'intérieur même de F. Ainsi, F peut exprimer sa propre procédure de décision équivalant à sa capacité à établir les preuves de ses théorèmes. Conséquemment, les métathéorèmes à propos de F peuvent se traduire dans le langage de F.

Gödel montra alors qu'on pouvait construire dans F une formule $\mathcal{G}$ telle que $\vdash_F \mathcal{G} \equiv (\forall x)\neg B(x,g(\mathcal{G}))$. Sous l'interprétation standard de F, $\mathcal{G}$ est équivalent à la formule qui affirme qu'il n'existe pas de nombre naturel qui soit le nombre de Gödel de la preuve de $\mathcal{G}$, autrement dit qu'il n'existe pas de preuve de $\mathcal{G}$ dans F. Ainsi, $\mathcal{G}$ affirme de lui-même qu'il n'est pas prouvable dans F. Nous reviendrons plus loin sur

[3] Mendelson 1997, p. 206

l'interprétation de $\mathcal{H}$ et sa parenté avec les paradoxes auto-référentiels tels que le paradoxe du menteur. Pour l'instant, on désignera $\mathcal{H}$ comme étant une proposition *gödelienne* pour F.

Ce qui suit établira que la proposition gödelienne $\mathcal{H}$ est indécidable, c'est-à-dire que non- $\vdash_F \mathcal{H}$ et que non- $\vdash_F \neg\mathcal{H}$ .Supposons dans un premier temps que F est consistant. Supposons en outre que $\vdash_F \mathcal{H}$ et que $g(p)$ est le nombre de Gödel de la preuve de $\mathcal{H}$ dans F. Ainsi, nous avons $\vdash_F B(g(p), g(\mathcal{H}))$. Mais par élimination de l'équivalence établie plus tôt, $\vdash_F (\forall x)\neg B(x,g(\mathcal{H}))$. Mais lorsque x prend la valeur de $g(p)$, nous avons $\vdash_F \neg B(g(p),g(\mathcal{H}))$, ce qui génère une contradiction. Donc,

(1) si F est consistant, alors non- $\vdash_F \mathcal{H}$.

Cette première conclusion aura une importance accrue dans la preuve du second théorème. Dans un second temps, supposons que F est ω-consistant et que $\vdash_F \neg\mathcal{H}$. Par élimination de l'équivalence, nous avons $\vdash_F \neg(\forall x)\neg B(x,g(\mathcal{H}))$ et par équivalence des quantificateurs, $\vdash_F (\exists x)B(x,g(\mathcal{H}))$. Or F est aussi consistant et donc non- $\vdash_F \mathcal{H}$, c'est-à-dire qu'il n'existe pas de preuve de $\mathcal{H}$ dans F. Donc, $B(n,g(\mathcal{H}))$ doit être faux pour tout nombre n, et ainsi $\vdash_F \neg B(n,g(\mathcal{H}))$ pour tout n. Par ω-consistance, nous déduisons que $\vdash_F \neg(\exists x)B(x,g(\mathcal{H}))$, ce qui génère à nouveau une contradiction. Seconde conclusion :

(2) si F est ω-consistant, alors non- $\vdash_F \neg\mathcal{H}$.

De (1) et de (2) nous concluons :

(3) si F est ω-consistant, alors $\mathcal{H}$ est indécidable dans F.

Le premier théorème de Gödel vient d'être démontré.

Pour ce qui est du second théorème, nous devons tout d'abord considérer n'importe quel théorème de F, soit $0\neq1$. Autrement dit, $\vdash_F 0\neq1$. Maintenant, si en plus $\vdash_F \neg(0\neq1)$, alors F est inconsistant. Par ailleurs, si non- $\vdash_F \neg(0\neq1)$, F est consistant.[4]

[4] Il suffit qu'il existe une seule formule qui ne soit pas un théorème de F pour que celui-ci ne soit pas inconsistant.

Ainsi, l'assertion selon laquelle $\neg(0\neq1)$ n'est pas un théorème est équivalente à l'assertion de la consistance de F.[5] Mais il est possible d'exprimer la première assertion dans F comme suit : $(\forall x)\neg B(x,g(\neg(0\neq1)))$. Comme cette formule est équivalente à l'assertion de la consistance de F, nous pouvons introduire la nouvelle notation « Con_F » pour exprimer cette formule de façon abrégée. Or, puisque F est formalisé à l'intérieur même de F, la proposition (1) s'exprime par la formule « $Con_F \supset \mathcal{G}$ ». Cette formule est aussi un théorème de F. Mais voici où le bât blesse : si $\vdash_F Con_F$, alors par modus ponens $\vdash_F \mathcal{G}$, et ainsi F deviendrait inconsistant.[6] Voilà donc la conclusion du second théorème de Gödel :

(4) si F est consistant, Con_F ne peut être démontré par F.

L'incomplétude de F se présente donc sous deux aspects. Tout d'abord, selon le premier théorème de Gödel, la proposition $\mathcal{G}$ n'est pas prouvable, mais comme $\mathcal{G}$ affirme d'elle-même qu'elle n'est pas prouvable dans F, $\mathcal{G}$ est vraie sous l'interprétation standard de F. Autrement dit, F est incapable de prouver tout ce qui est vrai à son sujet. Il importe de noter que F est *essentiellement* incomplet s'il est consistant, c'est-à-dire que toute extension de F demeure incomplète. Il ne suffit pas d'ajouter simplement $\mathcal{G}$ aux axiomes de F pour rendre F complet. Car, ce faisant, F n'est plus F mais devient plutôt F^*. L'arithmétisation de F^* sera différente de celle de F, générant une proposition gödelienne $\mathcal{G}^*$ différente de $\mathcal{G}$, mais tout de même indécidable. Ainsi, F^* demeure incomplet.

Pour sa part, le second théorème d'incomplétude affirme que pour tout système formel F « suffisamment fort »[7], s'il a la propriété d'être consistant, il ne peut démontrer lui-même qu'il possède bel et bien cette propriété. F est donc incomplet dans la mesure où on aurait souhaité (en particulier dans le cadre du programme de Hilbert) qu'il puisse

---

[5] Nous avons utilisé ici $\neg(0\neq1)$, mais nous aurions pu utiliser toute négation d'un théorème de l'arithmétique. Autrement dit, il existe une infinité de formules équivalentes à l'assertion de la consistance de F.

[6] Pour s'en convaincre, on n'a qu'à reprendre le raisonnement de la conclusion (1) du premier théorème de Gödel.

[7] « Suffisamment fort », « suffisamment vaste », ces expressions peuvent paraître vagues, mais nous avons déjà vu qu'elles peuvent être rendues précises.

démontrer une caractéristique aussi décisive concernant l'ensemble des théorèmes qu'il peut dériver. Une preuve mathématique de la consistance de l'arithmétique de Peano est possible, mais elle utilisera des notions et des méthodes qui ne seront pas formalisables dans F représentant la théorie des nombres naturels. Il en va ainsi entre autres des preuves de Gentzen de 1936 et 1938.[8]

1.2 Systèmes formels et machines de Turing

En tant que tel, le système formel F précédent paraît ridiculement limité quand il s'agit de répondre aux desseins du Mécanisme. Les discussions entourant les enjeux du Mécanisme postulent plus souvent l'existence d'une machine ou d'un ordinateur sophistiqué dont le comportement serait suffisamment apparenté à celui d'un humain pour rendre la thèse mécaniste digne d'intérêt. Pourtant, il existe un lien entre toute machine actuelle ou future construite dans le cadre des recherches en intelligence artificielle et tout système formel capable de démontrer minimalement tous les théorèmes de F.[9] Ce lien se fonde sur l'équivalence entre les systèmes formels et ce qu'on appelle les machines de Turing, qui sont des versions idéalisées d'ordinateurs concrets. Cette partie sera consacrée à la définition des machines de Turing et à l'explication de leur fonctionnement.

Avant d'aller plus loin, introduisons la notion d'algorithme. De façon générale, un algorithme est une procédure vouée à la résolution de problèmes d'une classe spécifique de manière systématique ou *effective*, c'est-à-dire selon un ensemble fini de règles précises. Si le problème en question a une solution, la procédure doit y parvenir en un nombre fini d'étapes et de façon déterministe, c'est-à-dire qu'à chaque étape de la résolution, la prochaine étape à accomplir est bien déterminée, jusqu'à ce que le résultat complet et final soit établi.

Maintenant, une machine de Turing se veut la réalisation mathématique d'un algorithme. On dit alors que la machine *calcule* le résultat (output) d'un problème mathématique à partir de données de départ (inputs). Le fonctionnement d'une telle machine fut imaginé par Alan Turing dans les années 1930. Il s'agit d'un automate

---

[8] Gentzen 1969

[9] L'expression « système formel » réfèrera désormais à tout système formel répondant à cette caractéristique minimale.

abstrait défini mathématiquement que l'on imagine capable de lire et d'écrire sur un ruban de longueur indéfinie, divisé en cases et qui, au départ, contient une suite finie de symboles (un par case) suivie de cases vides. Le ruban sert de médium à la fois pour entrer la chaîne finie des données codifiées que l'on veut qu'elle traite (input), pour stocker certains résultats intermédiaires durant le déroulement des opérations et pour afficher le résultat de son calcul une fois terminé. L'automate parcourt le ruban case par case en suivant pas à pas une procédure intégrée à la machine sous forme de table d'états internes. À chaque étape de la procédure, la table tient compte à la fois de ce qui peut être inscrit dans la case observée et de l'état interne discret de la machine afin de déterminer à la fois la prochaine action à poser (lire, écrire, effacer, parcourir le ruban dans un sens ou dans l'autre) et le prochain état interne de la machine (une certaine position dans la table des états internes). L'exécution de ce calcul par étapes discrètes se poursuit jusqu'à ce que la machine atteigne un état interne signalant l'arrêt du calcul. À ce moment, le résultat complet du calcul devrait être affiché sur le ruban. La machine est alors prête à recommencer un autre calcul.

Il existe une correspondance certaine entre la notion d'algorithme et la notion de calculabilité par l'entremise de machines de Turing. Plus qu'une vague correspondance, la thèse de Church, qui sera examinée plus loin, suggère une véritable équivalence entre les deux notions, si bien qu'aujourd'hui les notions d'algorithme, de procédure algorithmique ou mécanique, de calcul et de machine de Turing sont pratiquement interchangeables sans que l'on s'en formalise outre mesure. Le reste de ce mémoire profitera de cette opportunité pour varier le vocabulaire.

Bien que la table des états internes d'une machine de Turing fasse partie intégrante de sa structure, il est possible de reproduire sur un ruban la suite des instructions qu'elle peut contenir. Ainsi, des machines dites universelles peuvent calquer le fonctionnement de n'importe quelle machine de Turing, y compris elle-même. On n'a qu'à inscrire comme donnée de départ sur le ruban le code d'une machine particulière à imiter, puis ajouter à la suite les données de départ que celle-ci doit traiter.

Nous avons déjà noté qu'une machine de Turing est une version idéalisée, abstraite, purement mathématique d'un ordinateur concret. Notons par ailleurs que dans l'éventualité où nous pourrions à volonté étendre la capacité de stockage d'un

ordinateur, les ordinateurs personnels que nous utilisons chaque jour seraient de véritables machines de Turing universelles concrètes, au sens où ils pourraient théoriquement exécuter le calcul de n'importe quelle machine de Turing.

1.3 Correspondance conceptuelle entre machine de Turing et système formel

Il existe une équivalence fondamentale entre les machines de Turing et les systèmes formels appartenant à la classe à laquelle s'appliquent les théorèmes de Gödel. Les fonctions récursives sur lesquelles repose le processus d'arithmétisation des systèmes formels développé pour la première fois par Gödel sont équivalentes à la calculabilité des machines de Turing. Autrement dit, tout problème mathématique dont la solution peut être obtenue de manière algorithmique peut être formulé soit en terme de machine de Turing, soit sous forme de systèmes formels. Cela implique que le processus de démonstration de tous les théorèmes générés à partir d'un système formel peut devenir une procédure mécanique. Le raisonnement qui permet de démontrer le théorème de Gödel peut s'appliquer autant aux systèmes formels qu'aux machines de Turing. Ainsi, pour toute machine de Turing définie pour prouver les théorèmes d'un système formel équivalent, si l'ensemble de sa production est consistant, alors elle ne peut démontrer sa consistance.

Maintenant que nous sommes assez familiers avec le théorème de Gödel associé aux systèmes formels suffisamment forts, la question se pose de savoir quelle correspondance existe-t-il entre les notions associées aux systèmes formels et celles associées aux machines de Turing. Chacune d'entre celles-ci étant équivalente à un système formel particulier, il doit bien exister une façon d'exprimer les différents concepts liés au théorème de Gödel en tenant compte du mode de fonctionnement des machines de Turing.

Tout d'abord, voyons comment le raisonnement gödelien s'adapte aux machines de Turing. C'est à Turing, évidemment, que nous devons ce qui est convenu d'appeler le problème de l'*arrêt* et qui s'inspire directement des méthodes développées par Gödel, qui lui-même s'était inspiré de Cantor et de sa méthode de diagonalisation. L'arrêt d'une machine de Turing correspond à la fin du calcul qu'elle a effectué à partir bien sûr d'une donnée de départ qui figurait sur le ruban. Lorsqu'une machine s'arrête, il ne reste plus

qu'à consulter le ruban pour obtenir le résultat du calcul. On néglige ici le fait que la machine puisse se bloquer pour des raisons qui n'ont pas à voir directement avec la conclusion de l'exécution du calcul. Inversement, si une machine ne s'arrête jamais, cela voudra dire qu'il n'existe pas de solution au calcul qu'incarne la machine étant donné une certaine valeur de départ. Par exemple, si on demande à une machine $T(n)$ convenablement programmée de trouver le plus grand nombre naturel en commençant par un nombre n arbitraire, elle ne s'arrêtera jamais.[10]

Une machine que nous observons calculer sans pouvoir déterminer si elle s'arrêtera un jour n'est pas très satisfaisante car elle laisse toujours planer le doute sur la possibilité d'un arrêt imminent! C'est pourquoi nous supposerons l'existence d'un calcul A[11] que l'on applique à une machine $T(n)$, où n est la donnée de départ, et dont l'arrêt signifie que $T(n)$ ne s'arrête jamais. Nous supposerons bien sûr que le calcul A ne se trompe pas; de toutes façons, si A se trompe, nous pouvons en principe le vérifier, car il existera une valeur n pour laquelle $T(n)$ s'arrêtera.

Existe-t-il un calcul tel que A dont nous pouvons être certains qu'il ne se trompe pas et qui détermine si le calcul de la machine qu'on lui soumet ne s'arrête jamais? À cette question, Turing répond par la négative, et voici comment. Tout d'abord, il est possible d'énumérer tous les programmes[12] de machines de Turing possibles dont la valeur de départ est n :

$$T_1(n), T_2(n), T_3(n), T_4(n), T_5(n), T_6(n), \ldots, T_q(n), \ldots$$[13]

Considérons maintenant $A(q,n)$ le calcul qui s'arrête seulement si la machine $T_q(n)$ ne s'arrête jamais. Par diagonalisation, nous poserons que $q=n$. Nous avons maintenant :

(4) Si $A(n,n)$ s'arrête, alors $T_n(n)$ ne s'arrête pas.

---

[10] Évidemment, elle ne s'arrête pas théoriquement, alors qu'en pratique, si on construisait une telle machine, celle-ci finirait par briser (et donc s'arrêterait) au bout d'un certain temps d'utilisation.

[11] Nous suivons ici Penrose 1995 en désignant par « calcul A » toute machine de Turing dont la fonction est d'évaluer l'arrêt de toute machine $T(n)$, y compris elle-même.

[12] Par programme, on entendra la traduction de la table d'états internes d'une machine particulière en une suite de symboles pouvant être transcrite sur un ruban et pouvant être lue par une machine de Turing universelle. Suivant Penrose 1989, tout programme sera en relation bi-univoque avec un nombre binaire, celui-ci pouvant être à son tour converti en nombre naturel.

[13] L'énumération des programmes, basée sur la suite des entiers naturels 1,2,3,..., ne donne pas nécessairement que des machines qui fonctionnent. Pour différentes raisons, ces machines peuvent être défectueuses (Penrose 1995, p.73), mais il existe une méthode qui permet de les identifier comme telles et de ne conserver que les machines qui donnent des résultats ayant une réelle signification.

Cependant, comme l'énumération établie plus haut contient toutes les machines possibles, il existe une machine $T_k(n)$ dont le calcul est identique à $A(n,n)$. Mais avec $n=k$, nous avons $A(k,k) \Leftrightarrow T_k(k)$, ce qui entraîne que :

(5) Si $A(k,k)$ s'arrête, alors $T_k(k)$ ne s'arrête pas,

ou encore,

(6) Si $T_k(k)$ s'arrête, alors $T_k(k)$ ne s'arrête pas.

Cependant, comme l'antécédent implique sa négation, le conditionnel (6) est équivalent à son conséquent. Autrement dit, $T_k(k)$ ne s'arrête pas. Mais alors $A(k,k)$ non plus, puisqu'il est équivalent à $T_k(k)$. Donc $A(k,k)$ est incapable de détecter que $T_k(k)$ ne s'arrête pas, nous venons de démontrer l'incomplétude de $A(k,k)$.

Nous venons de présenter une première correspondance conceptuelle générale entre système formel et machine de Turing. La démonstration de Turing fait un usage équivalent de la méthode de diagonalisation qu'emploie Gödel dans ses théorèmes. Ainsi, on obtient un objet mathématique auto-référentiel comparable à la proposition indécidable $\mathcal{H}$ pour un système formel. Le calcul $A(k,k)$ devient son propre objet de calcul, puisque k est justement la valeur numérique du programme de la machine T_k qui est équivalente au calcul A pour une donnée de départ k. L'indécidabilité de la proposition $\mathcal{H}$ pour le système formel a pour équivalent l'indétermination de l'arrêt du calcul $T_k(k)$ pour le calcul $A(k,k)$.

La consistance est une autre caractéristique essentielle des systèmes formels qui nous intéressent. Cependant, on ne peut parler de la consistance d'une machine de Turing sans apporter quelques précisions. Tout d'abord, une contradiction n'affecte pas une machine de Turing comme elle affecte un système formel. Peut-on seulement parler de contradiction parmi les opérations de la machine? Est-ce que cela veut dire qu'à un moment précis la machine effectue à la fois une opération et son contraire? Au pire, une machine affligée d'une telle contradiction se bloquerait, « déchirée » entre deux états internes, mais on imagine mal qu'elle se mette à prouver n'importe quoi comme cela est permis dans un système formel inconsistant. En ce sens, la consistance est une notion qui n'a pas d'écho direct en ce qui concerne le fonctionnement interne des machines de Turing.

Néanmoins, on peut concentrer notre attention sur la production des machines de Turing, c'est-à-dire l'ensemble des résultats qu'elle produit au terme de son calcul. En ce sens, on pourrait qualifier une machine d'inconsistante si elle parvient à un résultat et, plus tard, à un résultat contradictoire. Par exemple, imaginons une machine dont la tâche spécifique et limitée serait de prouver les théorèmes d'un système formel inconsistant comme le ferait un logicien humain. Il y a tout à parier qu'une fois la contradiction mise à jour, la machine, tout comme le logicien humain, pourrait démontrer n'importe quelle proposition. Est-ce à dire que la machine est inconsistante? Dit-on du logicien humain qui utiliserait le même système formel qu'il est lui aussi inconsistant? On réalise que pour qualifier une machine d'inconsistante, il faut que sa table d'états internes ne lui permette qu'une seule tâche bien définie, à savoir la production de théorèmes d'un certain système formel inconsistant. Il faut donc établir clairement que ce n'est pas parce qu'une machine produit un ensemble inconsistant de théorèmes qu'elle est en elle-même inconsistante. Ainsi, par définition, aucune machine de Turing universelle (donc aucun ordinateur) n'est limitée à la production de théorèmes d'un seul système formel, au contraire les machines universelles peuvent reproduire le fonctionnement de n'importe quelle machine. Elles ne sont donc pas à strictement parler inconsistantes, même si leurs productions peuvent l'être. Dorénavant, lorsque nous parlerons de machine inconsistante, nous référerons à ces machines dont la table d'états internes est limitée à la production de théorèmes d'un système formel inconsistant, tout en sachant que le fonctionnement d'une telle machine peut être reproduit par une machine universelle.

De toute façon, en ce qui concerne le problème de l'arrêt des machines de Turing, on préfère parler de *sûreté*[14] plutôt que de consistance. Une machine A est dite sûre lorsqu'elle conclut que tel calcul $C_q(n)$ ne s'arrête jamais seulement si $C_q(n)$ ne s'arrête effectivement pas. En d'autres termes, une machine sûre ne se trompe pas, elle ne donne pour résultat[15] que des vérités. Tout comme la consistance d'un système formel doit être établie avant de pouvoir conclure que sa proposition gödelienne est vraie, on doit s'assurer de la sûreté de la machine A avant de conclure à l'incomplétude de cette machine de Turing qu'il existe une machine qui ne s'arrête pas et que A ne peut

[14] *Soundness* en anglais.
[15] Une machine de Turing ne donne de résultat que si elle s'arrête.

déterminer qu'elle ne s'arrête pas. En ce qui nous concerne, et en particulier dans l'argument de Penrose présenté plus tard dans le chapitre 6, la sûreté des machines de Turing est la condition suffisante qui équivaut à la consistance des systèmes formels pour pouvoir conclure à l'incomplétude. En fait, la sûreté est une condition plus forte que la consistance, puisque la sûreté d'une machine implique la consistance des résultats qu'elle produit, mais l'inverse n'est pas vrai. La raison en est que contrairement à la sûreté, la consistance ne garantit pas par définition la vérité des résultats sous une interprétation donnée. Si une machine est sûre, ses résultats sont vrais, ils ne peuvent donc générer une contradiction, la machine est donc consistante.

La conséquence la plus générale que l'on peut tirer de l'équivalence entre machine de Turing et système formel est que l'on n'a plus à se soucier des limitations inhérentes à la manière de formaliser du système formel dans lequel un certain problème est posé. Le théorème de Gödel qui, à l'origine, avait été développé pour certains systèmes formels spécifiques tels que celui des *Principia Mathematica*, gagne ainsi en généralité et devient applicable à tout système axiomatique représentable par une machine de Turing. Cette universalité sera implicitement invoquée lors des discussions qui suivront, notamment concernant la possibilité de découvrir le système formel incomplet à la base de toute machine.

CHAPITRE 2.

LES DEUX VERSIONS DU MÉCANISME

2.1 La thèse du Mécanisme

Dans le contexte de ce mémoire, le Mécanisme désignera la thèse qui soutient que l'esprit, son fonctionnement ou son comportement, peut être adéquatement reproduit ou simulé par une machine pouvant être conçue et construite par des humains. Le Mécanisme postule donc l'existence future de machines *pensantes* qui satisferaient les exigences de sa thèse. En ce sens, le Mécanisme est intimement associé aux recherches contemporaines en intelligence artificielle. On s'entend ordinairement pour dire qu'une telle machine serait un ordinateur suffisamment puissant. Il apparaît donc que « machine » réfère de manière abstraite à une machine de Turing, puisque tout ordinateur peut être représenté sous forme de machine de Turing. Cela expose donc les machines pensantes à des limitations similaires à celles imposées par le théorème de Gödel aux système formels. Une représentation complète de l'esprit est-elle possible face à de telles limitations? Est-ce toute la thèse du Mécanisme qui est réfutée par le théorème de Gödel? C'est ce que je tenterai de découvrir dans ce qui suit. Pour l'instant, je m'attarderai à présenter les tenants et aboutissants de la thèse mécaniste.

Une première mise au point s'impose. Il importe que le Mécanisme puisse, d'une certaine manière, prêter le flanc à une application du théorème de Gödel, sinon toute discussion devient vaine. Comme nous venons de le constater, le fait que l'on utilise actuellement des ordinateurs, c'est-à-dire des machines digitales ou numériques, pour représenter le fonctionnement de l'esprit rend notre parti pris assez plausible. Cette mise au point vise surtout à conjurer la possibilité que seule une machine essentiellement analogique, qui évolue donc de manière continue, puisse représenter le fonctionnement de l'esprit. Ce qu'il faut souligner ici est qu'il semble toujours possible de reproduire ou de simuler le continu physique par une représentation discrète, de sorte qu'un être humain n'est pas à même de discerner entre le continu et sa représentation discrète. En ce sens, toute mesure obtenue de manière analogique peut être remplacée par une approximation numérique suffisamment précise de telle sorte que la différence entre le

résultat d'un « calcul » analogique et le résultat d'un calcul numérique est indiscernable et donc négligeable.

2.2 Les racines philosophiques du Mécanisme

Le Mécanisme trouve ses sources philosophiques dans une forme de naturalisme souvent teintée de physicalisme. D'un point de vue ontologique, il s'appuie implicitement sur la conviction que tout comportement de la matière obéit strictement à des lois physiques. En ce sens, la construction matérielle d'une machine pensante ne devrait pas échapper à une description complète de son fonctionnement et de son comportement en termes physiques. D'un point de vue épistémologique, le Mécanisme assume que la description de tout phénomène naturel peut faire l'objet d'une formalisation mathématique en termes de lois physiques. Notre connaissance actuelle des lois de la nature laisse croire que le comportement de la matière est calculable, au sens où il peut être précisément modélisé ou simulé par un calcul effectué par une machine de Turing.

Cela nous amène à considérer le fonctionnement de notre esprit dans une perspective mécaniste. Constatant que notre cerveau est le siège de notre esprit, le Mécanisme conçoit qu'une certaine classe suffisante de phénomènes mentaux repose sur le comportement de la matière qui constitue notre cerveau. Ainsi, reproduire le fonctionnement de l'esprit ou simuler adéquatement son comportement devient possible au moyen d'une machine de Turing. À chaque phénomène mental correspond un pendant purement mécanique suffisamment conforme pour établir une équivalence entre les deux.

Cette conviction semble trancher dans un certain sens la question du réductionnisme, à savoir que l'on peut carrément se passer des concepts purement mentaux et s'en remettre uniquement sur les théories physiques pour expliquer le comportement humain. Mais comme nous le verrons bientôt, la distinction entre les interprétations forte et faible du projet de l'intelligence artificielle ne permet pas de conclure que le Mécanisme conduit à l'élimination pure et simple des concepts mentaux.

2.3 Deux interprétations de l'intelligence artificielle (IA)

Le Mécanisme est une des rares thèses philosophiques dont la plausibilité est directement reliée aux progrès réalisés dans un domaine scientifique particulier. Le Mécanisme est en effet le principal fondement philosophique des recherches en intelligence artificielle. Le projet de l'IA est véritablement la construction concrète d'une machine pensante. De façon générale, l'IA soutient qu' « au fur et à mesure que l'intelligence des machines progressera, les mécanismes qui la sous-tendent convergeront progressivement vers les mécanismes sous-tendant l'intelligence humaine »[16]. En d'autres termes, les phénomènes mentaux qui constituent l'intelligence humaine sont reproductibles par une machine de Turing, et plus cette machine réussira à bien reproduire ces phénomènes mentaux, plus l'intelligence d'une machine pensante se développera, plus son algorithme deviendra équivalent à l'algorithme qui est à l'œuvre dans notre esprit. Cette thèse de l'IA est assez vague, et plusieurs de ceux qui s'y intéressent ont senti le besoin de distinguer ce projet en deux versions suivant qu'il intègre ou non une forme de réductionnisme concernant la nature des phénomènes mentaux.

L'interprétation *forte* de l'IA est assurément la plus réductionniste des deux. Elle soutient que tout phénomène mental se réduit à une manipulation mécanique de symboles, autrement dit, à un algorithme calculable. En d'autres termes, le cerveau n'est pas le support privilégié et unique de la faculté de penser en général. Un support électronique, fait de circuits électriques, suffit pour héberger ce qui constitue l'essence du phénomène de la pensée, à savoir un algorithme sous forme de programme informatique reposant tout de même sur la capacité de calcul et de mémoire de son support. Il existe donc une interdépendance entre l'algorithme et le support, mais ce n'est pas pour autant une relation exclusive. Le même algorithme peut avoir plusieurs supports de plusieurs types, qu'ils soient biologiques ou électroniques, pourvu seulement qu'ils soient assez puissants, c'est-à-dire qu'ils aient les ressources suffisantes pour exécuter l'algorithme dans toute son ampleur.

Qu'en est-il donc des concepts psychologiques tels que la conscience de soi ou la faculté de jugement du vrai et du faux? Selon les tenants de l'IA forte, si on reconnaît

---

[16] Hofstadter 1985, p. 649

chez la machine des comportements qui laissent croire qu'elle possède bel et bien ces concepts (ou facultés) psychologiques qu'ordinairement l'on accorde volontiers aux humains qui présentent les mêmes comportements, alors elle les possède *réellement*, dans le même sens que *nous* les possédons. Si la machine affirme qu'elle a conscience d'elle-même, qu'elle exerce sa propre volonté, son propre libre-arbitre, qu'elle peut juger du vrai et du faux, on n'a d'autres choix que de la croire.

Dans le cas où il s'avère que l'algorithme ne contient pas à proprement dit une partie expressément construite pour rendre compte de chacune de ces facultés (l'algorithme ne contient pas une section modulaire qui sert spécifiquement à reproduire tous les comportements associés à un concept psychologique particulier), on peut admettre que ces facultés aient pu « émerger » de la complexité même de l'algorithme et de son exécution. Cette position émergentiste demeure réductionniste dans la mesure où les concepts psychologiques n'ont pas de réalité propre et irréductible, ils sont réductibles à des phénomènes physiques organisés selon un algorithme ayant atteint un certain niveau de complexité. On concède que cette position recèle pour l'instant un peu de mystère, mais cela ne devrait pas empêcher le projet de l'IA de se réaliser.

Ce qu'il faut retenir de cette interprétation, c'est qu'il ne fait aucun doute que l'esprit est en soi une machine algorithmique. Rien ne permet plus de croire en la spécificité de l'esprit, si ce n'est que son support (le cerveau) n'est pas électronique mais bien organique. Il est tout autant possible de construire un cerveau électronique équivalent au cerveau organique et doué des mêmes facultés.

L'interprétation forte de l'IA est en grande partie redevable à ce qu'on appelle le « test de Turing ». À la question « Est-ce que les machines peuvent penser? », Turing (1950) ne se risque pas à répondre et propose plutôt de mettre à l'œuvre des machines dans un jeu d'imitation. Le jeu mettrait en présence deux personnes, dont l'une est l'interrogateur, et une machine. L'interrogateur se trouve dans une salle séparée de celle fermée où sont réunis l'ordinateur et l'autre humain, désigné chacun par un pseudonyme. Le but de l'interrogateur est d'identifier correctement quel pseudonyme désigne la machine. Pour ce faire, il n'a droit que de poser des questions écrites à chacun des deux et il reçoit également leurs réponses par écrit. Évidemment, la stratégie de l'ordinateur est de tromper l'interrogateur et de réussir à se faire passer pour un humain.

Au moment de la rédaction de son article, Turing était plutôt optimiste et estimait qu'un demi-siècle de recherche en IA seraient suffisants pour atteindre un taux de réussite concluant. Dans ces circonstances :

> The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machine thinking without expecting to be contradicted.[17]

Ce que propose ici Turing est avant tout une théorie sur les concepts. L'usage d'un concept participe à sa définition, celle-ci évolue par conséquent dans le temps. Si aujourd'hui nous ne sommes pas prêts à admettre généralement que les machines peuvent penser, c'est que l'extension du concept n'inclut pas les machines. Le test que propose Turing influencerait notre usage du concept de pensée et celui-ci finirait par s'appliquer à une certaine classe de machines. Il en irait ainsi pour les autres concepts que l'on réserve aujourd'hui exclusivement à l'esprit humain, ils sont tous déterminés par l'usage que nous en faisons. Il s'agit d'un réductionnisme au sens où une fois qu'une classe de machines mérite de figurer dans l'extension d'un concept, ce concept peut être défini en termes purement mécaniques, si bien que ce qui faisait la singularité irréductible des phénomènes mentaux s'estompe. Cela revient à signer l'arrêt de mort de cette soi-disant caractéristique exclusive et restrictive qui empêchait les phénomènes mentaux d'être réduits à une description algorithmique.

Tournons-nous maintenant du côté de l'interprétation *faible* de l'IA qui est, à juste titre, moins ambitieuse dans ses prétentions. Le but est toujours de reproduire le fonctionnement de l'esprit et son comportement, mais « reproduire » doit être pris ici dans un sens béhavioriste, c'est-à-dire que ce ne sont que les comportements externes de la machine qui en bout de ligne nous intéressent, il n'y a aucun engagement ontologique concernant la réalité des concepts psychologiques. L'algorithme qui simule les manifestations publiques du fonctionnement de l'esprit est comparable à ce qui se passe réellement dans l'esprit. Si la machine nous dit qu'elle peut juger du bien et du mal, nous n'avons aucune raison de croire qu'elle possède véritablement cette faculté. D'un point de vue éthique par exemple, la machine peut très bien agir conformément au bien au moment même où elle affirme agir par devoir, mais cela ne fait pas d'elle un agent

[17] Turing, 1950, pp.13-14

moral équivalent à un agent humain véritablement déchiré entre le bien et le mal. Comme l'IA faible ne se prononce pas directement sur la réalité des concepts psychologiques chez les humains, la conscience et les différentes facultés mentales peuvent très bien être irréductibles à une description purement mécanique, et peut-être reposent-ils même sur une certaine spécificité biologique de notre cerveau. L'IA faible n'est donc pas aussi réductionniste que l'interprétation forte.

Ce qu'il faut donc retenir ici, c'est que l'interprétation faible ne considère pas que l'esprit est une machine pensante. L'esprit possède des caractéristiques internes qui lui sont propres. Aucune machine algorithmique n'est équivalente à l'esprit d'un point de vue interne, mais elle peut très bien lui être équivalente du point de vue de son comportement externe.

Si on reprenait le test de Turing, on conclurait de la réussite des machines qu'elles simulent bien les réponses que donnent les humains, mais cela ne suffit pas pour admettre qu'elles pensent. On pourrait s'attarder longuement sur l'interprétation que l'on peut accorder à la réussite des machines au test de Turing. John Searle a proposé un argument, dit de la chambre chinoise, qui met en doute les conclusions triomphalistes de Turing. Imaginons un homme qui ne parle que le français dans une chambre fermée et qui reçoit des questions rédigées en chinois par l'entremise d'une fente donnant sur l'extérieur. L'homme répond aux questions en suivant les instructions d'un algorithme écrit en français. Cet algorithme est si performant que l'homme arrive à tromper la plupart des interlocuteurs chinois sur sa propre ignorance du chinois. On peut facilement imaginer que l'homme pourrait être remplacé par un ordinateur exécutant le même algorithme. Il s'agit à peu de chose près d'un véritable jeu d'imitation, dont le but du jeu est de se faire passer pour un locuteur compétent en chinois. Et maintenant, la question cruciale : est-ce que l'homme, de même qu'un ordinateur qui suivrait le même algorithme, comprend la langue chinoise? La réponse de Searle est catégorique : non, ils ne comprennent pas le chinois, et par conséquent non, la manipulation mécanique de symboles ne suffit pas pour qu'il existe une représentation interne adéquate de la signification qui permette la prise de conscience par un humain de ce qu'il accomplit

lorsqu'il répond à une question.[18] Même si cet argument met sérieusement en doute la plausibilité de l'interprétation forte de l'IA, il laisse plutôt intact sa version faible.

2.4 Deux interprétations, deux types de limitations?

Cette distinction entre les deux interprétations de l'IA donne l'impression qu'il peut exister deux versions du Mécanisme. Si cela s'avère vrai, il faudra montrer comment le théorème de Gödel impose des limitations à chacune des deux versions. Mais la distinction est peut-être trompeuse, car elle ne distingue pas de façon générale le processus par lequel un ordinateur détermine les affirmations qu'il doit exprimer lorsqu'il est questionné. Dans les deux cas, la machine *calcule* ses réponses en fonction des questions. L'usage d'un algorithme est commun à toutes les machines, et la distinction entre IA forte et faible est avant tout de nature philosophique, elle concerne la nature de nos concepts mentaux. La question serait sans doute tranchée si nous parvenions à construire une machine pour laquelle on pourrait décider de façon empirique si elle se conforme aux exigences de l'IA forte plutôt qu'à celle de l'IA faible, ou vice versa. Mais pour l'instant, nos connaissances des « mécanismes » de la pensée demeurent fort partielles et ne laissent pas présager d'une telle possibilité.

En ce qui concerne les discussions sur les limitations imposées par le théorème de Gödel, le test de Turing peut-il intervenir? Nous constaterons plus tard que ces discussions doivent supposer que nous pouvons déterminer si la machine se trompe réellement lorsqu'elle nous répond, et que par le fait même elle ne doit pas « vouloir » nous tromper, comme c'est le cas avec le test de Turing. L'enjeu des discussions concernera la capacité des machines à affirmer toutes les vérités auxquelles nous, humains, avons accès en principe et pouvons affirmer. Si une machine affirme une fausseté, nous ne voulons pas avoir à nous demander si la machine *se* trompe ou qu'elle *nous* trompe de façon délibérée. Nous devons pouvoir déterminer sans aucune hésitation si la machine est en soi sûre ou non, si elle est dépourvue d'« intentions » malicieuses.

Maintenant, nous avons vu que chaque version détermine la nature de l'esprit, à savoir s'il est une machine pensante ou non. Comme la thèse du Mécanisme et les

[18] Évidemment, cette conclusion a été longuement débattue… mais cela n'a que très peu à voir avec le théorème de Gödel!

arguments que nous examinerons sont de nature philosophique, même si ces derniers reposent sur un théorème purement logique, et comme chaque version contient des considérations philosophiques différentes de l'autre, il y a fort à parier que les limitations soient différentes pour chaque version. Alors, quels sens et quelles portées accorder aux *limitations* imposées par le théorème de Gödel? Il faut comprendre que nous serons confrontés dans ce qui suit à plusieurs positions adverses concernant les limitations de l'esprit. Il y a ceux pour qui le théorème de Gödel n'a carrément rien à voir avec le débat sur le Mécanisme, c'est-à-dire qu'il ne permet pas de déterminer si en soi l'esprit est une machine ou non. C'est la position de la plupart des spécialistes de l'IA qui ne croient pas qu'un théorème purement logique puisse limiter ce que nous pouvons faire en pratique avec des machines. D'autres comme Lucas et Penrose sont plutôt persuadés que l'esprit n'est pas une machine algorithmique et que le théorème de Gödel est pertinent et même décisif en ce qui concerne le Mécanisme tel que soutenu aujourd'hui. Ainsi, les limitations qui sont imposées aux machines pensantes ne concernent pas directement l'esprit. Enfin, certains tel Benacerraf croient que l'esprit peut être une machine et que s'il en est ainsi, l'esprit est limité en ce qui concerne ce qu'il peut connaître, c'est-à-dire croire comme vrai.

Maintenant, comment évaluer chacune de ses positions, comment déterminer quels arguments sont décisifs? Notons tout d'abord que les trois arguments que nous allons examiner prennent pour hypothèse de départ que le Mécanisme est vrai, même si cela n'est pas toujours évident. Lucas est plutôt ambigu sur cette hypothèse de départ, Penrose beaucoup moins, mais il semble ne pas en tirer toutes les implications. Leur façon de formuler leurs arguments commence par suggérer qu'une machine pensante peut vraiment reproduire le comportement de l'esprit, qu'elle peut donc énoncer toutes les vérités de l'esprit. Mais l'esprit est-il pour autant une machine? En d'autres termes, dans quelle interprétation de l'IA sont-ils? Si l'esprit est une machine, il est sujet aux mêmes limitations, y compris celles du théorème de Gödel. En outre, les propriétés de l'esprit sont équivalentes à celles de la machine. S'il n'est pas une machine, comme le soutient l'IA faible, les limitations de l'esprit et de la machine sont assurément différentes, de même pour leurs propriétés respectives. Cela aura plusieurs conséquences sur la force des arguments que nous examinerons en temps et lieu.

2.5 Algorithmes ascendants et descendants

Un petit mot maintenant sur la façon dont nous pouvons construire un algorithme qui reproduirait le comportement humain. Il existe deux types distincts d'algorithme : les descendants et les ascendants. Évidemment, ces deux types d'algorithme peuvent être combinés pour donner des algorithmes mixtes, afin de combler les lacunes de chaque type. Nous y reviendrons.

De façon générale, les algorithmes descendants se veulent des structures de règles fixes et bien déterminées dont l'application est rigoureuse et sans aucune possibilité de changement. Lorsqu'on lui présente à plusieurs reprises un même input à traiter, on peut être sûr que le résultat sera toujours le même output. Si ce résultat n'est pas jugé satisfaisant (il est erroné ou insuffisamment précis), la tâche incombe aux programmeurs de retravailler l'algorithme, de modifier « manuellement » ses règles, pour qu'il produise un meilleur résultat, car la structure des algorithmes descendants ne peut pas varier par elle-même. C'est en grande partie sur ce point que divergent les algorithmes ascendants. Ceux-ci intègrent certaines (méta-)règles qui permettent de modifier la structure opératoire de l'algorithme sans intervention humaine. Cela permet à l'algorithme d'améliorer ses performances en corrigeant les erreurs qu'il a commises. Bien sûr, un être humain peut intervenir pour indiquer à l'algorithme ses erreurs, mais il ne lui indique pas directement comment il doit modifier sa structure opératoire une fois les erreurs identifiées, jusqu'au moment où l'algorithme ne fait plus d'erreur.

Les algorithmes ascendants sont beaucoup plus flexibles que ceux descendants, car d'une certaine manière il peut *apprendre* sans nécessairement nécessiter de « connaissances préalables » concernant la tâche qu'il doit effectuer. Mais cela ne les rend pas pour autant de manière absolue plus performants que les algorithmes descendants. Ceux-ci sont particulièrement performants, au sens où ils sont carrément plus performants que la plupart des humains, en ce qui concerne les problèmes mathématiques et les jeux dont les règles sont bien définies, mais les possibilités d'actions sont très nombreuses, comme c'est le cas pour le jeu d'échecs. Les algorithmes ascendants, quant à eux, sont souvent associés aux réseaux de neurones artificiels, qui

sont particulièrement efficaces dans la reconnaissance de patterns lorsque les critères qui permettent de traiter l'information sont vagues, insuffisamment précisés ou clairs.

Cette distinction parmi les deux types d'algorithmes ne devrait pas nous faire perdre de vue dans ce qui suit qu'ils sont tous les deux *exécutables par un ordinateur*. Autant les réseaux de neurones artificiels que les programmes de jeu d'échecs sont développés uniquement dans le cadre d'une utilisation informatique. En ce sens, autant les algorithmes descendants qu'ascendants peuvent être représentés sous forme de machine de Turing. On en conclut donc qu'en principe, les deux types d'algorithmes sont concernés par le théorème de Gödel.

En somme, qu'avons-nous accompli dans ce chapitre? Principalement, nous avons établi une distinction entre deux versions de la thèse mécaniste. Bien qu'elles aient des racines philosophiques communes, elles n'ont pas pour autant les mêmes conséquences sur la conception de l'esprit humain. Cela aura une importance cruciale lorsque viendra le temps de réfuter le Mécanisme, car la portée des limitations imposées par le théorème de Gödel n'est pas la même pour chaque version.

CHAPITRE 3.
PHILOSOPHIE ET VÉRITÉ MATHÉMATIQUE

3.1 Le statut de la vérité mathématique

La vérité mathématique jouera un rôle crucial dans les arguments contre le Mécanisme qui seront bientôt présentés. Chacun des arguments s'intéresse particulièrement à l'ensemble des vérités qu'une machine pensante pourrait affirmer. Est-ce que cet ensemble est égal à celui des humains? Existe-t-il au moins une vérité que la machine ne pourra jamais affirmer? Si tel est le cas, la machine ne peut être équivalente à l'esprit. Comme le théorème de Gödel est un résultat de logique mathématique, nous serons spécialement intéressés par la recherche d'une proposition mathématique vraie mais indémontrable pour chaque machine. Mais quelle est la nature de la vérité mathématique? Peut-elle être saisie par l'intuition humaine ou doit-elle être le résultat d'une démonstration dans un système formel? Ces différentes possibilités sont à la base de différentes conceptions des mathématiques que nous examinerons à l'instant.

Selon une de ces conceptions, la compréhension de la vérité mathématique pourrait très bien être exclusive à l'esprit humain. Nous le verrons, la conception réaliste est aussi associée de façon privilégiée au théorème de Gödel. Dans un certain sens, on pourrait considérer qu'il s'agit là d'un argument visant à limiter la portée du Mécanisme. Voilà ce qui nous attend dans ce chapitre.[19]

3.2 Deux philosophies des mathématiques

Pour les besoins de notre cause, nous nous contenterons de distinguer deux grandes philosophies des mathématiques sans entrer dans les nuances de chacune.[20] La définition de chacune des positions sera présentée de telle sorte que leur incompatibilité mutuelle puisse être bien évidente.

---

[19] Ce chapitre est grandement inspiré par les propos exposés dans Tieszen 1994 et dans Wang 1974.
[20] Tieszen (1994) inclut aussi l'intuitionnisme, mais cet apport n'a pas vraiment d'influence sur les conclusions qui sont utilisées ici.

Le *formalisme* soutient que les mathématiques se réduisent à la manipulation syntaxique de signes concrets et finis selon un ensemble fini d'axiomes et de règles d'inférence. Il propose essentiellement que les vérités mathématiques proviennent des théorèmes de tous les systèmes formels que nous pouvons construire. La signification que l'on doit accorder aux combinaisons de symboles qui composent les formules mathématiques ne réfère fondamentalement qu'à ces symboles en tant que tels, concrets et finis. Cette caractéristique distingue le formalisme des deux autres positions pour lesquels la signification des symboles utilisés dans une preuve mathématique renvoie à des entités abstraites, mentales ou non. Il n'y a rien de plus à comprendre dans une démonstration qu'un agencement particulier de symboles mathématiques, auxquels nous humains accordons une signification qui n'est pas essentielle pour affirmer qu'il s'agit bien là de mathématiques. Cette philosophie des mathématiques est communément associée au Mécanisme, car une machine correctement programmée pour construire des preuves mathématiques suivant un système formel dispose de tout le savoir-faire pour affirmer des vérités mathématiques au même sens que les humains.

Le formalisme se justifie parfois en faisant valoir que la théorie des ensembles (le système axiomatique ZF ou l'une de ses variantes) est suffisante pour reconstruire l'ensemble des théories dans les différents domaines des mathématiques. Autrement dit, tout le savoir mathématique actuel peut être traduit dans les termes de la théorie des ensembles. Alors, pourquoi chercher la vérité plus loin? Elle se trouve dans les démonstrations et les réfutations que l'on peut faire à l'intérieur du système formel ZF, sans rien de plus.

Le formalisme tel qu'il peut être défendu aujourd'hui est sans aucun doute l'héritier du programme de Hilbert. Celui-ci souhaitait fonder la totalité des mathématiques au moyen de méthodes finitistes. On sait maintenant, grâce au théorème de Gödel, que ce programme n'est pas réalisable dans sa version originale. Certaines versions moins radicales sont actuellement mises de l'avant, mais cela ne nous concerne pas vraiment.

Contrairement au formalisme, le *réalisme mathématique*, aussi appelé *platonisme*, soutient que la signification des propositions mathématiques vraies réfère à des objets mathématiques abstraits qui existent de façon indépendante de l'esprit. Nos

états mentaux intentionnels servent de pont entre les formules mathématiques écrites sur papier et les entités abstraites auxquelles les premières se réfèrent, ce qui nous permet de juger de leur vérité. La vérité mathématique est donc fondée sur l'existence d'entités abstraites ou « idéales » dans un monde indépendant de notre esprit et du monde physique, un monde qui les transcende, un monde où ils existent sans localisation spatio-temporelle concrète. Par un mystère que nous ne sommes pas prêts d'élucider, notre esprit humain a accès à ce monde comme s'il s'agissait d'un sixième sens intuitif, ce qui nous permet de comprendre la vérité mathématique. Notre intuition nous aide à *découvrir* les entités mathématiques et leur signification, elle ne les *invente* pas. Les systèmes formels ne sont que des outils formels qui rendent plus précis et valides les théorèmes mathématiques que nous avions conjecturés et qui nous rassurent ainsi sur la cohérence de nos vérités mathématiques. Le réalisme est donc en opposition radicale avec le formalisme. Le formaliste invente ses vérités mathématiques à la pointe de son crayon à l'aide de symboles concrets, alors que le réaliste les découvre grâce à la signification abstraite qu'il accorde à ces symboles.

3.3 Le formalisme et le théorème de Gödel

Rappelons ce que le formaliste attend d'une preuve mathématique. Une preuve est une suite de signes concrets qui s'agencent selon une procédure formalisée, précise et déterminée à l'avance et qui se conclut par une formule que l'on appelle théorème. Les propositions obtenues au terme de cette suite sont tout aussi « réelles » que les éléments primitifs qui les composent. Comme Hilbert le propose[21], les signes concrets doivent être donnés à notre faculté de perception (et non de raisonnement) antérieurement à toute pensée, c'est-à-dire que l'interprétation abstraite des signes concrets ne doit intervenir que postérieurement à l'élaboration de la preuve. La signification abstraite de ces signes se réduit à leur réalité concrète. Bien sûr, il n'est pas dit que nous devons éliminer purement et simplement le recours à la signification abstraite, mais nous devons réaliser qu'il s'agit simplement d'un artifice ou d'un raccourci pour nous aider à construire des preuves mathématiques.

---

[21] Tieszen 1994, p.182

Les machines de Turing satisfont les critères du formalisme (à la Hilbert) quant à la manipulation syntaxique de signes concrets, finis et discrets. Mais toute machine de Turing est équivalente à un système formel auquel on peut appliquer le second théorème de Gödel. Celui-ci démontre que la consistance d'un système formel F, représentée par le prédicat $Con(F)$, n'est pas démontrable dans F si celui-ci est consistant. Autrement dit, la démonstration de la consistance de F ne peut se faire au moyen de signes concrets, finis et discrets permis par le système F. En ce sens, seuls des objets mathématiques abstraits pour le système F permettent de prouver la consistance de F, tels que des ordinaux transfinis (dans le cas de la preuve de Gentzen).

Comme tout système formel sera affecté de la même façon par le second théorème de Gödel, l'approche formaliste semble inadéquate à représenter l'ampleur de la vérité mathématique. Il existera toujours un concept *abstrait* (ici, la consistance) qui échappera au pouvoir démonstratif d'un système formel consistant. Ce concept est *abstrait* précisément dans le sens où il est *au-delà* de ce que peuvent démontrer les signes concrets et finis. L'apparent échec[22] de l'approche formaliste laisse donc le champ libre au réalisme qui permet la compréhension de la signification abstraite des vérités mathématiques. Cela ouvre aussi la voie à une forme de rigueur informelle qui permettrait justement de concevoir des preuves tout à fait convaincantes mais qui ne satisferaient pas les critères du formalisme. Ainsi, lorsque nous démontrons le second théorème de Gödel, c'est notre compréhension de la signification abstraite des symboles employés par le système formel qui nous laisse croire qu'il existe bel et bien une preuve de la consistance du système formel au-delà de ce système formel.

3.4 Gödel et le réalisme mathématique

Gödel était un réaliste mathématique convaincu. Il considérait que ce réalisme avait été essentiel à la découverte de ses célèbres théorèmes d'incomplétude. Comme il le note lui-même :

---

[22] J'utilise l'adjectif « apparent » car le formalisme n'est pas mort, et bien qu'il ne rende pas parfaitement compte du théorème de Gödel, il est encore relativement satisfaisant pour plusieurs mathématiciens. Selon eux, le formalisme est suffisant pour *faire* des mathématiques, à défaut de pouvoir justifier adéquatement tous les concepts mathématiques.

> How indeed could one think of expressing metamathematics in the mathematical systems themselves, if the latter are considered to consist of meaningless symbols which acquire some substitute of meaning through metamathematics.[23]

Les systèmes mathématiques obtiennent leur interprétation grâce à des considérations métamathématiques, et celles-ci se réduisent à une expression dans un système qui n'a aucune signification. Le serpent se mord la queue. Donc, les considérations métamathématiques sont des coquilles vides et ne donnent qu'une illusion d'interprétation. La signification des entités mathématiques doit provenir d'ailleurs si nous sommes capables de concevoir une telle chose que le théorème de Gödel, car celui-ci, comme nous l'avons vu, procède d'une arithmétisation de la syntaxe nécessaire à la construction de propositions indécidables sous forme arithmétique.

Pour le réaliste, il n'y a rien d'étonnant dans les implications du théorème de Gödel. En fait, il s'agit d'un résultat qu'il espère secrètement, car à ses yeux il sonne le glas du projet formaliste, comme nous venons de le voir. Pour Gödel, la découverte de son premier théorème repose essentiellement sur un principe heuristique, celui du concept transfini de la « vérité objective mathématique »[24], duquel découlent les résultats prouvables de façon finitaire, et non l'inverse. Gödel admet que la découverte de son théorème n'est pas totalement impossible par les moyens finitistes, mais ceux-ci auraient rendu sa découverte beaucoup plus ardue, car il apparaît dans cette perspective comme un défaut de construction qu'on aurait souhaité ne jamais voir apparaître.

3.5 Conceptions de l'esprit

Nous concentrerons nos efforts à déterminer maintenant quelle conception de l'esprit doit être associée au réalisme et au formalisme. En fait, il s'agit de cerner ce qui rend l'esprit capable de déterminer un grand nombre de vérités mathématiques. Chacune des deux philosophies n'accorde pas la même importance à la compréhension de la signification des vérités mathématiques et ne conçoit pas de la même façon leur justification, ce qui impose des conditions différentes sur le fonctionnement de l'esprit.

Tout d'abord, dans la perspective formaliste, l'esprit pourrait « se contenter » d'être une machine, ce qui serait suffisant pour bien affirmer les vérités mathématiques,

---

[23] Wang 1974, p. 9
[24] Wang 1974 p. 9

Tout d'abord, dans la perspective formaliste, l'esprit pourrait « se contenter » d'être une machine, ce qui serait suffisant pour bien affirmer les vérités mathématiques, et par conséquent il serait soumis aux mêmes limitations que celles des machines. Cela rappelle évidemment l'interprétation forte de l'IA.[25] L'algorithme de notre esprit, qu'on imagine fort sophistiqué et complexe, permet la compréhension[26] et la justification des vérités mathématiques en calculant de façon non consciente[27] le fait que telle proposition mathématique vraie découle d'une application adéquate des règles et des axiomes d'un système formel, soit explicite (celui étudié à ce moment par un mathématicien par exemple), soit implicite (celui qui « gouverne » le fonctionnement de notre cerveau, formé à partir de notre expérience quotidienne et ayant développé les fonctions nécessaires à une compréhension naïve de la géométrie et de l'arithmétique par exemple). En d'autres termes, nos certitudes à l'égard des vérités mathématiques n'est le résultat que d'un calcul non conscient, mais bien concret et fini.

Il faut préciser à ce moment que lorsque Turing développa son concept de machine, il avait l'ambition de décrire le processus qui était à l'œuvre chez l'humain lorsque celui-ci calculait, ou plus généralement, faisait des mathématiques.

> What Turing did was to analyze the human calculating act and arrive at a number of simple operations which are obviously mechanical in nature and yet can be shown to be capable of being combined to perform arbitrarily complex mechanical operations.[28]

Cette analyse qui s'avéra somme toute assez concluante rend par le fait même plus plausible la thèse de l'IA forte. Si la description aussi bien externe qu'interne du comportement d'un être humain lorsqu'il fait des mathématiques peut se faire en termes purement mécaniques, c'est-à-dire de machine de Turing, pourquoi chercher plus loin ce qui fonde notre compréhension des mathématiques? Pour parvenir à établir cette description mécanique, Turing a dû s'appuyer sur deux principes : la détermination et la finitude. Imaginons un être humain qui fait une addition sur une feuille de papier. Le premier stipule que la détermination du prochain acte à poser (dans une suite visant le

---

[25] La version faible ne sera pas envisagée dans ce qui suit car elle n'engage à aucune conception particulière de l'esprit.

[26] Sans doute que dans la perspective formaliste, notre usage du langage naturel (incluant sa compréhension) se réduit à l'usage d'un algorithme.

[27] Non consciente au sens où l'activité de nos neurones, qui fourniraient le support matériel de ce calcul, n'est pas perceptible par une simple introspection.

[28] Wang 1974, p. 91

Chaque acte est donc déterminé au moment *t* de façon univoque par le calculateur qui tient compte à la fois d'un seul symbole lu et son état d'esprit au moment *t*.

Si le concept de symbole peut être aisément précisé du fait qu'il est discret, il en va autrement de « l'état d'esprit ». On imagine bien que l'état d'esprit d'un calculateur humain peut être parasité par toutes sortes de considérations impertinentes à son calcul. Mais cela n'empêche pas pour autant le calculateur de décider de son prochain acte (s'il doit lire ou écrire quelque chose sur sa feuille de papier). Il doit donc être possible et même nécessaire de rendre cet état d'esprit sous forme d'une attitude conditionnelle formelle, atteint par la succession des actes précédents et qui indique ce qui doit être accompli lors de la prochaine étape, dépendamment du prochain symbole qui sera lu. Cette caractéristique rend le processus de calcul humain suffisamment similaire à ce qui peut être accompli par un algorithme dont le rôle même est de fournir la marche à suivre pour résoudre un problème. La succession des états d'esprit du calculateur est représentée par une table d'états interne de la machine qui essentiellement encode un algorithme.

Le principe de la finitude, comme son nom l'indique, impose des limitations aux capacités de notre cerveau dans son exécution du calcul. Notre cerveau étant physiquement d'une grosseur limitée, contenant un nombre fini de neurones servant de support matériel à l'exécution de l'algorithme, on en déduit qu'il doit exister une limite supérieure à sa capacité de percevoir et de stocker des informations. Il en va de même pour son nombre d'états d'esprit qui, s'il était infini, obligerait une fermeture logique arbitraire des états d'esprit et obligerait le cerveau à posséder un pouvoir infini de discernement, ce qui est contraire aux principes actuels de la physique.[29] En termes de machine de Turing, cela signifie que seul un nombre fini de symboles peuvent être lus ou écrits pour chaque étape et que la table des états internes est finie elle aussi.

Ces deux principes, s'ils sont exacts, donnent une solide base pour la conception formaliste/mécaniste de l'esprit. Cependant, il n'est pas clair que même si ces principes s'appliquent aux cerveaux, ils s'appliquent d'une manière similaire et imposeraient ainsi les mêmes limitations qu'aux machines.[30] Plus important encore, une telle conception de

---

[29] Wang 1974, p. 93
[30] Tieszen 1994, p. 191

les mêmes limitations qu'aux machines.[30] Plus important encore, une telle conception de l'esprit est loin de résoudre la question de l'incomplétude issue du théorème de Gödel. Si l'esprit est une machine, et s'il est consistant, alors en principe une vérité mathématique lui échappe et il ne peut démontrer toutes les vérités. En somme, l'approche formaliste laisse présager l'existence d'une limitation imposée par le théorème de Gödel aux machines pensantes, incluant l'esprit.

Qu'en est-il de la conception réaliste de l'esprit? Celle-ci contraste avec la conception formaliste principalement à cause de son rôle d'intermédiaire entre le monde physique et le monde des entités mathématiques abstraites. Dans cette optique, les états mentaux sont fondamentalement intentionnels, c'est-à-dire qu'ils font appel à une signification qui dépasse la simple manipulation de symboles discrets. Les formules mathématiques écrites sur papier n'ont aucune valeur si elles ne sont pas interprétées par nos états mentaux. Ceux-ci permettent d'extraire les vérités mathématiques d'un monde qui transcende la réalité physique et mécanique. Bien qu'ils en sachent très peu sur la véritable ampleur de ce monde abstrait, les réalistes considèrent souvent que l'esprit a en principe un accès privilégié à la totalité de ce qu'il contient. Gödel n'hésite pas à suggérer l'existence d'un organe physique qui nous permettrait d'accéder au monde des Idées mathématiques :

> I conjecture that some physical organ is necessary to make the handling of abstract impressions (as opposed to sense impressions) possible, because we have some weakness in the handling of abstract impressions which is remedied by viewing them in comparison with or on occasion of sense impressions. Such sensory organ must be closely related to the neural center for language.[31]

Mais cette suggestion ne nous concerne pas outre mesure. Un tel organe ne pourrait-il être reproduit par une machine? Aussi intéressante qu'elle puisse paraître, nous ne tenterons pas de répondre à cette question ici. Ce que l'on doit surtout retenir, c'est que le réaliste est porté à croire qu'aucune vérité mathématique ne peut en principe échapper complètement à l'esprit humain. En particulier, contrairement aux machines pensantes, il n'est pas soumis aux limitations du théorème de Gödel. Son fonctionnement n'est pas celui d'un algorithme car ce dernier n'est pas par nature intentionnel.

---

[30] Tieszen 1994, p. 191
[31] Wang 1996, p. 233

La conception réaliste de l'esprit n'entre pas nécessairement en conflit avec le principe de finitude défini plus tôt. Le réaliste ne voit aucun inconvénient à considérer que le nombre d'états mentaux du cerveau est fini à chaque instant. Ce serait plutôt le principe de détermination qui poserait problème. Pour un être humain, les mathématiques ne se résument pas à une simple manipulation de symboles selon une marche à suivre déterminée à l'avance et dépourvue d'intentionnalité.

Il semble bien qu'un aspect relié à l'intentionnalité soit ce qui distingue les deux conceptions de l'esprit. Mais cela pourrait peut-être se révéler plutôt flou à la lumière de certaines considérations de l'IA forte. Rappelons que celle-ci considère qu'une machine pourrait éventuellement reproduire parfaitement le fonctionnement de l'esprit, au sens où elle possèdera, au même titre que nous, les mêmes facultés « mentales » que l'on accorde à l'esprit, que ce soit la conscience de soi ou la faculté de juger le vrai du faux, le bien du mal. En somme, les facultés internes des machines pensantes, tout comme l'esprit, seraient pourvues d'intentionnalité. La question serait maintenant de savoir ce qui empêche une machine d'avoir accès elle aussi au monde platonicien. Dans cette perspective, nous nous retrouvons pratiquement à cheval entre les deux conceptions de l'esprit, comme si la conception formaliste pouvait se permettre d'être fondamentalement intentionnelle, ce qui semble *prima facie* contradictoire. Une façon expéditive de résoudre cette contradiction serait d'associer l'approche formaliste à l'IA faible, ce qui éliminerait pratiquement la conception formaliste de l'esprit et laisserait intact la conception réaliste. On pourrait aussi éliminer la conception réaliste, et par le fait même l'aspect intentionnel des vérités mathématiques.

3.6 La thèse de Church

Alonzo Church a mis de l'avant dans les années 1930 une proposition qui n'est pas en soi susceptible d'être prouvée mathématiquement, mais qui a une profonde influence sur notre manière de concevoir les preuves mathématiques. Qu'est-ce qui fait d'une preuve mathématique un raisonnement valide? Existe-t-il des méthodes qui peuvent nous garantir que nos raisonnements mathématiques sont valides? À cet effet, la notion de méthode *effective* suggère que l'on peut résoudre une classe de problèmes mathématiques suivant une procédure systématique s'appuyant sur des règles

élémentaires et fixes. La thèse de Church, ainsi nommée par S. C. Kleene, propose que l'on identifie la notion intuitive et vague de méthode (ou fonction) *effective* à la notion de récursivité qui, elle, est bien définie en mathématiques. En identifiant cette notion plutôt imprécise à la notion de fonction récursive (ou, de façon équivalente, la notion de machine de Turing), la « procédure systématique » devient mécanique (calculable au sens de Turing), donc tout à fait précise et satisfaisante. À cet égard, c'est bel et bien l'analyse qu'a fait Turing du comportement d'un calculateur humain qui a formellement convaincu aussi bien Gödel que Church lui-même du bien-fondé de la thèse de Church. La thèse de Church apparaît d'autant plus convaincante qu'il semble impossible de concevoir ce que pourrait être une méthode effective finie qui ne serait pas récursive. Ce qu'elle propose paraît si évident et même trivial pour certains spécialistes qu'il ne semble nécessaire que de la mentionner au passage sans envisager que l'on puisse la contester. Pour ceux-ci, la thèse de Church n'affirme rien de plus que les problèmes mathématiques doivent être résolus par des méthodes mathématiques.

Cependant, la thèse de Church ne fait pas l'unanimité pour autant, elle a été contestée à plusieurs reprises. En ce qui nous concerne, la thèse de Church peut être vue comme ayant une profonde influence sur les conceptions de l'esprit déjà étudiées. En effet, le recours à la thèse de Church a été implicite en ce qui concerne la conception formaliste de l'esprit. Grossièrement, si on conçoit le processus de la pensée comme étant essentiellement une succession de raisonnements informels mais tout de même effectifs qui nous permettent de résoudre de manière efficace une quantité impressionnante de problèmes de la vie quotidienne et si la thèse de Church est vraie, alors le fonctionnement de l'esprit qui sous-tend ces raisonnements informels est **mécanique**, l'esprit est donc une machine. Autrement dit, si toute fonction cognitive est **effective**, la thèse de Church nous oblige à conclure que toute fonction cognitive est **mécanique**.

Comme il a été établi plus tôt, le réalisme mathématique ne peut endosser une telle conclusion à propos de l'esprit. Et cela concerne notre usage de significations abstraites d'objets mathématiques indépendants de notre esprit. En effet, plusieurs résultats mathématiques se fondent sur des méthodes ou des concepts non constructifs, tels les nombres ordinaux transfinis. Les mathématiciens travaillent à partir de la

signification de ces nombres sans qu'il existe de méthode calculable pouvant les démontrer. Le réaliste ne nie pas nécessairement que le fonctionnement de l'esprit est gouverné par des règles, seulement il doit exister une classe de « fonctions » mentales effectives qui ne sont pas récursives. La conception réaliste se voit obligée de rejeter la thèse de Church.

En terminant, rappelons que nous avions affaire dans ce chapitre à une tentative indirecte de « réfutation » du Mécanisme. La vraie nature des vérités mathématiques est disputée entre deux philosophies des mathématiques, à savoir le formalisme et le réalisme. Pourtant, seule la seconde semble rendre compte adéquatement de la vérité des propositions gödeliennes associée chacune à un système formel, mais ne pouvant être démontrée à l'intérieur de ce système. Comme le Mécanisme semble plutôt lié au formalisme, nous en déduisons que le Mécanisme ne permet pas de rendre compte de toutes les vérités mathématiques. L'esprit humain serait doué d'une intuition qui lui donne accès aux entités mathématiques du monde platonicien et lui permet de former des propositions mathématiques vraies à propos de ces entités, alors que l'on répugne à accorder aux machines une telle faculté, car on les croit dépourvues d'intentionnalité. Cette critique du Mécanisme a une répercussion sur la thèse de Church, car la perception de la vérité des propositions gödeliennes seraient le fait de méthodes effectives qui ne seraient pas des mécanismes récursifs.

CHAPITRE 4.
L'ARGUMENT DE LUCAS

4.1 Réfuter le Mécanisme

Réfuter ou limiter la thèse mécaniste sur la base d'un théorème de logique mathématique est assez audacieux. Les différentes réfutations ou limitations s'appuient principalement sur ce que nous, humains, pouvons connaître versus ce que peuvent connaître les machines algorithmiques. Existe-t-il des limites *formelles* à notre connaissance humaine? Comme nous le constaterons, le problème ici est que, trop souvent, des limites *matérielles*, à savoir nos ressources intellectuelles restreintes et peu fiables et la finitude de notre existence humaine, viennent embrouiller la portée des limites formelles. Pour compliquer un peu le problème, nous avons vu que le Mécanisme se décline en deux versions, la réfutation doit donc être double.

L'article de J. R. Lucas, « Minds, Machines and Gödel », publié en 1961, doit être considéré comme le premier véritable effort pour présenter une argumentation systématique en faveur d'une limitation imposée par le théorème de Gödel à la thèse du Mécanisme, et en cela il est devenu un véritable classique. En fait, plus qu'une simple limitation, Lucas soutient que le théorème de Gödel fournit un argument permettant de prouver que le Mécanisme est carrément faux, à savoir « minds cannot be explained as machines »[32].

Déjà en 1959, Ernest Nagel et James R. Newman concluaient leur livre de vulgarisation sur le théorème de Gödel[33] en faisant part de leur conviction que les machines à calculer ne remplaceraient pas l'intelligence humaine, et ce sur la base du résultat de Gödel qui montrait bien qu'aucune méthode axiomatique (système formel) reproduisant la théorie des nombres naturels n'est capable de résoudre tous les problèmes (mathématiques). Pour chaque méthode axiomatique que l'on pourrait incorporer à une machine, il existerait un nombre infini de problèmes mathématiques qu'elle ne pourrait résoudre. Bien qu'ils supposaient que l'intelligence humaine est

[32] Lucas 1964, p. 43
[33] Nagel 1989, pp. 92-95

limitée d'une façon ou d'une autre, elle semble posséder une structure de règles d'opérations bien plus puissante que celle des machines.[34]

Pour sa part, comment Lucas entend-il réfuter le Mécanisme? Si l'on se fie à Lucas, il suffit de montrer que l'esprit humain est fondamentalement différent de n'importe quelle machine qui aura la prétention de représenter parfaitement l'esprit : « it is enough [...] to show that the machine is *not the same* as a mind »[35]. Par « fondamentalement différent », il faut entendre que l'esprit humain comprend au moins une vérité que ne peut prouver[36] la machine pensante. Il ne s'agit pas ici pour l'esprit d'être quantitativement supérieur à la machine, c'est-à-dire de posséder de façon absolue un plus grand nombre de vérités que la machine, mais simplement d'en posséder au moins *une* que la machine ne peut affirmer.

Est-ce vraiment suffisant? N'oublions pas ici qu'il existe deux versions du Mécanisme, chacune ayant des considérations différentes sur la nature de l'esprit. Est-ce que l'argument de Lucas réfute les deux versions? Sinon, à laquelle des versions ses critiques s'adressent-elles? À quelles objections fera face Lucas? C'est ce que nous verrons dans ce chapitre.

Avant d'aller plus loin, soulignons que l'argument de Penrose, que nous examinerons plus tard, est à bien des égards similaire à celui de Lucas. C'est pourquoi dans ce qui suit les propos de Lucas seront teintés de ceux de Penrose[37]. Ce dernier traitant certaines questions de façon plus complète ou plus rigoureuse, plutôt que de revenir plus tard sur les mêmes questions soulevées ici, j'ai choisi d'intégrer certaines considérations de Penrose dans ce chapitre-ci.

---

[34] Prudemment, Nagel et Newman se contentaient de parler des machines conçues à leur époque, mais si on considère qu'autant le fonctionnement des machines de leur époque que celui des machines d'aujourd'hui est basé sur la théorie des machines de Turing, leur remarque demeure encore actuelle.
[35] Lucas 1964, p. 49
[36] On conviendra que la machine affirme une proposition, autrement dit la considère comme vraie, si et seulement si elle est le résultat d'un calcul algorithmique. En d'autres termes, la machine *n'affirme* que ce qu'elle *prouve* ou *démontre*, et vice versa. Bien sûr, on n'exclut pas la possibilité qu'il puisse y avoir une erreur dans la preuve, mais toute affirmation de la machine sera accompagnée d'une preuve, d'un calcul. Du point de vue de la machine, ce qui est vrai (ce qui est affirmé) est équivalent à ce qui est démontrable.
[37] Penrose 1994

4.2 Vraie mais non démontrable

Selon Lucas, il existe pour chaque machine pensante une vérité qu'elle ne peut affirmer. Quelle est donc cette vérité qui échappe aux machines? Considérant que toute machine est équivalente à un système formel, on se doute bien que Lucas a en tête la proposition gödelienne. Nous avons déjà vu dans le premier chapitre que celle-ci est indécidable pour le système formel dont elle est issue. Pourtant, sous l'interprétation standard, elle est vraie. De l'avis de Lucas, seul l'esprit humain est à même de voir *toute* proposition gödelienne comme étant *vraie*[38], alors que chaque machine sera incapable de démontrer sa propre proposition gödelienne et, par conséquent, de l'affirmer. C'est ainsi que Lucas montre que l'esprit est fondamentalement différent de toute machine.

Comme si ce n'était pas suffisant, Lucas ajoute un second volet à son argument pour bien montrer que toute machine construite par un humain est faillible[39] et que le Mécanisme est donc faux. Imaginons un tenant de la thèse mécaniste soutenant qu'il a construit une machine pensante, une machine bien définie qui se veut le modèle mécanique de l'esprit. En principe, une telle machine, pourvue d'une capacité de raisonnement équivalente à celle d'un humain, et en particulier à celle de Lucas, devrait pouvoir affirmer les mêmes propositions que Lucas peut démontrer et considérer comme vraies. Or Lucas trouve à partir du système formel de cette machine une proposition indécidable, une proposition que la machine ne peut affirmer sous peine de devenir inconsistante. À ce moment, le tenant du Mécanisme est libre de modifier sa machine pour qu'elle puisse pallier à cette faille, c'est-à-dire accomplir ce dont elle était incapable. Il présente alors à Lucas une nouvelle machine. Lucas, appliquant à nouveau le premier théorème de Gödel à la nouvelle machine (ou plutôt, à son système formel), obtient une nouvelle proposition gödelienne que ne peut démontrer la machine mais qui est pourtant vraie. La machine est renvoyée au laboratoire où le tenant du Mécanisme est libre d'apporter toute modification qu'il juge nécessaire pour rendre sa machine équivalente à l'esprit humain. Ce petit jeu se poursuit jusqu'à ce que ou bien Lucas

---

[38] C'est-à-dire qu'elle respecte bel et bien ce qu'elle affirme d'elle-même, à savoir qu'elle n'est pas démontrable par le système formel et, par extension, par la machine.

[39] Une machine sera considérée *faillible* ou *vulnérable* dans un sens qui reflète son incomplétude. Une machine n'est pas faillible parce qu'elle se trompe ou fait une erreur de calcul, c'est plutôt parce qu'elle ne reproduit pas adéquatement le comportement de l'esprit humain en étant incapable d'affirmer la vérité de sa proposition gödelienne que peut saisir et affirmer l'esprit humain. Cette vérité constitue une *faille* ou une *vulnérabilité* pour toute machine pensante.

admette que la machine n'a plus de failles, ou bien que le Mécaniste reconnaisse que ses efforts sont vains. Dans le premier cas, le Mécaniste a prouvé sa thèse, dans le second, il est réfuté. Selon Lucas, en principe, on pourra toujours trouver une faille dans le fonctionnement de n'importe quelle machine que pourrait construire le Mécaniste grâce au théorème de Gödel, et le Mécaniste ne peut que perdre son pari. Puisque l'esprit humain est fondamentalement différent de toute machine qui peut être construite, le Mécanisme est donc réfuté.

Ce second volet a le mérite de montrer de façon plus évidente le caractère « dialectique » (c'est Lucas qui emploie ce terme) de son argument. En fait, ce qu'il veut mettre en évidence, c'est le côté *raisonnement par l'absurde* de son argument. À cet égard, Penrose exploite beaucoup mieux cette façon d'argumenter. Pour l'essentiel, il s'agit de prendre les hypothèses de la thèse à réfuter et de montrer qu'elle mène à une contradiction. Une des hypothèses est donc fausse (peu importe laquelle), mais comme la thèse ne peut se passer de l'une de ces hypothèses, c'est toute la thèse qui est réfutée.[40] Et c'est bien ce que soutient Lucas:

> The argument is a dialectical *reductio ad absurdum*. We take certain premisses the physical determinism[41] supplies, and show that, in virtue of Gödel's theorem, they lead to a contradiction; from which we argue that the thesis of physical determinism itself is the one to be rejected.[42]

Remettons cette citation dans le contexte du jeu entre le tenant du Mécanisme et Lucas lui-même. Le Mécaniste prétend que sa machine est l'égal de l'esprit humain, ce qui implique qu'elle peut, en particulier, affirmer (considérer comme vraies) les mêmes

---

[40] Examinons un exemple classique de raisonnement par l'absurde : l'argument d'Euclide selon lequel il n'existe pas de plus grand nombre premier. Il s'agit d'un raisonnement par l'absurde : on pose une certaine prémisse et on montre qu'elle entraîne une contradiction. On se doit donc de rejeter la prémisse. Dans le cas de l'argument d'Euclide, on suppose qu'il existe un nombre premier p supérieur à tous les autres. On considère maintenant le nombre N qui est le produit de *tous* les nombres premiers jusqu'à p et auquel on ajoute 1. Autrement dit, $N = ((2 \times 3 \times 5 \times \ldots \times p) + 1)$. N est donc plus grand que p, mais il n'est pas possible de le diviser sans reste par aucune combinaison de nombres premiers. Comme tout nombre est soit premier soit il ne l'est pas, N est alors soit un nombre premier plus grand que p, ou alors il ne l'est pas, auquel cas il doit exister un nombre premier plus grand que p qui permet de diviser N. Mais cette conclusion contredit directement notre hypothèse de départ. Donc, il n'existe pas de nombre premier plus grand que tous les autres. (Et cette conclusion est valide même pour les nombres excessivement grands, si grands qu'on pourrait imaginer qu'il n'y ait pas assez de temps dans l'univers pour simplement les écrire. Ainsi la conclusion est valide peu importe qu'il soit *en pratique* possible ou non de déterminer N. Cette dernière remarque prendra toute son importance quand nous examinerons les objections à l'argument de Lucas.)

[41] En ce qui nous concerne, le déterminisme physique dont parle Lucas réfère au Mécanisme.

[42] Lucas 1970, p.149

propositions que Lucas juge vraies. Considérant la spécification même du fonctionnement de la machine, Lucas répond qu'il peut trouver une faille dans la machine, c'est-à-dire qu'il existe en principe, en vertu du théorème de Gödel, une proposition indécidable qu'elle ne peut affirmer mais que Lucas, lui, peut saisir comme étant vraie. À toute machine que pourrait présenter le Mécaniste en prétendant qu'elle est sans failles, peu importe sa complexité, Lucas peut lui découvrir *en principe* une faille.

Cette dernière phrase nous introduit à une distinction qui sera cruciale pour la suite des discussions entre ce qu'on peut faire *en principe* et ce qu'on peut faire *en pratique*. Le raisonnement par l'absurde sert d'assise à un argument de principe, et cela implique que les contraintes matérielles qui pourraient nous empêcher de remplir l'une des conditions de l'argument sont totalement négligées. Par exemple, le fait qu'il n'y ait pas assez de temps dans l'univers pour écrire une suite de $2^{2^{65536}}$ « 1 » ne peut valoir comme objection au fait qu'*en principe* l'on peut énumérer une suite de « 1 » plus longue que la précédente. Cela met au défi toute imagination, et on peut facilement perdre le jugement de ce qui compte comme un argument valide et de ce qui est vide de sens. Les mathématiciens sont habitués à réfléchir à propos de raisonnements par l'absurde, mais il faut ajouter qu'ils travaillent sur des entités abstraites, alors qu'en ce qui concerne notre sujet, nous avons affaire à une situation du monde « réel ». Archimède prétendait être capable de soulever la Terre à l'aide d'un levier assez long et d'un point d'appui. En principe, il avait raison, alors qu'en pratique, son « exploit » n'est pas réalisable. Les prétentions de ce genre ont-elles une quelconque valeur ou conduisent-elles infailliblement à une sorte d'aporie entre les arguments de principe et les objections de pratique? Continuons l'analyse pour en décider.

Que réfute exactement l'argument? Réfute-t-il les deux versions de l'IA? Considérons tout d'abord l'IA forte. Pour celle-ci, l'esprit est une machine, il est donc soumis à des limitations équivalentes à celles d'une machine pensante. En d'autres termes, l'esprit n'est pas en mesure de démontrer sa propre proposition gödelienne. L'esprit est donc lui-même incomplet, dans la mesure où lui échappe la vérité de sa proposition indécidable. En somme, l'esprit n'est guère mieux qu'une machine pensante. Et ce n'est certainement pas le raisonnement par l'absurde de Lucas, qui doit prendre

pour hypothèse que l'esprit est une machine, qui y changera quelque chose, car s'il y a bien une faille chez les machines pensantes, la même faille se retrouve dans l'esprit humain. Nous sommes forcés de conclure que l'IA forte n'est pas réfutée par l'argument de Lucas.

Avant de passer à l'IA faible, il faudrait signaler qu'on trouve au passage dans les propos de Lucas ce qui pourrait bien être une tentative de réfutation de l'IA forte. [43] En effet, selon lui, soutenir que l'esprit est une machine devient totalement vide de sens s'il nous est impossible de spécifier un quelconque algorithme de l'esprit. Affirmer que l'homme est une machine sans pouvoir spécifier quoi que ce soit revient à tenter de définir une nouvelle signification au mot « machine » en incluant l'esprit dans son extension sans pour autant que la définition de l'esprit ait quelque chose en commun avec celle de la machine. Cette tentative n'a rien de substantiel selon Lucas. Ces propos montrent à quel point Lucas ne réalise pas que l'utilisation d'un raisonnement par l'absurde peut être une arme à double tranchant. C'est justement tout l'intérêt d'un tel argument que de ne pas avoir à tout spécifier, y compris l'algorithme de l'esprit, pour en déduire une contradiction. Si la version forte inclut parmi ses hypothèses que l'esprit est une machine, il faut tenir compte de ses conséquences tout au long du raisonnement, et non pas seulement discréditer l'hypothèse en dénonçant le fait qu'elle n'est pas précisément déterminée.

Si l'argument de Lucas ne parvient pas à réfuter l'IA forte, y parvient-il pour ce qui est de l'interprétation faible? L'IA faible propose une machine pensante capable de simuler parfaitement le comportement ou les manifestations « publiques » de l'esprit, sans que l'on ait à se prononcer sur la nature du fonctionnement interne de l'esprit. Si l'esprit juge que telle proposition mathématique est vraie, alors la machine se doit d'être en mesure d'affirmer ce théorème. On ne se demande pas si la machine « juge » de la vérité de cette proposition au même sens que les humains le font. Ces conditions sont favorables à l'argument de Lucas. L'esprit n'étant pas une machine, il n'est pas soumis au théorème de Gödel, contrairement aux machines pensantes. En principe donc, il existe pour chaque machine une proposition indécidable que l'esprit peut découvrir et saisir comme étant vraie, mais il n'existe pas une telle proposition pour l'esprit. L'IA

[43] Lucas 1970, p.152

faible échoue à fournir une machine pensante équivalente à l'esprit en termes de comportement externe.

4.3 Les objections

L'argument de Lucas a suscité une impressionnante vague de protestations et d'objections. Ce qui suit se veut une liste des catégories d'objections auxquelles a dû faire face Lucas. Cette liste ne comprend pas toutes les objections possibles, certaines ne m'ont pas paru dignes de mention. Je crois bien rassembler ici les objections les plus sérieuses, ce qui nous permettra de mieux juger si Lucas peut véritablement se déclarer vainqueur du débat. Par ailleurs, certaines objections formulées contre Penrose auraient pu être adaptées pour l'argument de Lucas, mais celui de Penrose étant plus rigoureux, il sera plus intéressant de voir comment Penrose répond, ce qui nous donne une idée de ce qu'aurait pu répondre Lucas.

Avant d'aller plus loin, je souligne à nouveau la cruciale importance de la distinction entre ce qui peut se faire *en principe* et ce qui peut se faire *en pratique*. Une grande partie du débat tournera autour de cette distinction. Il n'est pas dit non plus qu'une catégorie n'empiète pas sur une autre.

Voici les catégories que nous devrons analyser dans le contexte de l'IA faible :

Catégorie *i*) la consistance : Les conclusions des premier et second théorèmes de Gödel sont des conditionnels, dont la condition suffisante est que le système formel doit être consistant. Les objections concernent donc les possibilités que l'esprit soit consistant[44] et que nous puissions le savoir, que la machine pensante soit un système formel consistant et que nous ayons la capacité de prouver la consistance de ce système.

Catégorie *ii*) la complexité : Et si le système formel était trop complexe pour que l'on puisse découvrir sa proposition gödelienne? La trop grande complexité

[44] Il peut paraître étrange de parler de la consistance de l'esprit. Dans le contexte qui nous préoccupe actuellement, la consistance de l'esprit réfèrera à l'ensemble des propositions mathématiques considérées vraies par l'esprit, peu importe la manière (intuition ou démonstration formelle) dont elles sont obtenues. Cela inclut donc toutes les propositions gödeliennes possibles.

pourrait aussi interférer avec notre capacité à juger de la consistance du système, ce qui nous renvoie à la catégorie *i*.

Catégorie *iii*) l'« informalité » : L'argument de Lucas fait usage d'une double rigueur : celle formelle du théorème de Gödel, et une autre informelle lorsqu'il s'agit de voir la vérité de la proposition gödelienne pour un esprit qui n'est pas une machine. Certains soutiennent que l'argumentation informelle n'est pas recevable.

Catégorie *iv*) la finitude : l'humain est un être qui a une durée de vie limitée, et donc des ressources intellectuelles limitées dans le temps. L'ensemble des vérités qu'il peut affirmer est donc fini et une machine pourrait facilement les mémoriser, y compris les propositions gödeliennes.

Commençons l'analyse immédiatement par cette dernière catégorie. D'entrée de jeu, il faut souligner qu'en se situant implicitement dans un contexte de test de Turing, cette objection ne tient pas vraiment compte de l'enjeu que soulève le théorème de Gödel. Ce qui compte pour le test, c'est de tromper un évaluateur sur la nature de la machine pensante, alors que la question qui nous intéresse est plutôt de savoir si un algorithme pourrait arriver à simuler le comportement de l'esprit humain. Ce que l'on propose ici n'est qu'un stockage de propositions vraies et de leur démonstration, plutôt qu'une authentique capacité à démontrer et à mener des raisonnements à partir d'un ensemble restreint de règles.

Cette catégorie peut être abordée sous un autre angle. Elle questionnerait en fait notre capacité à mener des raisonnements sur ce qui excède notre finitude, en particulier les raisonnements qui demanderaient plus de temps que l'existence entière de l'humanité. Le simple fait de spécifier l'algorithme d'une machine pensante serait selon toute probabilité excessivement long, il en irait de même pour la procédure permettant de trouver sa proposition gödelienne? Nous avons déjà en partie répondu à cette question en signalant la distinction principe/pratique. Pour l'instant, nous n'irons pas plus loin,

car cette question s'apparente maintenant à la catégorie de la complexité qui sera traitée plus tard.

4.4 La consistance de l'esprit

Attaquons maintenant la première catégorie. La question de la consistance est tout à fait cruciale. Quelles raisons nous permettent de croire que le système formel de la machine pensante est consistant? À la lumière des propos de Lucas, on réalise que le système formel « hérite » de cette propriété du fait que l'esprit l'est lui-même. Autrement dit, si l'esprit humain est consistant, alors toute machine pensante doit l'être elle aussi. La machine est consistante pour les mêmes raisons (quand elles s'appliquent) que l'esprit l'est. La question est donc maintenant de savoir quelles raisons avons-nous de croire que l'esprit est consistant.

Lucas est convaincu que l'ensemble des propositions mathématiques que tout mathématicien considère vraies est en principe consistant et que, par conséquent, le raisonnement humain est consistant. Tout d'abord, le fait est que nous humains avons pratiquement horreur des contradictions, nous ne tolérons pas facilement qu'un raisonnement puisse mener à admettre une chose et son contraire. Nous savons distinguer le vrai du faux, mais plus encore nous préférons le vrai au faux. Il s'agit qu'un discours nous apparaisse le moindrement contradictoire pour qu'il soit discrédité et qu'il perde tout son pouvoir de persuasion. Par ailleurs, Lucas s'appuie implicitement sur un théorème du calcul des prédicats [$\neg A \supset (A \supset B)$] pour montrer que le raisonnement humain n'a rien à voir avec les conséquences que l'on peut tirer d'une paire de propositions contradictoires, à savoir que l'on peut affirmer n'importe quoi, que toute proposition devient acceptable :

> If we really were inconsistent machines, we should remain content with our inconsistencies, and would happily affirm both halves of a contradiction. Moreover, we would be prepared to say absolutely anything – which we are not.[45]

Ce recours peut paraître un peu déplacé dans la mesure où la conclusion de Lucas ne s'applique que dans un contexte d'IA forte. Mais si l'esprit n'est pas une machine

[45] Lucas 1964, p.53

pensante, pourquoi devrait-il se comporter comme celle-ci devant une contradiction? On en conclut que l'argument de Lucas ne concerne pas l'IA faible.

Néanmoins, il est vrai que parfois les humains endossent des ensembles de propositions qui se révèlent inconsistants, et cela se produit en particulier en mathématiques. Par exemple, on n'a qu'à penser à Frege et à ses fondements de l'arithmétique. Cependant, la façon dont il a réagi lorsque Russell lui apprit l'existence de ce qui est devenu son célèbre paradoxe illustre comment l'humain réagit devant l'inconsistance : il faut y pallier. Il est indéniable que nous sommes faillibles, c'est-à-dire que nous ne prenons pas conscience immédiatement de nos erreurs et des contradictions qui se cachent dans nos raisonnements. Il peut s'écouler un certain temps avant la découverte d'une contradiction dans un système formel. Mais cela n'empêche pas le fait qu'un système inconsistant ne nous satisfait aucunement, et que nous souhaitons toujours éliminer les contradictions une fois qu'elles sont découvertes.

Dans un assez long passage, Lucas traite de la possibilité qu'un système formel inconsistant puisse tout de même se retrouver au sein de notre esprit, selon une suggestion de Putnam.[46] Remarquons que nous changeons ici de perspective, nous passons de l'IA faible à l'IA forte. Voyons tout de même ce qu'a à dire Lucas. Deux situations sont envisageables : le système formel contient certaines règles qui « bloquent » la démonstration d'une proposition contredisant une autre proposition déjà démontrée; la contradiction est si subtile, si recherchée qu'elle permet la démonstration de la proposition indécidable associée au système et sans interférer avec la « consistance » de toutes les autres démonstrations. La première situation laisse présager un ensemble de règles fort diverses pour stopper les preuves avant qu'elles ne conduisent à une contradiction. Une règle pourrait forcer l'élimination de la preuve la plus longue. Elle pourrait aussi restreindre l'utilisation des axiomes ou des règles d'inférence dans certains cas particuliers. Ces suggestions peuvent paraître astucieuses, mais Lucas les considère fort peu satisfaisantes : « Even the less arbitrary suggestions are too arbitrary. »[47] En effet, comment pourrait-on justifier de telles règles d'arrêt? Comment justifier que le résultat de la preuve la plus courte soit vrai et que celui de la

[46] Lucas 1964, pp. 52-56
[47] Lucas 1964 p. 54

longue ne le soit pas alors que la longueur d'une preuve est dépendante des axiomes de départ? Un autre ensemble d'axiomes pourrait bien rendre plus longue la preuve auparavant plus courte. Et comment *justifier* la restriction d'un axiome ou d'une règle d'inférence plutôt qu'un autre? Selon Lucas, pour ce qui est de la recherche concernant les systèmes formels, s'il est démontré qu'un axiome conduit à une contradiction, alors cet axiome est complètement éliminé et non seulement restreint. Autrement, aucun consensus n'est possible, chacun pouvant développer ses propres restrictions du système inconsistant qui n'ont rien à voir avec celles des autres. Comment s'assurer alors d'un consensus sur le jugement de ce qui est vrai et de ce qui est faux? De l'avis de Lucas, une règle d'arrêt serait ainsi loin d'être satisfaisante.

La seconde situation propose qu'une contradiction existerait bel et bien mais qu'elle serait si « recherchée » qu'elle laisserait croire que le système est véritablement consistant tout en permettant de démontrer la proposition indécidable. Lucas répond que même si la théorie des nombres naturels fait partie de tout système formel équivalent à une machine pensante, nous n'aurons jamais une pleine certitude que la théorie n'est pas inconsistante. Lucas soutient que tant que nous n'avons pas découvert cette contradiction si spéciale, il nous est permis de croire que le système est consistant. Si une contradiction était un jour découverte, elle ne serait pas tolérée longtemps, les mathématiciens tenteraient aussitôt de redéfinir un nouveau système consistant qui rejoint nos concepts intuitifs sur les nombres naturels. Dans l'éventualité où il s'avérait qu'aucun système contenant l'arithmétique simple puisse être exempt de contradictions, Lucas devient incroyablement pessimiste et affirme que nous devrions non seulement abandonner la totalité des mathématiques mais aussi la totalité de la pensée. Cette dernière remarque, que Lucas ne prend nullement la peine d'expliquer, pourrait en un sens apporter de l'eau au moulin du Mécanisme en affirmant que l'étude de l'esprit dépend fondamentalement de la consistance d'un système formel qui mécaniserait notre compréhension intuitive de la théorie des nombres naturels. Mais il faut se rappeler que nous sommes dans l'interprétation forte de l'IA, et que cette conclusion est en un sens légitime, car toute machine pensante étant inconsistante, tous les raisonnements seraient possibles, toute vérité et sa négation seraient bonnes à dire.

Revenons à la situation de l'IA faible. Ces considérations de Lucas ont certainement le mérite de rendre plausible l'hypothèse que l'esprit est consistant. Mais tout ce que propose Lucas, c'est d'assumer que l'esprit est consistant. Cela ne revient pas au même que de savoir si un système formel est consistant. Nous devons en faire la preuve. Mais on peut aisément imaginer que le système formel d'une machine pensante sera extrêmement complexe. Saurons-nous faire la preuve de sa consistance? C'est dans cette optique que s'inscrivent les critiques de Putnam.

En 1960, Putnam[48] répondait à ce qui n'était alors que les vagues suggestions de Nagel et Newman en s'appuyant sur la consistance du système. Il faisait valoir que ce que le premier théorème de Gödel permet réellement de démontrer n'est pas directement une proposition indécidable et pourtant vraie, mais plutôt un conditionnel. En cela, autant le système formel que le logicien peuvent arriver à cette conclusion, c'est-à-dire que le théorème de Gödel peut très bien être représenté dans le système formel. Ce n'est que $\mathcal{M}(F)$ qui ne peut pas être démontré, et non le conditionnel au complet. Mais le logicien ne peut démontrer $\mathcal{M}(F)$ lui non plus à moins de *démontrer* au préalable que F est consistant, ce qui est peu plausible si F est très complexe.

Lucas ne répond pas directement à la question de la complexité de la preuve de la consistance. En fait, nous avons déjà esquissé la réplique en montrant que le raisonnement par l'absurde ne se préoccupe que des questions de principe. Si le système formel paraît trop complexe, ce n'est sûrement que pour des raisons matérielles. En principe, il n'y a pas de raison pour qu'une équipe de mathématiciens humains ne puisse pas prouver la consistance du système formel, si on lui donnait suffisamment de temps et de ressources. De l'avis de Lucas, le tenant du Mécanisme qui accepte de jouer le jeu de la confrontation entre sa machine et Lucas ne peut se prétendre être en principe agnostique en ce qui concerne la consistance du système.

Que cela ne tienne, Putnam a présenté une version formelle de l'argument de Lucas[49] et indique sa faiblesse :

---

[48] Putnam (1964) p.77
[49] Lucas 1970, pp. 156-157

(a) Si le système formel F n'est pas consistant, alors F n'est pas égal à moi.

(b) Je peux voir que « si F est consistant, alors $\mathcal{H}(F)$ » est vrai.

(c) Si F est consistant, alors je peux voir que « $\mathcal{H}(F)$ » est vrai.

(d) Si F est consistant, alors F ne peut pas voir que « $\mathcal{H}(F)$ » est vrai.

(e) Si F est consistant, alors F n'est pas égal à moi.

Putnam souligne que (b) n'implique pas (c) et que la proposition (c) pourrait très bien être contredite (en raison par exemple de la complexité de F ou de $\mathcal{H}(F)$). Et comme l'usage du prédicat de vérité ne se conforme à aucune règle formelle[50], il semble impossible de *démontrer* (c). Donc le raisonnement de Lucas est invalide.

La réplique de Lucas est assez astucieuse. Ce qui rend la réfutation de (c) plausible selon Lucas tient principalement au fait qu'il s'appuie sur un regard extérieur à la situation entre Lucas et la machine. En effet, dans une telle perspective, la négation de (c) est bien possible :

(¬c) Même si F est consistant, je *ne peux pas voir* que « $\mathcal{H}(F)$ » est vrai.

Il faut comprendre ici que la négation de la possibilité de voir la vérité relèverait, comme dans (d), d'une impossibilité logique. Cependant, du point de vue de Lucas, (¬c) est inadmissible. Il serait contradictoire qu'il soutienne à la fois (b) et (¬c). Au mieux, Lucas pourrait admettre (b) et (¬c)* : Même si F est consistant, alors je *peux ne pas voir* que « $\mathcal{H}(F)$ » est vrai. Cependant je *peux* tout autant *voir* que « $\mathcal{H}(F)$ » est vrai. En d'autres termes, (¬c)* n'interdit pas (c), au contraire, s'il est possible d'admettre (¬c)*, alors il doit être possible d'admettre aussi (c). Dans la perspective de Lucas, aucune impossibilité logique ne peut nier directement (c), et (c) exprime justement cela d'une certaine façon. Ainsi, l'objection de Putnam ne tient plus, le raisonnement est valide.

L'objection de Putnam n'est valide que si un point de vue externe peut constater que F est consistant sans avoir à le dévoiler à Lucas. Mais la validité de cette objection s'estompe si Lucas joue à la fois son propre rôle de sceptique et le rôle du mécaniste. Il ne peut alors se cacher à lui-même la consistance de F, il ne peut donc s'interdire de voir

[50] C'est le célèbre théorème de Tarski de 1936, tel que rapporté dans Mendelson 1997, p. 217.

que « $\mathcal{M}(F)$ » est vrai. Cette version à un joueur de l'argument dialectique conduit tout droit à la contradiction tant espérée entre les propositions (a) et (e). En fait, il faudrait peut-être, suivant la suggestion de Lucas, revoir les propositions (a), (c), (d) et (e) pour mieux refléter la forme « monologuée ». (a) devient :

(a)' Je peux voir que (a) est vrai,

et ainsi de suite pour les autres. Sous la nouvelle forme de ces propositions modifiées, l'argument est résolument valide.

4.5 La curiosité de Whiteley

Voici une objection si originale qu'elle n'a pas été classée. Elle suggère une incomplétude de l'esprit même dans le contexte de l'IA faible. Existe-t-il une proposition informelle indécidable pour chaque être humain? C. H. Whiteley[51] a proposé une telle phrase : « Lucas ne peut affirmer la vérité de cette phrase sans se contredire ». Tout le monde peut affirmer la vérité de cette phrase, excepté Lucas qui ne peut affirmer que cette phrase est vraie sans se contredire. Il s'agit donc d'une proposition indécidable qui ne s'applique qu'à Lucas. En ce sens, Lucas est lui-même « incomplet », il ne peut affirmer toutes les vérités du monde. Pourtant, une machine pensante pourrait très bien affirmer la vérité d'une telle phrase. Ainsi, Lucas se retrouve aussi incomplet que les machines pensantes. Et comme on peut changer dans la phrase le nom de Lucas pour celui de n'importe quel être humain, tout être humain est donc nécessairement incomplet au même titre que les machines. Alors, match nul?

Lucas pourrait tenter de contester l'indécidabilité de la phrase. Est-elle indécidable au même titre que la proposition gödelienne? Car ni celle-ci ni sa négation ne **peuvent** être prouvées par le système formel associé. Lucas peut-il affirmer la négation de la phrase indécidable? S'il peut l'affirmer, alors la négation est vraie et la phrase indécidable est fausse. Mais si elle est fausse, Lucas ne peut l'affirmer (car Lucas n'affirme pas le faux). Mais c'est justement ce qu'affirme la phrase indécidable! Donc elle est vraie, et Lucas ne peut l'affirmer.

Les propositions indécidables rappellent très certainement le paradoxe du menteur. Mais nous devons réaliser ici que ce qui évite la proposition indécidable de

[51] Whiteley 1962, pp. 61-62

sombrer dans le paradoxe est qu'elle identifie un individu (ou un système formel) en particulier pour lequel elle s'applique. La phrase « Personne ne peut affirmer cette phrase sans se contredire » n'identifie personne en particulier, et par conséquent est un paradoxe. Car si elle est vraie, quelqu'un peut en principe affirmer cette phrase sans se contredire, mais ce faisant, elle devient fausse. Mais si cette phrase est fausse, personne ne peut l'affirmer, la phrase est donc vraie. La phrase est donc paradoxale. On pourrait donc conclure au passage que le paradoxe n'est pas loin lorsqu'on tente d'universaliser une proposition indécidable particulière.

4.6 La complexité des systèmes formels

Nous étudions maintenant la question cruciale de la pertinence des limites matérielles à l'argument de Lucas. Comme nous l'avons déjà vu, les raisonnements par l'absurde en mathématiques sont assez courants, et leur légitimité n'est pas contestée par ce qui peut se faire en pratique. Mais les mathématiciens traitent de l'existence d'entités abstraite, alors que le projet du Mécanisme est somme toute bien *concret*. Quelle valeur accordée à une objection de pratique face à un argument de principe à propos d'une existence concrète? À la limite, il s'agit presque d'une question d'attitude personnelle.

Rappelons tout d'abord une conclusion de Lucas : il existe pour chaque système formel associé à une machine pensante possible une proposition gödelienne indécidable que l'esprit peut saisir comme étant vraie. Ainsi, l'esprit comprend la vérité de la totalité des propositions gödelienne possible, ce qui n'est le cas d'aucune machine Cependant, affirmer qu'il existe une proposition indécidable pour chaque système formel ne revient pas au même que d'affirmer que « Cette proposition gödelienne $\mathcal{M}$ n'est pas démontrable par telle machine ». Une machine pensante pourrait aussi bien démontrer la première affirmation, alors qu'il n'est pas du tout certain que l'humain puisse en pratique déterminer exactement le contenu de la seconde affirmation. Ainsi, comme l'humain n'arrive pas à considérer comme vraie la proposition indécidable appropriée, l'esprit n'est pas en pratique différent de la machine. Hofstadter, dans son livre paru en 1979, s'intéresse de façon approfondie aux objections de pratique à l'argument de Lucas :

> À moins d'être doté d'un esprit un peu mystique, on doit simplement en conclure que chaque être humain atteindra les limites de sa capacité de gödeliser [face à des machines suffisamment complexes]. Mais au-delà, les systèmes formels ayant au moins cette complexité, bien qu'incomplets pour la raison évoquée par Gödel, seront tout aussi puissant que cet être humain.[52]

Hofstadter admet bien que la machine pensante sera incomplète en raison du théorème de Gödel, mais si cette vérité est inaccessible en pratique aux humains, aucun de nous ne peut donc prétendre pouvoir à coup sûr faire mieux que toute machine.

Nous noterons tout d'abord que Lucas présuppose que pour toute machine pensante, il soit possible d'avoir accès à son programme pour en déduire son système formel, condition nécessaire si l'on souhaite appliquer le raisonnement de Gödel à l'algorithme de la machine. Comme le souligne Benacerraf :

> If the machine is not designated in such a way that there is an effective procedure for recovering the machine's program from the designation, one may well know that one is presented with a machine but yet unable to go anything about finding the Gödel sentence for it. The problem becomes even more intractable if one supposes the machine to be oneself – for in case there may be no way of discovering a relevant index – of finding out one's own program.[53]

Clairement, nous ne sommes pas dans le contexte d'un test de Turing. Ce ne peut être en observant seulement les résultats auxquels parvient la machine que l'on peut déduire son programme, il faut avoir accès à son programme. Lucas pourrait répondre que cela ne pose aucun problème, car dans la version à un joueur de son jeu entre lui et le Mécaniste, celui qui cherche la proposition gödelienne est le même que celui qui a construit la machine, celui-ci devant bien connaître le programme de sa machine pensante. Cependant, Lucas n'envisage pas la possibilité d'avoir recours aux algorithmes ascendants, c'est-à-dire à une machine qui modifierait elle-même son propre code sans l'intervention d'un constructeur humain.[54] Cette possibilité sera traitée par Penrose plus tard.

---

[52] Hofstadter 1985, p.536

[53] Benacerraf 1967, p.28

[54] Benacerraf commente également ce qui est convenu d'appeler l'argument dialectique de Lucas, à savoir le jeu mené entre ce dernier et un Mécaniste. En supposant que Lucas arrive bel et bien à trouver une vulnérabilité (une proposition indécidable) pour chaque machine que construit le Mécaniste, Benacerraf souligne que cela ne prouve pas la thèse de Lucas, puisque le Mécaniste est assez limité en ce qui concerne ses capacités à construire des machines. Dans les faits, il ne pourra jamais construire toutes les machines possibles, ses capacités mentales le limitent. Donc, lorsque Lucas affirme pouvoir trouver une

Accordons à Lucas la possibilité d'avoir accès au programme de la machine. Comment s'y prendra-t-il pour déduire la proposition gödelienne du système formel associé à ce programme? Existe-t-il une procédure effective? Hofstadter propose de montrer les difficultés auxquelles devra se confronter Lucas dans l'élaboration de cette procédure. Un système formel se veut un système mécanique de manipulations de symboles, mais ce n'est pas du tout le genre de système qui est à la base de la programmation des ordinateurs. Quiconque connaît le fonctionnement d'un ordinateur et la façon dont il est programmé comprend aisément qu'il puisse exister plusieurs niveaux de manipulations de symboles. Au niveau le plus bas, on trouve le langage élémentaire que les composantes électroniques de l'ordinateur utilisent pour exécuter le programme. On appelle ce langage élémentaire *langage machine* dans lequel la formulation des instructions se fait sous forme de séquences de bits. À moins d'être un concepteur de langage informatique, personne ne programme directement en langage machine. On utilise plutôt un langage *évolué*, tel le C++, dont la formulation des instructions est apparentée à un langage naturel, en l'occurrence l'anglais. Les langages évolués des niveaux supérieurs intègrent un compilateur ou un assembleur qui permet de traduire les instructions en langage machine. Le langage machine et sa manipulation par le processeur central de l'ordinateur s'apparente directement avec le type de langage inscrit sur un ruban d'une machine de Turing universelle et sa manipulation par une table d'états internes. Le langage évolué, pour sa part, n'a rien à voir directement avec celui d'une machine de Turing, et les commandes et autres instructions qu'il peut contenir sont en fait des termes qui regroupent et unifient une suite plus ou moins longue d'instructions en langage machine dont le programmeur n'a pas à se soucier directement de la formulation exacte.

Le but de la courte présentation de la hiérarchie des langages informatiques[55] est de montrer qu'il existe plusieurs niveaux de description des actions (manipulation de

---

vulnérabilité pour toute machine possible, il a beau jeu car il ne sera confronté en fait qu'à des machines « limitées ». Selon Benacerraf, l'argument dialectique ne prouve rien. À mon avis, cette critique de Benacerraf montre plutôt que Lucas ne rend peut-être pas complètement justice à la thèse mécaniste. S'il n'est pas possible en pratique de construire une machine pensante par des moyens accessibles aux humains, on ne voit pas comment le projet de l'IA pourrait être mené à terme. Et cette limitation n'a rien à voir avec le théorème de Gödel. Il faut donc accorder aux humains la capacité pratique de construire toute machine pensante possible en ayant recours à toutes les méthodes accessibles.

[55] Pour une présentation plus complète, voir Hofstadter 1985, chapitre X.

symboles) que peuvent exécuter les ordinateurs. Sans qu'il soit nécessaire de définir avec précision chaque niveau, on comprend qu'il existe des niveaux inférieurs qui se rapprochent de la base matérielle de l'ordinateur et du fonctionnement de ses composantes, et des niveaux supérieurs (les langages évolués) de plus en plus éloignés du support matériel mais dont la manipulation de symboles et la description des actions sont beaucoup plus familières au commun des mortels. Il est à noter, et cela s'avère sûrement crucial pour le projet de l'IA, que les langages évolués permettent « de *définir* de nouvelles entités de haut niveau à partir des entités déjà connues, puis de les *appeler* par un nom »[56]. Le programmeur, et l'ordinateur à sa suite, peuvent donc définir eux-mêmes de manière relativement flexible des procédures de traitements d'informations spécifiques à certains domaines (appelés modules), ce qui permet de catégoriser un ensemble d'informations comme un tout de niveau supérieur plutôt que comme une somme de parties de niveau inférieur. Ainsi, par exemple, l'ordinateur pourrait observer un tableau comme un tout indissociable auquel il peut donner un nom plutôt que comme une somme de taches de couleurs.

Imaginons maintenant qu'une machine pensante soit construite et testée. On s'attend à ce qu'elle soit capable d'interagir avec ses interlocuteurs et qu'elle prétende à tout le moins avoir des croyances et pouvoir distinguer le vrai du faux. Autrement dit, elle doit être capable d'interagir avec le niveau du langage ordinaire, un niveau de description qui sera probablement le plus élevé. La question est maintenant de savoir à quel niveau de manipulation des symboles devons-nous nous attendre à « trouver » les manifestations de l'intelligence de la machine? Est-ce dans les niveaux inférieurs ou les niveaux supérieurs? On imagine très bien qu'en eux-mêmes, les niveaux inférieurs ne sont pas suffisants, car ce n'est pas à ces niveaux que se produisent l'interaction directe avec les interlocuteurs ainsi que la manipulation des concepts du langage humain.

Chaque niveau de langage a ses propres règles d'inférences et définit aussi en partie les règles des niveaux supérieurs.[57] Maintenant, imaginons que Lucas se propose de trouver la proposition indécidable de notre machine pensante. Sur quel niveau devra-

---

[56] Hofstadter 1985, p. 326

[57] Pour chaque niveau de langage, un programme ne s'écrit pas n'importe comment et ne permet pas de faire n'importe quoi. La traduction d'un niveau de langage à l'autre se fait selon des règles précises, quoique la correspondance ne soit pas toujours bijectionnelle (Hofstadter 1985, p. 327).

t-il se pencher pour espérer construire une telle proposition? Assurément, il se penchera sur le programme complètement assemblé (ou compilé) et donc rédigé en langage machine pour tenter d'en déduire un système formel. Nous imaginons facilement la très grande complexité d'un tel programme et l'ampleur surhumaine de la tâche qui attend Lucas. Probablement n'aura-t-il pas assez d'une seule vie! Et une fois le système formel découvert, encore faut-il l'arithmétiser pour obtenir la proposition gödelienne, ce qui est encore loin d'être évident.

Dans l'éventualité où Lucas (ou l'un de ses successeurs) finit par obtenir une proposition gödelienne malgré la très grande complexité du programme en langage machine, dans quelle mesure Lucas peut-il affirmer détenir une proposition vraie que ne peut démontrer la machine? La proposition qu'il a obtenue se rapporte au niveau le plus inférieur, alors que ce sont aux niveaux les plus élevés que l'on trouve les manifestations de l'intelligence des machines. Selon Hofstadter, si l'argument de Lucas fonctionne pour le niveau inférieur, il n'est cependant pas applicable aux machines ayant des niveaux supérieurs. Il sous-entend donc que l'argument de Lucas doit pouvoir s'appliquer directement aux niveaux supérieurs pour être valide. Autrement dit, ce n'est pas parce qu'on peut déterminer en pratique une proposition gödelienne pour le niveau inférieur que l'on pourra, même en principe, en déterminer une pour le niveau supérieur aussi.

La question que soulève Hofstadter est des plus cruciales : quelle répercussion l'incomplétude du niveau inférieur a-t-elle sur le niveau supérieur? Est-il possible que la proposition gödelienne de niveau inférieur puisse être traduite dans le langage du niveau supérieur? Sinon, Lucas peut-il toujours affirmer détenir une vérité que ne peut démontrer la machine pensante? Sans doute ici le problème est obscurci par le fait que Lucas fasse référence au système formel équivalent à la machine pensante plutôt qu'à la machine elle-même. Peut-être le problème serait-il plus évident s'il était formulé en termes de problème de l'arrêt des machines de Turing.

Cette question restera ouverte, je ne vois pas comment on pourrait y trouver une solution sans véritablement sombrer dans la spéculation dénuée de sens. Mais cette question nous amène vraiment au-delà de la limite de notre connaissance du problème. Pour l'instant, nous allons supposer que l'incomplétude du niveau inférieur génère une incomplétude du niveau supérieur. Nous continuerons notre examen de cette question

dans le contexte de l'IA forte en compagnie de Benacerraf et de Penrose dans les prochains chapitres.

Avant de passer à la dernière catégorie, notons que les considérations précédentes peuvent aussi s'appliquer à la question de la complexité de la preuve de la consistance du système formel. Par ailleurs, à la lumière de la distinction entre niveau inférieur et niveau supérieur, on peut se demander quelles seraient les répercussions de l'inconsistance du niveau inférieur sur le niveau supérieur. Encore une fois, les résultats sont difficiles à évaluer.

4.7 L'informalité de l'argument de Lucas

La notion de vérité qu'utilise Lucas dans son argument n'est pas formalisable dans le système formel F. Cette impossibilité s'appuie sur le célèbre théorème de Tarski. Lorsque Lucas soutient que $\mathcal{G}(F)$ est vrai mais qu'elle n'est pas prouvable dans F, cette proposition ne peut être montrée par F. Le prédicat de vérité ne peut à la fois s'appliquer à tous les théorèmes de F et se distinguer suffisamment de la notion de prouvabilité pour pouvoir s'appliquer en plus à la proposition indécidable. Une façon de rejeter l'argument serait justement de soutenir qu'il n'est pas convaincant car il n'est pas complètement formel. Si la vérité n'est pas un prédicat de F, qu'elle n'est pas équivalente à la notion de prouvabilité, qu'elle échappe à tout système formel, alors en quoi consiste-t-elle? Lucas admet qu'il ne peut répondre à la question de façon satisfaisante pour celui en quête d'une notion formelle de la vérité. Selon lui, notre faculté à juger de la vérité ne peut se passer d'un recours à l'intuition. La proposition indécidable nous apparaît vraie, et ce jugement n'est pas le résultat d'un calcul.[58]

À cet égard, Benacerraf examine l'expression « produire en tant que vrai »[59] qu'emploie Lucas pour désigner le fait qu'il saisit comme vraie la proposition gödelienne. Aux yeux de Benacerraf, Lucas semble avoir ici recours à une sorte de raisonnement informel. Mais il y a un aspect équivoque dans le fait que la machine

---

[58] Cette réponse déplaira au formaliste, mais elle est tout à fait plausible pour le réaliste. À cet égard, Lucas ne s'avance nullement à défendre une vision réaliste des vérités mathématiques. Il semble plus intéressé en ayant recours au théorème de Gödel à justifier certaines considérations éthiques telles que l'autonomie de la raison et le libre-arbitre face à la thèse déterministe du Mécanisme.

[59] « Produce as true »

pensante soit incapable de produire comme vraie une proposition indécidable alors que tout humain en est capable. Qu'est-ce que le premier théorème de Gödel prohibe la machine pensante de faire? Prouver $\mathcal{A}(F)$ à partir de ses propres règles et de ses propres axiomes. Lucas peut-il faire mieux en utilisant ces mêmes règles et axiomes? Non! C'est pourquoi Benacerraf croit que le fait de saisir $\mathcal{A}(F)$ comme étant vraie relève d'une forme de preuve informelle au sens où elle ne peut être formalisée dans F. La machine pensante est limitée en ce qui concerne ce qu'elle peut démontrer par preuve formelle, mais qu'en est-il des « démonstrations » informelles? Une machine peut-elle être douée d'une telle capacité à produire informellement comme vraies certaines propositions? Si oui, tant que la machine ne se convainc pas qu'elle détient une preuve *formelle* de $\mathcal{A}(F)$, le théorème de Gödel ne peut être aucunement invoqué pour limiter les capacités de la machine. Cette critique rejoint à certains égards celle de Hofstadter. Les différents niveaux de manipulations de symboles permettent différents niveaux de démonstrations, le niveau supérieur permettant quant à lui des raisonnements informels à propos de la vérité de certaines propositions.

Évidemment, Lucas fustige cette possibilité. Les machines pensantes ne peuvent faire que des démonstrations formelles :

> Machines cannot conjure up anything, but only act according to their input and the way they are wired up. [...] For a machine, and on the mechanist thesis for men too, 'producing as true' is simply and entirely the consequence of certain antecedent conditions, and therefore can be represented as being proved-in-a-formal-system of some sort.[60]

Selon Lucas, pour une machine, toute démonstration aux apparences « informelles » est en fait une démonstration formelle. Entre Lucas et Hofstadter, la discussion achoppe sur la possibilité pour une machine pensante à plusieurs niveaux de faire des raisonnements informels à son niveau supérieur.

Plusieurs critiques utilisent des arguments formels pour montrer qu'il n'existe pas de méthode formelle pour arithmétiser tout système formel et trouver sa proposition gödelienne ou encore qu'il n'est pas possible de prouver sa propre consistance. Mais ces arguments ont-ils une quelconque validité dans le contexte de l'IA faible? La réponse est

---

[60] Lucas 1968, pp. 147-148

non. Implicitement, toutes ces objections formelles ont pour but de montrer les limites *formelles* de l'esprit en assumant que celui-ci est une machine, qu'il est soumis aux mêmes limitations que celle-ci. Mais cela ne concerne pas la version faible de l'IA, et l'examen de ces objections est donc remis au chapitre suivant.

Au final, qu'est-ce que Lucas a accompli? Nous pouvons affirmer sans nous tromper que sa tentative de réfutation du Mécanisme ne remplit que la moitié de ses promesses. Il néglige complètement la thèse de l'IA forte en rejetant comme une pure fantaisie sémantique la possibilité d'étendre le concept de pensée pour y inclure les machines, faisant ainsi de l'esprit une machine pensante. Pourtant, le raisonnement par l'absurde qu'il utilise aurait très bien pu admettre une telle possibilité. C'est d'ailleurs ce que fera Benacerraf. Lucas tente de réfuter l'IA faible en montrant qu'aucune machine pensante ne pourrait être l'égale d'un esprit humain. Il existe pour chaque machine une proposition gödelienne vraie que ne peut démontrer la machine. Il faut bien voir qu'il s'agit d'un argument de principe, ainsi les capacités de l'esprit ne sont nullement limitées par le manque de temps ou de ressources matérielles. Mais pour la plupart des critiques, c'est ce qui est le plus difficile à avaler. Une machine pensante ne sera plus une entité purement abstraite une fois construite et ce jour venu, il ne faudra plus la comparer en principe à l'esprit, mais bien en pratique. Il faudra alors déterminer en pratique si elle est consistante et quelle est sa proposition gödelienne. La complexité du fonctionnement du cerveau, qu'on ne peut directement comparer au fonctionnement de la machine dans le contexte de l'IA faible, laisse tout de même présager aux critiques que cette tâche sera tout à fait surhumaine.

Deux autres objections se rapportent à la réelle portée d'une proposition gödelienne vraie. Whiteley propose une phrase que seul Lucas ne pourrait affirmer et ce, sans avoir aucunement recours au raisonnement formel du théorème de Gödel. Ainsi, les propositions gödeliennes ne seraient pas les seules à produire une incomplétude chez tout être pensant identifié individuellement. De son côté, Hofstadter remet en question la répercussion que pourrait avoir une proposition gödelienne pour une machine pensante. En distinguant les niveaux de descriptions des actions de la machine par le biais des différents niveaux de langages de programmation, on réalise que l'existence d'une proposition gödelienne au niveau inférieur ne produit peut-être pas une incomplétude au

niveau supérieur qui est celui où les comportements intelligents (et intelligibles pour un humain) se manifestent.

En somme, la réfutation de l'IA faible par Lucas bat de l'aile. La puissance de raisonnement que l'on doit accorder en principe à l'esprit humain laisse plutôt froids autant les philosophes que les spécialistes de l'IA. À leurs yeux, l'argument de Lucas apparaît beaucoup trop simpliste.

CHAPITRE 5.
L'ARGUMENT DE BENACERRAF

5.1 Benacerraf, critique de Lucas

À la fin du chapitre précédent, nous avons vu que Benacerraf s'intéressait particulièrement à la notion de preuve qu'emploie Lucas lorsqu'il utilise l'expression « produire en tant que vrai ». De l'avis de Benacerraf[61], cet usage est équivoque, à cheval entre la notion de démonstration formelle et celle de raisonnement informel. Mais il est possible d'interpréter l'argument pour éviter le double sens de « preuve ». Accordons à Lucas qu'il puisse prouver (en un certain sens qui reste à définir) $\mathcal{G}(F)$, la proposition gödelienne d'une machine pensante équivalente à un système formel F consistant. Si prouver une proposition signifie seulement montrer qu'elle est un théorème d'un système formel particulier, il n'y a rien là que F ne peut pas faire. En effet, il est possible d'énumérer l'ensemble des systèmes formels W_i (où i spécifie un système formel particulier) de telle sorte que la relation $R_{W_i}[(A_1,\ldots,A_n), (A_n)]$, telle que $A_1, \ldots, A_n$ est une preuve de A_n dans W_i, peut être traduit dans F si bien que si une telle relation est vraie, F peut la prouver. En particulier, si on ajoute simplement $\mathcal{G}(F)$ à F pour donner F_1, la proposition « $\mathcal{G}(F)$ est un théorème de F_1 » peut être prouvée aussi bien par Lucas que par F, ce qui ne permet pas de les différencier.

Mais si nous imaginons que les axiomes de chacun des W_i sont des *théorèmes* de Lucas, il est clair que Lucas peut *prouver* $\mathcal{G}(F)$ dans un sens qui échappe à F. Benacerraf propose de parler de *prouvabilité absolue* qui inclut tout ce que peut F peut démontrer ainsi que certaines propositions qu'aucun F ne peut démontrer. En ce sens, Lucas peut soutenir prouver des propositions vraies qui ne peuvent être prouvées par aucun système formel. Benacerraf en déduit que l'union des systèmes formels qu'il peut produire n'est pas un système formel en lui-même (en supposant bien sûr que l'union demeure consistante). Si véritablement il existait cette sorte de prouvabilité absolue dont pourrait se réclamer Lucas, alors il ne serait pas une machine puisqu'il serait

[61] Benacerraf 1967, pp. 20-21

essentiellement différent de tout système formel. Mais comme le remarque Benacerraf, Lucas ne fournit aucune raison de croire qu'une telle signification de la notion de preuve soit bel et bien celle qui est en jeu lorsqu'il soutient pouvoir produire comme vraies des propositions indécidables.

5.2 Benacerraf, complément de Lucas

Aux yeux de Benacerraf, l'argument de Lucas est loin d'être concluant. Comme nous le verrons, ce constat est légitime puisque Lucas néglige complètement la version forte de l'IA. Benacerraf propose une relecture de l'argument original dans laquelle il ne répugne pas à considérer l'esprit comme une machine. Ce faisant, il aborde de front la question des limitations imposées par le théorème de Gödel dans le contexte de l'IA forte. L'argument que met de l'avant Benacerraf demeure un raisonnement par l'absurde mais a l'avantage d'être beaucoup plus clair, plus formel, au sens où il se présente comme une suite de vingt propositions en langage formel qui, au bout du compte, dérivent une contradiction. Il démontre ainsi qu'une des prémisses du Mécanisme doit être fausse, mais que ce n'est pas nécessairement celle affirmant que l'esprit est une machine. Contrairement à l'argument de Lucas, celui de Benacerraf ne fait pas appel à une compréhension intuitive de la vérité de la proposition indécidable associée à une machine. La contradiction repose essentiellement sur la possibilité que l'on puisse prouver à la fois la consistance du fonctionnement de l'esprit et des machines pensantes.

Dans un premier temps, Benacerraf formalise le pouvoir déductif de son esprit. Soit

(1) $S = [\, x \mid \text{L peut prouver } x]$,

c'est-à-dire que l'ensemble S contient tout ce que L (un individu particulier) peut saisir comme vrai dans un sens qui ne se réduit pas nécessairement à ce qui peut être démontré dans un système formel quelconque. On assume par ailleurs, tout comme Lucas, que S ne contient que des vérités. Mais S, comme le conçoit Benacerraf, n'est pas fermé sous la logique du premier ordre avec identité. Autrement dit, S ne contient pas toutes les démonstrations mathématiques possibles, certaines d'entre elles étant possiblement trop complexes ou trop longues pour que L en vienne à bout. C'est pourquoi Benacerraf introduit S^* qui se veut le sur-ensemble de S fermé sous la logique du premier ordre

avec identité. S^* contient donc en principe l'ensemble des vérités auxquelles a accès dans sa vie un esprit humain normalement constitué plus toutes les combinaisons valides purement logiques auxquelles elles peuvent participer, alors que S est limité par des contraintes matérielles. Le reste du raisonnement s'appuyant surtout sur S^*, le pouvoir déductif complet d'un individu, la distinction précédente montre bien que nous avons affaire à un argument de principe plutôt que de pratique. Ainsi, dans ce qui suit, le savoir de L (incluant par définition ce qu'elle peut prouver) se rapportera à S^* plutôt qu'à S.

Comme S ne contient que des propositions vraies et que la logique du premier ordre préserve la vérité, S et S^* sont consistants. Mais ce qui vient d'être démontré par Benacerraf doit bien faire partie de S, et par conséquent de S^*. Autrement dit, Benacerraf sachant qu'il est consistant, toute proposition exprimant la consistance de son pouvoir déductif idéal fait ainsi partie de S^*. Pour l'instant, le raisonnement se veut informel et non pas le fruit d'un système formel, et en ce sens il échappe à une application du théorème de Gödel.

Jusqu'à maintenant, l'esprit n'était pas à proprement dit considéré comme une machine, et nous avons principalement établi que l'ensemble de la production déductive de l'esprit était consistante et qu'un esprit pouvait prouver la consistance de sa propre production du simple fait qu'il la sait composée uniquement de vérités. C'est à la neuvième étape de son raisonnement que Benacerraf formalise l'hypothèse que l'esprit est une machine. Il faut bien comprendre que si les conditions de 9 conduisent à une contradiction, alors aucune machine pensante ne peut satisfaire à la fois ces trois conditions :

> 9. Suppose there is a recursively enumerable set W_j such that
>
> a) '$Q \subseteq W_j$' $\in S^*$ [62]
>
> b) '$W_j \subseteq S^*$' $\in S^*$
>
> c) $S^* \subseteq W_j$ [63]

Puisque la thèse du Mécanisme soutient que le pouvoir déductif de l'esprit peut être représenté par une machine de Turing et que la production des théorèmes de celle-ci forme nécessairement un ensemble récursivement énumérable, le pouvoir déductif de l'esprit peut donc être représenté par un ensemble récursivement énumérable W_j, où j est un nombre naturel qui permet d'identifier individuellement chaque machine possible et

---

[62] La théorie Q est la théorie la plus faible adéquate pour l'arithmétique.

[63] Benacerraf 1967, p. 25

d'en spécifier le code (programme) complet de la machine. Cela est rendu possible par le fait qu'il existe une énumération de toutes les machines possibles au moyen de nombres de Gödel encodant tous les programmes possibles. Pour éviter certaines ambiguïtés, j'introduirai T_j comme étant la machine pensante spécifiée par j et permettant de produire W_j, l'ensemble récursivement énumérable.

La première condition 9(a) impose non seulement que W_j ait les capacités de représenter l'arithmétique simple, mais aussi que cela fasse partie du pouvoir déductif complet de L. Cela aura son importance plus loin dans le raisonnement de Benacerraf. La condition 9(b) entraîne que L puisse prouver que W_j est un sous-ensemble de son pouvoir déductif. Autrement dit, T_j ne peut démontrer plus que le pouvoir déductif idéal du critique, ce qui est une façon de s'assurer que W_j est bien consistant. Nous comprenons ici que la seule façon de s'assurer de la consistance de la machine pensante est de restreindre son pouvoir déductif à ce qui est vrai selon S^*. En ce sens, la machine ne peut pas en principe parvenir à des vérités qui dépassent carrément l'entendement humain. Il s'agit là très certainement d'une limitation imposée au projet de l'intelligence artificielle qui sera développée plus loin.

9(c) indique que le pouvoir déductif complet de L est un sous-ensemble de W_j, donc L ne peut démontrer plus que T_j. Il s'agit là à proprement dit de la thèse mécaniste selon Benacerraf[64] : la machine pensante démontre tout ce que l'esprit peut en principe prouver.

Le reste du raisonnement découle principalement des hypothèses 9(a), 9(b) et 9(c) ainsi que de l'application des deux théorèmes de Gödel. 9(a) garantissant que W_j est adéquat pour l'arithmétique simple, cela permet d'appliquer les deux théorèmes de Gödel de sorte que l'énoncé « $Con(W_j) \supset \mathcal{A}(W_j)$ » est un élément de W_j (W_j peut démontrer le second théorème de Gödel). Par ailleurs, selon 9b, W_j étant un sous-ensemble de S^*, il doit être consistant. Benacerraf s'emploie alors à démontrer au moyen de 9c qu'il existe une formule exprimant la consistance de W_j (dans le langage de W_j) que T_j peut démontrer. Suivant le second théorème de Gödel, cela entraîne que W_j est inconsistant. Or W_j étant un sous-ensemble de S^*, cela implique que S^* est lui-même

[64] Benacerraf croit à tort qu'il s'agit là de la vision de Lucas de la thèse mécaniste (p. 26). Lucas n'envisage jamais réellement cette version forte de la thèse mécaniste.

inconsistant, contredisant l'hypothèse de départ concernant la consistance de l'esprit de L.

Si on choisit de ne pas contester l'hypothèse de départ, on doit conclure qu'aucune machine pensante ne peut satisfaire la conjonction des hypothèses 9a, 9b et 9c. Benacerraf croit que selon Lucas, c'est 9c qui fait défaut. Benacerraf suggère que l'on pourrait très bien proposer de conserver 9c et de rejeter plutôt 9a ou 9b. Ainsi, si l'esprit de L est une machine (de Turing) dont la production est incluse dans W_j (hypothèse 9c), alors soit il ne peut prouver que W_j est adéquat pour l'arithmétique simple (rejet de 9a), soit il ne peut prouver qu'il peut démontrer tout ce que W_j contient (rejet de 9b). Cependant, le rejet de 9b semble plus plausible que celui de 9a, car il semble beaucoup plus aisé pour n'importe qui de démontrer 9a plutôt que 9b. D'où la conclusion de Benacerraf (jouant le rôle de L) :

> It seems to be consistent with all this that I am indeed a Turing machine, but one with such complex machine table (program) that I cannot ascertain what it is. In a relevant sense, if I am a Turing machine, then perhaps I cannot ascertain which one.[65]

En fait, la négation de 9b interdit à L de *savoir* qu'il peut démontrer tout ce que peut démontrer T_j. La seule façon de prévenir cela serait qu'il soit impossible pour L de spécifier au complet le code-programme de T_j, car s'il le connaissait, il n'aurait qu'à suivre le code pour arriver à démontrer lui-même tout ce que la machine peut prouver. C'est pourquoi Benacerraf suggère que le programme soit trop complexe pour être connu par un humain. De façon plus générale, en admettant que L est une machine de Turing, non seulement L ne peut déterminer laquelle il est, mais de plus L ne peut identifier aucune autre machine comme étant équivalente à la machine qu'il est.

Il s'agirait là d'une véritable limitation imposée par le théorème de Gödel sur l'interprétation forte de l'IA. Cette limitation n'implique pas la mort du Mécanisme comme le clamait Lucas au chapitre précédent. Ce résultat est tout de même significatif d'un point de vue philosophique. Selon Benacerraf, il est permis de croire que nous sommes bien des machines de Turing, mais d'un type impossible à identifier précisément. Nous sommes condamnés par le théorème de Gödel à une certaine ignorance sur notre nature profonde. En un sens, cela expliquerait pourquoi Lucas peut

---

[65] Benacerraf (1967), p.29

bien argumenter en faveur de sa propre consistance sans pouvoir toutefois utiliser toute la rigueur formelle de la machine qui sous-tendrait les raisonnements de son esprit. Nous y reviendrons.

5.3 Lucas, critique de Benacerraf

On s'en doute, la conclusion de Benacerraf ne plaît pas spécialement à Lucas. Mais sa réplique[66] lui offre l'opportunité de réfuter l'interprétation forte de l'IA, ce à quoi il n'était nullement parvenu avec son argument original. Aux yeux de Lucas, la conclusion de Benacerraf rend l'IA forte vide de contenu. En effet, comment peut-on soutenir que l'esprit est une machine tout en étant empêché de savoir quelle machine nous sommes? Selon Lucas, la conclusion de Benacerraf est coupable d'ω-inconsistance. En effet, pour toute machine qui pourrait lui être présentée comme étant celle qui représente son esprit, Benacerraf connaît un raisonnement (reposant sur le théorème de Gödel) qui l'oblige à rejeter cette machine. Pourtant, Benacerraf continue de croire qu'il est bien une machine, malgré qu'aucune d'entre elles ne peut lui apporter satisfaction. Pour Lucas, l'ignorance de la spécification du programme de la machine n'est pas due simplement à la complexité, elle doit être beaucoup plus fondamentale :

> The only way I can be absolutely sure of not knowing that I am any particular machine is by not being any machine whatever.[67]

La réplique de Lucas est sévère et montre que Benacerraf ne peut pas jouer la carte de l'ignorance sans rendre la thèse mécaniste complètement creuse ou à tout le moins, peu satisfaisante pour les spécialistes de l'intelligence artificielle. Si vraiment nous sommes empêchés de savoir quelle machine nous sommes, comment soutenir que le projet de l'IA forte peut réellement aboutir?

Le problème de la réplique de Lucas est que si dans le contexte de l'IA faible il pouvait toujours invoquer ses compétences informelles lorsqu'il s'agissait de prouver la consistance d'un système formel ou la vérité de la proposition gödelienne, maintenant, il ne peut y avoir recours. Comme notre esprit est une machine, les arguments formels qui pouvaient être rejetés dans l'IA faible à cause de leur impertinence sont ici admis. Quelles en sont les conséquences? L'argument de Benacerraf conclut implicitement

---

[66] Lucas (1968)
[67] Lucas (1968), p.151

qu'il doit être possible d'ignorer son algorithme tout en sachant que notre esprit est une machine, et c'est précisément ce que nie Lucas. Or, voilà qu'un argument formel nous donne de bonnes raisons de croire que la conclusion de Benacerraf est bel et bien possible.

Tout d'abord, voyons quelle forme pourrait prendre l'ignorance de Benacerraf. Il s'avère que L doit être empêché de connaître que $S^* \subseteq W_j$. Autrement dit, il doit être incapable de déterminer si toutes les propositions vraies de S^* sont incluses dans W_j. Mais si l'esprit est une machine, S^* est obtenue à partir d'un algorithme. Or ce qui permet à Lucas de dire que l'esprit est différent de toute machine est le fait que S^* peut contenir en principe les propositions gödeliennes de toutes les machines pensantes possibles. Mais pour cela, il doit exister un algorithme qui permettrait de trouver infailliblement une proposition gödelienne pour n'importe quelle machine, aussi complexe soit-elle. Le problème, c'est qu'il n'existe pas de tel algorithme. La raison en est que l'énumération des systèmes formels qui sont générés à partir de l'ajout systématique de nouvelles propositions gödeliennes repose sur l'usage des nombres ordinaux dont l'énumération n'est pas algorithmique.

Pour y voir plus clair, imaginons au départ un système formel F_0 relativement simple, dont nous pouvons prouver la consistance et pour lequel nous pouvons trouver une proposition gödelienne $\mathcal{H}(F_0)$. F_0 est donc incomplet. Plus encore, il est essentiellement incomplet, ce qui signifie que si l'on ajoute $\mathcal{H}(F_0)$ à F_0, on obtient un nouveau système formel F_1 qui demeurera consistant mais qui ne sera pas plus complet. En effet, il est possible d'obtenir une nouvelle proposition indécidable $\mathcal{H}(F_1)$ que l'on peut ajouter à F_1 pour obtenir F_2, et ainsi de suite. Imaginons maintenant que nous puissions intégrer la totalité de l'ensemble infini des propositions gödeliennes des systèmes formels précédents $\{\mathcal{H}(F_0), \mathcal{H}(F_1), \mathcal{H}(F_2), \ldots\}$ à un nouveau système formel, F_ω, où ω représente le premier nombre ordinal transfini.[68] F_ω est consistant, mais n'est pas pour autant complet. Il est encore possible d'obtenir une nouvelle proposition

[68] King 1996, p. 382. Comme il le souligne, ω n'est pas le plus grand nombre naturel, il représente plutôt la limite de l'énumération 1, 2, 3, ...

gödelienne $\mathcal{G}(F_\omega)$ que l'on ajoute au nouveau système formel suivant $F_{\omega+1}$. Et ça continue!

Le processus itératif de formation de nouveaux systèmes formels est non borné. Cependant, comme nous venons de le constater, il fait intervenir une notation reposant sur les nombres ordinaux transfinis (ω, $\omega+\omega$, ω^2, ω^ω, etc.). Or, comme le souligne Hofstadter, selon un théorème de Church et Kleene et déduit de la thèse de doctorat de Turing[69], « il n'existe aucun système de notation de nature récursive capable de donner un nom à tout ordinal constructif »[70]. En somme, il n'existe pas d'algorithme qui puisse générer l'énumération complète des ordinaux transfinis. En conséquence, comme le processus, appelé gödelisation, qui permet de trouver une proposition gödelienne repose sur la construction de cette suite, on en déduit qu'il n'existe pas d'algorithme permettant de mécaniser la gödelisation.

Remarquons que dans le contexte de l'IA faible, une telle conclusion vient renforcer la thèse de Lucas en venant confirmer que les machines de Turing n'ont pas les ressources pour trouver toutes les propositions gödeliennes possibles. Mais si l'esprit est une machine, cette conclusion vient plutôt confirmer la possibilité d'ignorer notre identification à une machine pensante en montrant une limitation à notre connaissance de l'ensemble des propositions vraies, une limitation qui serait partagée par toute machine pensante.

Cette conclusion de Benacerraf a tout pour plaire aux spécialistes de l'IA. Bien sûr, elle montre une limitation pour toute machine, mais comme elle est partagée par l'esprit humain, on ne peut plus proclamer ce dernier supérieur aux machines pensantes. La limitation partagée impose un nivellement qui conforte le projet de l'IA forte. Pourtant, lorsqu'on réalise que ce nivellement repose sur une forme d'ignorance inaliénable, c'est-à-dire l'incapacité d'identifier notre propre machine qui sous-tend les raisonnements de notre esprit, la réplique de Lucas apparaît tout à fait pertinente. Comment garantir cette ignorance inaliénable si ce n'est en niant totalement que nous sommes des machines pensantes? Mais cette réplique assume à tort que nous possédons une méthode mécanique qui nous permettrait d'identifier à coup sûr notre machine

---

[69] Turing 1939
[70] Hofstadter 1985, p.536

pensante. Or il s'avère qu'il n'existe pas d'algorithme qui permette de faire cette identification. Donc la réplique de Lucas tombe à plat.

Cette conclusion apparaît d'autant plus convaincante qu'elle ne semble pas affectée par les considérations de Hofstadter. En aucun cas nous avons senti le besoin de faire intervenir les différences de niveaux de manipulations symboliques pour réfuter les propos de Lucas. Autant l'argument de Benacerraf que l'objection apportée à la réplique de Lucas semblent se situer au niveau inférieur, celui-ci se rapprochant le plus du fonctionnement d'une machine de Turing. Bien sûr, nous aurions pu faire intervenir les considérations de Hofstadter en soulignant que dans le contexte de l'IA forte, la machine pensante est dotée des mêmes facultés de haut niveau que l'esprit humain. Or ce n'est pas à ce niveau que l'on peut appliquer le théorème de Gödel, ainsi autant la machine pensante que l'esprit ne sont pas soumis aux limitations issues du théorème de Gödel. De telles considérations sont peut-être acceptables, mais elles escamotent complètement la possibilité d'une application du théorème de Gödel et empêchent toute conclusion à la Benacerraf qui nous apparaît d'un intérêt philosophique remarquable.

CHAPITRE 6.
L'ARGUMENT DE PENROSE

6.1 L'effort de Roger Penrose

Le livre *Les ombres de l'esprit*[71] de Roger Penrose présente ce qui constitue sans doute à ce jour l'argument le plus exhaustif visant à conclure à l'échec de la thèse du Mécanisme en s'appuyant sur le théorème de Gödel. Ce livre succède à *The Emperor's New Mind*[72] dans lequel il développait une première version moins convaincante de son argument contre le Mécanisme. Penrose se reprend donc ici avec une double argumentation qui se révèle en fait être une synthèse de l'argument de Lucas et de sa réfutation de l'argument Benacerraf. Il faut dire que, comparativement à Lucas, l'effort de Penrose se démarque par l'ampleur et la profondeur de son analyse.

Contrairement aux auteurs des deux arguments précédents, Penrose est bien conscient d'une distinction entre les deux interprétations de l'IA. Pourtant, il ne semble pas réaliser toutes leurs implications sur la conception de l'esprit humain. Il se contente seulement d'en mesurer les conséquences sur les machines pensantes. Ce faisant, la réelle portée de sa double argumentation lui échappe.

L'objectif général de Penrose est de montrer que le Mécanisme et les deux versions du projet de l'intelligence artificielle qu'il engendre sont inadéquats à reproduire, simuler ou simplement expliquer le fonctionnement de l'esprit. À son avis, les théorèmes d'incomplétude de Gödel nous donnent la meilleure preuve que l'esprit n'est pas une machine ni ne peut être simulé par une machine. C'est pourquoi, selon Penrose, il faut en déduire que nos connaissances actuelles des lois de la physique sont incomplètes pour rendre compte de l'activité intégrale de notre cerveau, et les phénomènes qui nous restent encore à dévoiler et à comprendre seront essentiellement non calculables au moyen d'une machine de Turing.

Penrose connaît bien l'argument de Lucas et les objections auxquelles il a dû faire face. Parvient-il à en éviter les écueils? Pas vraiment. Ses considérations sont

---

[71] Penrose 1995. Titre original : *Shadows of the Mind.*
[72] Penrose 1989. Titre français : *L'esprit, l'ordinateur et les lois de la physique.*

beaucoup plus étendues que celles de Lucas, mais comme nous le verrons, il n'arrive pas à montrer de façon convaincante que les objections déjà vues ne sont pas légitimes, principalement parce que son argumentation confond toujours les implications des deux versions de l'IA.

Mais avant d'aller plus loin, nous allons nous pencher sur le platonisme de Penrose. Celui-ci a une influence sur la conception de l'esprit humain que peut avoir Penrose et montre comment cette conception *interdit* à Penrose de considérer jusqu'au bout l'hypothèse que l'esprit est une machine.

6.2 Le platonisme de Penrose

Le platonisme mathématique de Penrose l'oblige à faire de l'esprit un organe de perception des entités mathématiques, une perception qui serait semblable à celle des entités physiques de notre environnement. En bref, il explique que le monde physique, le monde matériel indépendant de notre perception humaine comporte tellement de régularités mathématiques, sa structure et son comportement s'accordent si bien à des descriptions mathématiques que ce monde doit bien être d'une façon ou d'une autre lié à un monde *platonicien* contenant toutes les vérités mathématiques possibles. La nature de ce lien reste mystérieux toutefois. Parallèlement, un autre mystère prévaut aussi au fait que notre esprit humain a un accès à ce monde platonicien grâce à sa raison et à son intuition. Penrose croit notre esprit capable de discerner les entités mathématiques garantes de nos vérités, rendant par le fait même les vérités mathématiques accessibles à notre esprit. Lorsque Penrose admet :

> Peut-être mes préjugés sont-ils erronés. [...] Peut-être existe-t-il des vérités mathématiques qui sont *fondamentalement inaccessibles* à la raison et à l'intuition humaines[73]

nous en déduisons qu'il a la conviction implicite que la totalité des vérités mathématiques est accessible à notre esprit. Et cette conviction influence son interprétation des conséquences du théorème de Gödel :

> 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

[73] Penrose 1995, p. 406

> monde mathématique platonicien. La vérité mathématique n'est pas déterminée arbitrairement par des règles d'un système formel d'origine humaine; elle a un caractère absolu, irréductible à tout système de règles algorithmiques. Cet a priori platonicien [...] fut à la base des motivations de Gödel.[74]

Si Gödel admettait sans peine son « a priori platonicien », il semble que celui-ci n'ait pas eu la même portée que celui de Penrose. En effet, Gödel hésitait à trancher la question du Mécanisme au moyen de ses théorèmes d'incomplétude et du platonisme qui a permis la découverte de ses derniers. Pour Penrose, seule une forme non algorithmique de conscience permet d'accéder à la totalité des vérités mathématiques. Les propositions mathématiques réfèrent plus ou moins adéquatement à des « entités » qui existent de façon bien définie dans le monde platonicien abstrait et éternel. L'intuition nous permet de « voir » avec plus ou moins d'évidence la vérité des propositions mathématiques, ce qui confère à l'intuition une légitimité, si ce n'est une nécessité, qui est égale, sinon plus grande que celle accordée aux démonstration formelle. L'intuition humaine transcende ce qui est calculable et permet de comprendre la vérité mathématique à un niveau qui dépasse celui des machines. Du coup, c'est elle qui nous guide lorsque nous développons des systèmes formels pour rendre compte d'une classe de propositions mathématiques en nous assurant qu'elles sont justes. Et c'est elle finalement qui a le dernier mot en ce qui concerne les propositions indécidables issues du théorème de Gödel.

Cependant, tout comme la réalité physique, on peut assumer que la réalité platonicienne ne se contredit pas. Le principe de la non-contradiction s'applique à ce monde, rien ne peut à la fois être et ne pas être une chose à un moment précis sous un même aspect. Ainsi, l'ensemble de nos croyances mathématiques vraies, c'est-à-dire qui s'accordent parfaitement aux entités de la réalité platonicienne, est nécessairement consistant. Il en découle que notre compréhension des mathématiques est elle-même consistante, et que tout algorithme souhaitant reproduire cette compréhension doit l'être aussi :

---

[74] Penrose 1995, p. 406

> One can see, therefore, how Penrose would be led to claim that to assert something, and to assert that it follows from one's own consistency, are essentially to say the same thing [...][75]

Cela permet à Penrose de se dispenser d'avoir à *prouver* la consistance de l'esprit humain, car son intuition lui garantit qu'il ne peut en être autrement, elle ne peut qu'être consistante si nous voulons qu'elle s'accorde à la réalité platonicienne. Ainsi, la vérité de toute proposition gödelienne associée à une supposée machine pensante se saisit intuitivement.

Penrose ne se gêne donc pas pour avoir recours à l'intuition humaine pour fournir les fondements de notre compréhension mathématique. Le sérieux problème d'un tel recours est que notre intuition s'avère souvent trompeuse, de nombreux mathématiciens ont fait des erreurs en se fiant à leurs intuitions. On n'a qu'à penser à Frege pour un exemple éloquent : il avait l'intuition (et la plupart d'entre nous avec lui) que pour toute propriété bien définie, il existe un ensemble d'objets qui possède cette propriété. Pourtant, le paradoxe de Russell met en échec cette intuition. C'est pourquoi il devient nécessaire de développer des systèmes formels qui évitent les paradoxes et dont l'interprétation standard rend univoque la notion de vérité qui lui est associée:

> Since, according to Penrose, we humans must assume that we do have such knowledge [that no consistent procedure can in any sense know that it is sound], this fact shows the superiority of informal intuition to formal definition. However, we believe such a position to be at odds with the very nature of science in general, and mathematics in particular. The whole purpose of introducing formal methods is to avoid the contradictions which arise from using "obvious" facts about our natural, intuitive notions.[76]

Cette critique n'est pas particulièrement originale et aurait très bien pu s'appliquer à Lucas. Elle a le mérite cependant de bien montrer comment le platonisme peut s'inscrire en porte-à-faux avec le projet général des mathématiques. Cela est suffisant pour jeter un doute sur la justesse des convictions de Penrose.

6.3 Individu ou communauté. La compréhension des mathématiques

---

[75] LaForte et al. 1998, p. 280
[76] LaForte et al. 1998, p. 272

Contrairement à l'argument de Lucas, Penrose est réticent à l'idée de se mettre lui-même en scène lorsqu'il s'agit de trouver la proposition gödelienne d'un système formel. Il suggère plutôt que cette tâche pourrait être menée par n'importe quel mathématicien en mesure de comprendre la démonstration de Gödel. En d'autres termes, ce ne sont pas les capacités d'un mathématicien en particulier qui sont ici en jeu, mais plutôt celles de la compréhension mathématique en général que partagent tous les mathématiciens.

> 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.[77]

Cette dernière suggestion irait très bien dans le sens de la conclusion de Benacerraf. Elle se place évidemment dans un contexte d'IA forte et fait appel implicitement au fait que la production réelle de théorèmes est limitée pour tout humain. Ainsi, l'algorithme de chacun pourrait être équivalent en ce qui concerne l'ensemble fini des théorèmes produits par chacun, mais ne pas être équivalent pour le reste. Chaque être humain serait pourvu d'un algorithme personnel et unique. On imagine facilement l'infinie complexité de la spécification de cet algorithme dont les détails intimes seraient trop compliqués à connaître pour pouvoir appliquer le raisonnement gödelien. En ce sens, l'algorithme propre à chacun serait sûrement inconnaissable. Bien que logiquement possible, cette suggestion apparaît fort douteuse lorsqu'il s'agit d'expliquer le fait que pour tout nouveau théorème découvert par un mathématicien particulier, la communauté des mathématiciens est en mesure de le comprendre et de l'accepter comme étant vrai. C'est simplement ce fait que veut mettre en lumière Penrose en insistant plutôt sur un algorithme universel de la compréhension générale, incluant celle des mathématiques. Si on peut arriver à construire une machine pensante, elle intégrera sûrement cet algorithme.

C'est ainsi que Penrose établit ici une distinction entre le contexte de découverte et le contexte de justification des vérités mathématiques. Peut-être chaque mathématicien fait-il appel à ses propres conceptions et expériences personnelles pour

---

[77] Penrose 1995, p.90

découvrir de nouvelles vérités, mais lorsque vient le temps de justifier ses travaux, il doit faire appel à une justification qui suscitera l'assentiment de ses collègues. Et en principe, une telle justification peut être comprise par n'importe quel mathématicien qui s'en donne la peine. Mais peut-on vraiment distinguer en pratique entre les algorithmes individuels de découverte et l'algorithme universel de justification pour tout mathématicien? Cette question demeure sans réponse. Penrose semble n'y accorder que très peu d'importance, en partie parce que les sujets de discorde entre mathématiciens sont à son avis assez peu nombreux (par exemple, l'existence d'ensembles dont la définition utilisent l'axiome du choix), ce qui permet de croire que chaque classe de mathématiciens qui adhèrent à un certain corpus de convictions mathématiques contestables par une autre classe fait appel à un algorithme de justification propre, ce qui entraîne qu'il existerait un nombre limité d'algorithmes de justification. Les classes possibles de mathématiciens étant limitées et peu nombreuses, les algorithmes divergents associés à chaque classe sont eux-mêmes peu nombreux et en grande partie équivalents. Selon Penrose, pour chaque classe, il sera possible de conduire un raisonnement gödelien pour montrer qu'elle n'utilise pas un algorithme sûr.

En passant d'une perspective individuelle à une perspective de compréhension universelle des mathématiques, Penrose souhaite réfuter la possibilité qu'un algorithme puisse reproduire la totalité des procédures en principe disponibles pour tout mathématicien pour décider de la vérité mathématique. Pour certains critiques, cette proposition est plus forte que la véritable thèse du Mécanisme, affirmant seulement que pour tout mathématicien, ses procédures personnelles sont le fruit d'un algorithme.[78] Autrement dit, celle-ci affirme uniquement que les procédures individuelles de chaque mathématicien sont algorithmiques, non pas que *toutes* les procédures en principe disponibles le sont. Ainsi, si Penrose souhaite véritablement réfuter le Mécanisme, il doit s'opposer directement à la proposition la plus faible, car la réfutation de la plus forte n'entraîne pas nécessairement la réfutation de la plus faible.

Cette objection nie en fait que l'on puisse étendre en pratique la compréhension mathématique limitée dans le temps d'un individu particulier à la compréhension mathématique humaine passée, présente et future. Et pourtant, c'est ce à quoi nous

[78] Laforte et al. 1998, p.277

découvrir de nouvelles vérités, mais lorsque vient le temps de justifier ses travaux, il doit faire appel à une justification qui suscitera l'assentiment de ses collègues. Et en principe, une telle justification peut être comprise par n'importe quel mathématicien qui s'en donne la peine. Mais peut-on vraiment distinguer en pratique entre les algorithmes individuels de découverte et l'algorithme universel de justification pour tout mathématicien? Cette question demeure sans réponse. Penrose semble n'y accorder que très peu d'importance, en partie parce que les sujets de discorde entre mathématiciens sont à son avis assez peu nombreux (par exemple, l'existence d'ensembles dont la définition utilise l'axiome du choix), ce qui permet de croire que chaque classe de mathématiciens qui adhèrent à un certain corpus de convictions mathématiques contestables par une autre classe fait appel à un algorithme de justification propre, ce qui entraîne qu'il existerait un nombre limité d'algorithmes de justification. Les classes possibles de mathématiciens étant limitées et peu nombreuses, les algorithmes divergents associés à chaque classe sont eux-mêmes peu nombreux et en grande partie équivalents. Selon Penrose, pour chaque classe, il sera possible de conduire un raisonnement gödelien pour montrer qu'elle n'utilise pas un algorithme sûr.

En passant d'une perspective individuelle à une perspective de compréhension universelle des mathématiques, Penrose souhaite réfuter la possibilité qu'un algorithme puisse reproduire la totalité des procédures en principe disponibles pour tout mathématicien pour décider de la vérité mathématique. Pour certains critiques, cette proposition est plus forte que la véritable thèse du Mécanisme, affirmant seulement que pour tout mathématicien, ses procédures personnelles sont le fruit d'un algorithme.[78] Autrement dit, celle-ci affirme uniquement que les procédures individuelles de chaque mathématicien sont algorithmiques, non pas que *toutes* les procédures en principe disponibles le sont. Ainsi, si Penrose souhaite véritablement réfuter le Mécanisme, il doit s'opposer directement à la proposition la plus faible, car la réfutation de la plus forte n'entraîne pas nécessairement la réfutation de la plus faible.

Cette objection nie en fait que l'on puisse étendre en pratique la compréhension mathématique limitée dans le temps d'un individu particulier à la compréhension mathématique humaine passée, présente et future. Et pourtant, c'est ce à quoi nous

[78] Laforte et al. 1998, p.277

convie Penrose, il n'y a pas de raison de croire qu'*en principe* la compréhension de tout être humain ne puisse être équivalente à la compréhension générale. L'objection présentée ici se révèle être une nouvelle instance de l'opposition entre principe et pratique, entre limite formelle (tout ce que l'humain peut possiblement concevoir) et limite matérielle (tout ce que l'humain peut réaliser étant donné les contraintes temporelles ou de complexité des problèmes). Si cette compréhension générale n'est pas algorithmique, comment la compréhension d'un être humain particulier, limitée il est vrai, mais potentiellement aussi vaste que la compréhension générale, pourrait être algorithmique? L'intelligence artificielle est à la recherche d'un algorithme qui permettra de décider pour tout raisonnement mathématique s'il est valide ou non, comme en est capable tout être humain, non seulement aujourd'hui mais pour les siècles à venir.

6.4 La conclusion $\mathcal{G}$

Comment Penrose s'y prend-il pour réfuter le Mécanisme en ses deux versions? L'argumentation de Penrose récupère dans un premier temps l'argument de Lucas. Penrose utilise cependant le problème de l'arrêt des machines de Turing en complément de l'incomplétude des systèmes formels en ce qui concerne leur proposition gödelienne. Comme nous l'avons vu dans le chapitre 1, le problème de l'arrêt montre qu'il n'existe pas de calcul A à la fois sûr et complet, c'est-à-dire un calcul qui pourrait déterminer à coup sûr si une machine ne s'arrête jamais. Si, comme Penrose, nous considérons que A doit reproduire la compréhension humaine, nous supposerons qu'il contient *toutes* les procédures algorithmiques possibles dont l'humain a disposé, dispose actuellement ou pourrait éventuellement disposer capable de déterminer si une machine ne s'arrête jamais. L'argument de Gödel-Turing montre alors que A ne peut détenir toutes les vérités. Le fait que $A(k,k)$ ne s'arrête pas lui échappe, mais cette vérité n'échappe pas à la compréhension humaine :

> [...] 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.[79]

[79] Penrose 1995, p.68

Ce résultat conclut donc qu'aucune machine pensante ne pourrait reproduire le comportement de l'esprit, conformément à la conclusion de l'argument de Lucas dans un contexte d'IA faible. Cependant, Penrose en déduit sa conclusion $\mathcal{G}$ qu'il veut décisive :

> $\mathcal{G}$: Ce n'est pas en utilisant un algorithme qu'ils savent sûr que les mathématiciens humains établissent la vérité mathématique.[80]

Selon lui, il suffit d'admettre cette conclusion pour réaliser que le Mécanisme est faux Nous constatons immédiatement que si cette conclusion réfute quelque chose, ce ne peut être que dans l'interprétation forte de l'IA, car l'IA faible n'affirme aucunement que les mathématiciens humains utilisent un algorithme pour établir la vérité mathématique. Elle affirme seulement qu'une machine pensante peut *se comporter* comme un mathématicien humain. En d'autres termes, la conclusion $\mathcal{G}$ n'est pas assez forte pour réfuter l'IA faible. Or, cette conclusion, Penrose la tire d'un argument pratiquement équivalent à celui de Lucas. Pourtant, nous avons bien vu au chapitre 4 que cet argument ne permet de réfuter que l'IA faible. Alors, comment Penrose peut-il extraire une conclusion qui concerne l'IA forte d'un argument qui ne s'adresse qu'à l'IA faible? Il ne le peut tout simplement pas, et nous devons conclure que Penrose a confondu les implications des deux versions de l'IA.

La conclusion $\mathcal{G}$ s'accorde plutôt bien avec la conclusion de Benacerraf. En effet, pour celui-ci, ce n'est pas en utilisant un algorithme *qu'ils savent sûr* que les mathématiciens humains établissent la vérité mathématique, c'est plutôt en utilisant un algorithme *qu'ils ne connaissent pas*. Son argumentation reprenant alors essentiellement la réplique de Lucas contre Benacerraf mais de façon plus complète[81], Penrose ne croit pas qu'on puisse utiliser un algorithme qu'on ne peut connaître.

La question se pose donc à savoir si Penrose ajoute à son argumentation des éléments qui n'ont pas jusqu'ici été envisagés et qui permettraient de réfuter de façon décisive chacune des deux versions de l'IA. Ce qui nous intéressera dans son argumentation sera donc limité aux considérations inédites de Penrose. En ce qui concerne l'IA faible, nous avons pu constater que l'argument de Lucas se heurtait à plusieurs catégories d'objections. Parmi elles, les considérations de Hofstadter sur les

---

[80] Penrose 1995, p. 68
[81] Penrose 1995, pp.152 et ss.

niveaux de manipulations de symboles questionne la réelle portée sur le niveau supérieur d'une incomplétude au niveau inférieur. Penrose n'y accorde aucune importance. Quant à l'objection de la trop grande complexité de l'algorithme d'une machine pensante, il se contente dans un premier temps de répéter dans le sillage de Lucas qu'un argument de principe ne peut être entravé par des objections de pratique. Le problème, c'est qu'une machine pensante n'est pas une entité purement abstraite, on ne construit pas *en principe* une machine pensante, on la construit en pratique. Nos limitations matérielles sont donc pertinentes, malgré ce qu'en pense Penrose.

Cependant, Penrose a plus d'un tour dans son sac. Dans la seconde partie de son argumentation, Penrose récupère la réplique de Lucas à Benacerraf pour réfuter à la fois l'IA forte et l'IA faible. Si, comme nous l'avons vu, la réplique de Lucas ne fonctionne pas dans le contexte de l'IA forte, nous n'en savons encore rien pour l'IA faible. Elle consisterait à dire que pour toute machine pensante, elle ne pourrait accepter le fait qu'elle est construite à partir d'un quelconque algorithme, puisque cela signifierait qu'elle est incomplète si moindrement elle croit être sûre. Mais comme l'esprit n'est pas soumis aux limitations des machines, il pourrait connaître la proposition gödelienne de toute machine. L'algorithme est connaissable pour les humains alors qu'il ne l'est pas pour les machines pensantes. Cet argument inspiré de la réplique de Lucas appliqué dans le contexte de l'IA faible se distingue-t-il vraiment de l'argument original de Lucas? À mon avis, il ne s'agit que d'une sorte de *corollaire*, contrairement à ce que semble penser Penrose. En admettant que la machine soit consciente de son état, elle se rend compte elle-même de son incomplétude sans pouvoir démontrer que les humains sont eux aussi limités de la même façon. En somme, la machine parvient à la même conclusion en utilisant le même raisonnement. L'application à l'IA faible de la réplique de Lucas ne fournit rien de plus à Penrose que son argument original ne pouvait déjà fournir.

Pour ce qui est de l'IA forte, Penrose innove-t-il? Sa présentation visant à montrer l'impossibilité d'ignorer son propre algorithme est un peu plus fouillée. Mais les nouveaux éléments qu'il apporte n'échappent pas vraiment aux critiques formelles déjà formulées dans le chapitre précédent. À cet égard, sa contribution est marginale.

Il s'avère donc que la contribution originale de Penrose ne concernera que l'argument contre l'IA faible. Penrose argumente contre les objections de la catégorie de la consistance (ou de la sûreté, ce qui revient au même) en montrant qu'en principe nous n'avons aucune raison de croire que nous sommes empêchés de connaître la sûreté de l'algorithme. En ce sens, Penrose explore deux possibilités : 1) nous connaissons la spécification de l'algorithme de la machine pensante, mais nous ne savons pas s'il est sûr; 2) nous ne connaissons pas la spécification de son algorithme.[82]

La première possibilité rappelle une suggestion de Gödel lui-même :

> En revanche, sur la base de ce qui a été démontré jusqu'ici, rien n'interdit que 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.[83]

Rappelons seulement que nous la considérons dans la perspective de l'IA faible. Nous ne pouvons en aucune façon démontrer formellement l'équivalence des comportements entre la machine pensante et l'intuition mathématique, ni même détenir une « preuve » informelle, à savoir à la limite une bonne raison de croire que la machine est équivalente, car le simple fait de croire qu'elle est équivalente nous oblige à croire qu'elle est sûre et donc incomplète.[84] Mais cette restriction apparaît peu plausible aux yeux de Penrose, il soutient même qu'une telle restriction serait néfaste au projet de l'IA. Voyons pourquoi.

Toute machine pensante étant équivalente à un système formel, Penrose propose ici de chercher parmi les caractéristiques du système formel ce qui pourrait le rendre non sûr.[85] Un système formel se définit par ses axiomes et ses règles d'inférence. Pour juger de l'équivalence de la machine avec notre compréhension des vérités mathématiques, nous devrons vérifier si les *théorèmes* qu'elle produit sont vrais. Nous supposerons qu'en principe nous, humains, pouvons déterminer de façon indéniable la vérité des

---

[82] Penrose 1995, pp. 121 et ss. En fait, Penrose explore trois possibilités, mais sa première possibilité n'est pas particulièrement pertinente à notre propos.
[83] Penrose 1995, p. 118
[84] Évidemment, plusieurs critiques vont contester qu'une preuve informelle, intuitive nous permette de conclure validement à l'équivalence de la machine.
[85] Clairement, Penrose ne fait aucun cas des considérations de Hofstadter.

théorèmes du système formel. Mais comme l'indiquait Gödel plus haut, nous ne pouvons démontrer que le système formel ne donne que des théorèmes vrais. Cependant, les axiomes étant eux-mêmes des théorèmes relativement simples et évidents, nous pourrons sans doute admettre un jour sans réserve leur vérité. Si donc l'équivalence du système formel est indémontrable formellement ou même informellement, cela sera imputable à au moins une des règles d'inférence qui sera considérée comme fondamentalement douteuse. Mais bizarrement, elle sera aussi pratiquement indispensable et ce, pour tout système formel qui pourrait prétendre à représenter notre compréhension mathématique.

De l'avis de Penrose, nous aurions affaire à un miracle quasi-permanent : comment en effet une règle fondamentalement douteuse pourrait produire des théorèmes vrais à tout coup? Une telle situation apparaît fort déraisonnable aux yeux de Penrose :

> Il me semble qu'il serait extrêmement imprudent de la part d'un adepte de l'IA [forte ou faible] de placer tous ses espoirs dans la découverte d'une procédure algorithmique [...] dont l'existence 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.[86]

Selon Penrose, le tenant du Mécanisme se voit priver de l'ultime argument qui aurait établi l'équivalence de comportements tant recherchée entre l'esprit et la machine.

Cet argument de Penrose n'est pas moins vulnérable aux objections de la catégorie de la complexité. Chalmers, par exemple, conteste que nous puissions aussi aisément que le propose Penrose décomposer l'algorithme en un ensemble d'axiomes et de règles d'inférence de telle sorte que nous puissions finir par constater que les axiomes sont vrais et les règles sont valides. De l'avis de Chalmers, certaines règles pourraient s'avérer beaucoup trop complexes, pratiquement aussi complexes que le système formel lui-même, pour qu'on puisse juger de leur validité.

---

[86] Penrose 1995, p. 127 En fait, Penrose semble aller trop loin dans cette affirmation, mais tout dépend du sens qu'il accorde à « explicite ». Si cela signifie seulement que nous connaissons tous les axiomes et les règles d'inférence du système formel, alors il n'a pas montré que la construction explicite de la procédure algorithmique est hors de portée de notre intelligence. Au contraire, nous avons convenu avec Gödel que nous pourrions la découvrir empiriquement, mais que nous ne pourrions démontrer qu'elle est vraiment la procédure que nous cherchions, celle équivalente à notre compréhension. Dans un second sens, « explicite » référerait à la démonstration de l'équivalence de comportements entre la machine et l'esprit, ce qui serait effectivement hors de portée de notre intelligence.

La deuxième possibilité suppose que l'algorithme soit tout à fait inconnaissable, que nous ne pourrions jamais arriver à déterminer tous ses détails, il y aurait toujours quelque chose qui échapperait à l'intelligence humaine. C'est carrément la possibilité de spécifier l'algorithme qui est remise en question : « ce serait l'immense complexité des détails infimes de la spécification de l'algorithme *F* censé sous-tendre la compréhension mathématique qui soustrairait cet algorithme à la connaissabilité humaine »[87]. Cette complexité n'est pas imputable à la taille de l'algorithme, celle-ci ne constitue pas une limite qui échappe à ce que peut faire en principe un être humain. Nous avons affaire à une complexité qui excèderait les capacités intellectuelles possibles de l'être humain, et cela n'a rien à voir avec le fait que ces capacités sont en pratique limitées dans le temps.

Le problème se pose alors de savoir comment un tel algorithme pourrait être construit? De la même façon que le cerveau humain a été construit, pourrait-on répondre, c'est-à-dire par un processus d'évolution et de sélection « naturelles » mettant en jeu au début des machines primitives qui peuvent apprendre jusqu'au jour où une machine extrêmement évoluée pourra reproduire adéquatement la compréhension humaine, et même probablement dépasser les capacités intellectuelles de ses créateurs humains originels. De génération en génération, pourrait-on croire, les machines amélioreront leurs performances, les algorithmes se complexifieront à un point tel que les humains ne seront plus capables de les comprendre.

Penrose conteste cette possibilité : « si les procédures initiales de l'IA sont algorithmiques et connaissables, tout algorithme résultant *F* sera également connaissable »[88]. Penrose explique que si on conçoit l'évolution des machines comme le résultat de deux sources d'interactions, à savoir les « pressions » exercées par leur environnement (facteurs externes) et leurs mécanismes internes d'apprentissage (facteurs internes), on devra analyser si ces deux sources de l'évolution peuvent générer à un certain moment une complexité telle qu'elle excède les capacités intellectuelles humaines. Cela ne va pas sans rappeler l'un des paradoxes du sorite, à savoir si on regroupe un à un des grains de sable, à partir de combien de grains peut-on appeler cet ensemble un tas. En principe, un être humain pouvant suivre tous les développements de

[87] Penrose 1995, p. 133
[88] Penrose 1995, p. 134

machine qui se font suivant un algorithme, ce qu'on recherche ici serait une source d'évolution si complexe qu'elle en serait incalculable, un humain ne pourrait même plus concevoir la machine résultante.

Tout d'abord, examinons la source environnementale, qui peut se distinguer en deux types. L'environnement artificiel concerne tout ce que la machine pourra apprendre de l'enseignement dirigé d'un éducateur humain. L'environnement naturel regroupera donc tout ce qui sera source d'apprentissage qui ne proviendra pas d'un humain. Le plus important est sans doute que les deux environnements peuvent être simulés par une machine dont les outputs serviraient d'inputs à la machine pensante. L'éducateur humain peut être remplacé par une machine (il s'agit de la thèse même du Mécanisme) ainsi que l'environnement naturel. Il n'est pas nécessaire ici, nous dit Penrose[89], de simuler l'environnement naturel réel, plus simplement faut-il qu'il soit typique, et c'est justement la prétention de la physique de pouvoir suffisamment comprendre les lois de la nature pour pouvoir simuler tout environnement à l'aide de machines. Cette précision permet d'éluder la question des systèmes physiques chaotiques. Nous en savons assez pour en faire une simulation adéquate, mais nous ne pouvons simuler le système réel car nous manquons de données initiales. Mais cela n'a pas d'importance tant que nous pouvons simuler par ordinateur un système qui reproduit un environnement typique. En somme, l'environnement, tant artificiel que naturel, ne peut être qu'une source de développement pour les machines pensantes qui peut être traduite sous forme algorithmique et par le fait même, qui ne sera jamais si complexe qu'elle excèdera les capacités intellectuelles d'un humain.

Si on considère maintenant la deuxième source d'interactions, à savoir les règles internes des machines permettant l'apprentissage, on peut distinguer à nouveau deux types de procédures d'apprentissage. Par procédure d'apprentissage, on entend toute procédure qui, à partir d'un algorithme dont les performances ont été évaluées, permet de générer un nouvel algorithme dont les performances seront supérieures aux précédentes. Chaque procédure fait appel à un algorithme ascendant différent. La première procédure est liée aux réseaux de neurones artificiels, elle permet à une machine de corriger les erreurs qu'elle a commises en renforçant ou en affaiblissant les

[89] Penrose 1995, p. 143

connexions entre les neurones artificiels (devenant ainsi une machine dont l'algorithme est différent), ce qui permet en bout de ligne d'améliorer ses performances. Une autre procédure consiste à créer des algorithmes « génétiques », c'est-à-dire qu'à partir d'un algorithme de base relativement performant, plusieurs autres algorithmes sont générés par un processus plus ou moins aléatoire, chacun étant une version plus ou moins altérée de l'algorithme de base. On sélectionne alors le plus performant de cet ensemble d'algorithmes, qui deviendra à son tour le géniteur de nouveaux algorithmes. C'est deux types de procédures produisent des algorithmes ascendants, et comme nous l'avons déjà établi, les algorithmes ascendants sont entièrement algorithmiques au sens où ils peuvent tous être exécutés par un ordinateur et ce, même si les algorithmes génétiques intègrent des composantes issues de procédures aléatoires ou pseudo-aléatoires.[90]

La question est maintenant de savoir si les algorithmes ascendants, dont l'évolution tient compte d'un environnement qui peut être entièrement simulé sous forme algorithmique, peuvent devenir si complexes qu'ils excèderaient l'entendement humain. Mais le processus évolutif est lui-même algorithmique, c'est-à-dire en principe connaissable par un mathématicien humain. En effet, que l'on soit en présence d'un réseau de neurones artificielles ou d'algorithmes génétiques, les différentes possibilités d'évolution sont entièrement connaissables à chaque nouvelle étape, étant donné que les règles d'évolution sont incluses dans la structure même de l'algorithme ascendant. Et comme l'environnement donne un feedback qui est lui-même algorithmique, rien n'est inconnaissable. Rien ne permet de croire qu'à une étape quelconque de l'évolution de la machine pensante, un élément inconnaissable pourra être intégré à la structure de l'algorithme :

> Ces procédures [ascendantes] 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![91]

---

[90] Bien que les phénomènes aléatoires ne soient pas algorithmiques, Penrose considère qu'on peut les remplacer en pratique par des procédures pseudo-aléatoires.

[91] Penrose 1995, p. 149

> *soient* connaissables par l'homme; sinon, toute tentative de construction du robot sera irrémédiablement vouée à l'échec![91]

En tant que tel, l'argument pourrait s'arrêter ici. Penrose a « montré » que tout algorithme est connaissable, la deuxième possibilité n'est qu'une chimère. La réfutation de la première possibilité a quant à elle montré que tout algorithme de machine pensante connaissable est nécessairement sûr. C'est pourquoi Penrose se sent justifié d'affirmer que si l'algorithme est connaissable pour l'homme, il ne l'est pas pour la machine : celle-ci ne peut connaître son propre algorithme. Mais comme nous l'avons déjà vu, il ne s'agit pas là d'un nouvel argument en sa faveur, ce n'est en fait qu'un corollaire de son argument principal inspiré de Lucas.

En réfutant les deux possibilités au moyen de considérations inédites, Penrose a-t-il réussi à renforcer l'argument de Lucas? Autrement dit, arrive-t-il à éviter les objections auxquelles faisait face Lucas? Étant donné que nous sommes dans le contexte de l'IA faible, seules deux catégories d'objections sont recevables. En ce qui concerne la catégorie de la consistance, peut-être en effet la réfutation de la seconde possibilité offre-t-elle un argument en faveur de la consistance de l'algorithme de la compréhension mathématique. En décortiquant entre facteurs externes et internes le processus d'évolution, Penrose montre que les facteurs externes sont algorithmiques et sûrement aussi non contradictoires que la réalité qu'ils dépeignent. Mais comme l'un de ses facteurs externes concerne l'apprentissage acquis d'un éducateur humain, on peut se demander si l'on ne tombe pas dans une pétition de principe : on chercherait à montrer la consistance d'une machine pensante en s'appuyant sur l'apprentissage que l'on fait d'un esprit humain que l'on *assume* consistant.

Pour ce qui est de la catégorie de la complexité, l'argumentation repose encore sur la distinction entre principe et pratique. En d'autres termes, rien de nouveau sous le soleil qui pourrait avantager Penrose. En fait, c'est même le contraire qui se produit. Penrose fait une concession qui pourrait très bien s'avérer fatale pour sa thèse. À plusieurs reprises en effet, il concède que : « les capacités de raisonnement du robot peuvent éventuellement *excéder* celles de l'être humain »[92] :

---

[91] Penrose 1995, p. 149
[92] Penrose 1995, p. 153

> [...] l'« algorithme inconscient et inconnaissable *F* » 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 [...]
>
> Nous supposons – conformément aux points de vue [de l'IA faible et forte] – que notre robot pourrait, *en principe*, [...] 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.[93]

Le problème, c'est que si nous supposons que la machine excède nos capacités, alors nous n'avons plus les moyens de vérifier la vérité de certaines propositions mathématiques. Et par le fait même, comment pouvons-nous vérifier que l'algorithme est sûr? Nous pouvons savoir qu'il est sûr pour toutes les vérités que nous sommes capables de concevoir, mais au-delà, nous sommes obligés de faire confiance à la machine. Pouvons-nous induire du fait qu'elle a démontré seulement des vérités dans le domaine du connaissable qu'elle démontrera seulement des propositions vraies dans le domaine de l'inaccessible? Le doute est possible et les défenseurs de l'IA faible pourraient y voir un échappatoire aux conséquences du théorème de Gödel.

Le doute est-il pour autant plausible? À cet égard, nous pourrions servir à Penrose sa propre médecine. Nous avons déjà vu que Penrose souhaitait montrer que les éléments caractéristiques d'un système formel ne rendent pas plausible la possibilité qu'un algorithme connaissable d'une machine pensante ne soit pas sûr. Si dans cette situation il pouvait paraître en effet peu crédible qu'une règle d'inférence soit considérée fondamentalement douteuse tout en permettant de démontrer uniquement des théorèmes que nous-mêmes pouvions juger vrais, il en va autrement ici. Car il n'est plus question ici qu'un humain puisse juger lui-même si un supposé algorithme ne donne que des vérités, puisqu'il s'agit d'une machine qui excède l'intelligence humaine. En admettant que tous les axiomes et les règles d'inférence sont connaissables, est-il plausible qu'une fois combinés en un système formel, ils aient des conséquences dont nous ne pouvons

[93] Penrose 1995, p. 152

juger de la vérité, de sorte que le système formel devient douteux? Penrose croyait que si nous pouvions douter de quelque chose, ce ne pourrait être que d'(au moins) une règle d'inférence. Celle-ci pourrait maintenant être douteuse justement parce qu'elle a des conséquences que nous ne pouvons évaluer. Ainsi, elle pourrait très bien paraître justifiée en ce qui concerne les théorèmes à notre portée, ce qui élimine les démonstrations miraculeuses que Penrose trouvait peu crédibles, et paraître douteuse en ce qui concerne certains des théorèmes qui nous échappent. Et Penrose n'offre aucune raison de croire que nous serions obligés d'assumer le contraire, à savoir que les théorèmes hors de notre portée sont bel et bien vrais.

Nous sommes plongés dans un doute qui rappellera à plusieurs le paradoxe sceptique de Wittgenstein tel que présenté par Kripke.[94] La différence est qu'ici nous avons affaire non seulement à l'intelligence humaine mais aussi à une autre forme supérieure d'intelligence, celle des machines pensantes.

On pourrait supposer, comme le fait Penrose, qu'une communauté de machines pourrait s'entendre sur les théorèmes vrais qui échappent à notre intelligence. Est-ce que cela garantirait que les machines pensantes ne disent que des théorèmes vrais? Comme la question est indécidable pour les humains, pourrait-on demander aux machines évoluées leur avis sur leur propre sûreté? Le problème dans ce cas-ci, c'est que le corollaire de l'argument de Lucas peut intervenir, et les machines sont obligées de douter de leurs propres convictions.[95] Aussi étrange que cela puisse paraître, cela n'est pas contradictoire dans ce cas-ci, puisque personne n'est en mesure de démontrer leur sûreté, contrairement au cas où les machines auraient un savoir qui n'excède pas celui des humains. Cela ne vient alors que confirmer notre propre doute sur la sûreté des machines pensantes. Nous ne sommes nullement avancés.

En somme, ce n'est pas tant les machines pensantes qui sont limitées par le théorème de Gödel que l'argument de Penrose. Celui-ci ne peut en effet soutenir que les capacités des machines pourront excéder un jour celles des humains. En supposant que l'algorithme de notre compréhension permette la démonstration par des machines

---

[94] Kripke 1982

[95] Les limites formelles qui ont été développées dans le cadre de l'IA forte (chapitre 5) peuvent ici être invoquées car elles ne s'appliqueraient pas aux humains, seulement aux machines dont les capacités excèdent celles de l'humain.

éliminer en principe ce doute sans donner une signification à la vérité qui va au-delà de la connaissance humaine. Voilà qui est philosophiquement peu banal et même troublant.

En somme, Penrose est-il plus convaincant que Lucas dans sa réfutation du Mécanisme? Tout d'abord, je suis tenté de dire que Penrose est en quelque sorte aveuglé par son platonisme et la conception a priori de l'esprit qui en découle. Cela l'empêche de bien voir la portée de l'IA forte qu'il néglige de la même manière que Lucas, bien qu'il s'en défende. Il confond pourtant la cible visée dans sa conclusion $\mathcal{G}$, l'IA forte, et la cible véritablement atteinte par son argument, l'IA faible. Bien qu'il connaisse les limitations et les objections formelles à son argument dans ce contexte, il ne saisit pas correctement leur pertinence dans l'hypothèse que l'esprit est une machine.

En ce qui concerne l'IA faible, l'analyse que fait Penrose est à plusieurs égards beaucoup plus proche de la réalité des recherches en IA que celle de Lucas. En montrant quel processus, entièrement algorithmique, permettrait une évolution des capacités des machines pensantes jusqu'au stade d'une compréhension supposément équivalente à celle des humains, Penrose soutient que si les processus initiaux qui régissent l'évolution des machines sont connaissables et algorithmiques, tout algorithme qui en résultera sera nécessairement en principe connaissable pour l'être humain. Mais selon Penrose, cet algorithme serait nécessairement perçu comme étant sûr, et cela est suffisant pour en déduire que la proposition gödelienne de la machine est vraie et que la machine ne peut la démontrer.[96] Évidemment, cette conclusion repose sur un argument de principe qui permet à Penrose d'esquiver les objections de pratique. En d'autres termes, nous n'avons pas réellement progressé depuis Lucas. Tous deux adoptent une position qui se rapproche du paradoxe du sorite : il semble impossible de déterminer quand la complexité de l'algorithme deviendra insurmontable comme il semble impossible de déterminer combien de grains de sable permettent de former un tas.

L'argument de Penrose a peut-être un certain mérite en se centrant sur une perspective universelle de compréhension des mathématiques plutôt que sur une perspective individuelle comme le fait Lucas. Ainsi, il peut éviter la réplique de Whiteley. La phrase de Whiteley ne pouvant plus identifier personne en particulier, elle

[96] De manière équivalente dans la perspective du problème de l'arrêt, on pourrait en déduire que la machine ne s'arrête pas alors qu'elle le devrait.

perspective individuelle comme le fait Lucas. Ainsi, il peut éviter la réplique de Whiteley. La phrase de Whiteley ne pouvant plus identifier personne en particulier, elle sombre dans un paradoxe comme nous l'avons vu dans le chapitre 4. Par ailleurs, pour ce qui est de la critique de Hofstadter, notons que Penrose ne s'y intéresse aucunement.

En dernier lieu, il apparaît que l'argument de Penrose est limité à son propre insu à ce que le savoir des machines pensantes n'excède pas celui des humains. Sans cette limitation, nous ne sommes plus en mesure de juger si les machines pensantes sont sûres, et elles-mêmes ne pourront se fier sur leur propre jugement comme nous l'indique la réponse à la réplique de Lucas qui se trouve à la fin du chapitre 5.

7. Conclusion

Le véritable enjeu de ce mémoire est avant tout la réfutation du Mécanisme. Celui-ci soutient de façon générale que le comportement de l'esprit humain peut être reproduit par une machine numérique. Si nous arrivons à montrer que des limitations peuvent être imposées aux machines pensantes par le théorème de Gödel concernant ce qu'elles peuvent affirmer, sans qu'aucune limitation ne frappe l'esprit humain, alors nous avons montré que les machines pensantes ne sont pas l'égal de l'esprit humain. La thèse du Mécanisme est donc réfutée. Plusieurs sont tentés de placer ces discussions dans le cadre d'un test de Turing. Cependant, nous avons vu que ce test oblige la machine à mentir sur sa propre nature ou à faire des erreurs pour paraître plus « humaine ». Mais une telle attitude est incompatible avec l'analyse que nous devons mener. Nous devons être assurés que la machine ne fait pas d'erreurs volontairement et qu'elle paraisse sûre si elle l'est vraiment.

Or il s'avère que le Mécanisme a deux visages. Il existe deux interprétations de l'intelligence artificielle, ce domaine de recherche appliquée dont l'objectif ultime est justement de construire une machine pensante égale en tout point à l'esprit humain. Le projet de l'IA est donc la concrétisation de la thèse mécaniste. Mais chaque interprétation de l'IA a des conséquences différentes sur la conception de l'esprit humain. Pour l'IA faible, l'esprit n'est pas une machine, les mécanismes et facultés de l'esprit sont étrangers à ceux d'une machine, alors que pour l'IA forte, il y a une équivalence entre les mécanismes de l'esprit et de la machine, donc l'esprit est une machine. Cette distinction influence grandement la portée des arguments qui ont été examinés.

Un argument préliminaire s'intéresse au statut de la vérité mathématique. Nous avons retenu deux philosophies des mathématiques particulièrement pertinentes à notre propos. Le formalisme conçoit la vérité mathématique comme le résultat de la manipulation de symboles concrets au sein de divers systèmes formels. En cela, le formalisme peut être associé au Mécanisme. Le réalisme, quant à lui, considère que la vérité mathématique se rapporte à des entités mathématiques abstraites qui existent dans un monde indépendant de l'esprit. On comprend aisément que de ces deux philosophies émergent deux conceptions de l'esprit humain. Pour la première, l'esprit peut très bien

être une machine dont le jugement de la vérité s'appuie sur un calcul. Pour la seconde, l'esprit est doué d'intentionnalité, c'est-à-dire que la signification des propositions mathématiques vraies qu'il peut affirmer se rapporte à des entités abstraites. Comme les machines sont réputées dépourvues d'intentionnalité dans leur fonctionnement, l'esprit ne peut être une machine. Or, de l'avis même de Gödel, il semble que l'obtention même de son théorème soit redevable à son platonisme et que la vérité des propositions gödeliennes soit liée à leur signification abstraite et à la possibilité de « sortir » du système formel dans lequel elle est formulée. On en déduit donc qu'une machine n'est pas en mesure de bien rendre compte de la vérité des propositions gödeliennes et que la thèse mécaniste est incorrecte. Cependant, cet argument n'est pas des plus convaincants, en grande partie parce qu'il repose sur des conceptions philosophiques qui demeurent vagues, intuitives et insatisfaisantes à l'égard de la nature de la vérité mathématique.

J. R. Lucas a proposé en 1962 un argument visant à montrer que le Mécanisme est faux et qui est devenu un véritable incontournable. Le théorème de Gödel montre que pour chaque système formel suffisamment fort, il existe une proposition gödelienne vraie mais indémontrable par le système formel. Comme toute machine est équivalente à un système formel, on en conclut que pour chaque machine il existe une vérité qu'elle ne pourra affirmer, contrairement à l'esprit humain qui peut affirmer toute proposition mathématique dès lors que sa signification lui indique qu'elle est vraie. Comme il s'agit d'une limitation qui ne semble s'appliquer qu'aux machines et non à l'esprit, l'argument mène à une réfutation du Mécanisme.

Ce raisonnement peut paraître convaincant, mais comme nous l'avons vu, il néglige un aspect crucial de la question, à savoir qu'il existe deux versions du Mécanisme. Il en ressort que si cet argument est pertinent dans le cadre de l'IA faible, il passe complètement à côté de l'IA forte, puisque cet argument ne considère pas sérieusement l'hypothèse selon laquelle l'esprit est une machine. Maintenant qu'il a été clairement établi que l'argument de Lucas ne s'adresse qu'à l'IA faible, nous pouvons évaluer les objections qu'il a suscitées. Dans ce contexte, les objections les plus pertinentes concernent notre capacité à établir la consistance du système formel de la machine et à trouver la proposition gödelienne associée. La trop grande complexité d'un tel système formel pourrait très bien avoir le dessus sur nos capacités intellectuelles,

nous privant ainsi de concrètement connaître une vérité que ne peut démontrer le système formel. Ces objections dévoilent un grand enjeu du débat, à savoir ce que les humains peuvent faire en principe et ce qu'ils peuvent faire en pratique. L'argument de Lucas s'en veut un de principe, car il repose sur un raisonnement par l'absurde. En principe, rien n'est à l'épreuve de l'esprit humain, ni la complexité d'un système formel, ni le temps qu'il faudrait pour analyser ce système formel. En pratique, nous savons bien qu'il en va autrement, comme nous l'avons vu, mais cela n'a pas vraiment d'importance pour Lucas. L'ensemble des propositions mathématiques que tient pour vraies l'esprit est consistant, les machines pensantes se doivent de l'être aussi, et cela suffit pour montrer l'existence d'une proposition gödelienne vraie mais non démontrable.

Une autre objection proposée par Whiteley suggère pour chaque être humain particulier une phrase qu'il ne pourrait affirmer, ce qui serait une sorte d'équivalent informel de la proposition gödelienne. Cette objection est notamment pertinente au sens où Lucas s'en remet aux capacités des individus (lui-même en particulier) pour comparer le comportement des machines pensantes et leur trouver une vulnérabilité. Hofstadter quant à lui propose une hiérarchisation des niveaux de manipulations symboliques de la machine pensante et suggère qu'une proposition gödelienne au niveau inférieur, celui du langage machine, n'aurait pas nécessairement une réelle répercussion au niveau supérieur, celui où se manifestent les comportements intelligents.

En somme, Lucas n'atteint pas son objectif. Il écarte la possibilité de l'IA forte sans justification convaincante. Pour ce qui est de l'IA faible, il faut être prêt à concéder que ce qu'on peut faire en principe prime sur ce qu'on peut faire en pratique et ce, aussi bien aujourd'hui que le jour où une machine pensante sera effectivement construite.

Pour sa part, Benacerraf propose un argument qui comble en un sens la lacune de Lucas, c'est-à-dire qu'il s'intéresse à l'hypothèse de l'IA forte. Assumant que l'esprit est une machine, il peut se permettre d'utiliser un raisonnement formel dont la rigueur est peu reprochable. Sa conclusion est remarquable d'un point de vue philosophique : aussi bien les machines pensantes que l'on construirait que l'esprit humain sont soumis à une limitation qui les empêche d'identifier en principe comme en pratique le système formel équivalent à l'algorithme qui serait à l'œuvre dans leurs cerveaux aussi bien organique qu'électronique. Cette conclusion va dans le même sens que les objections qui ont été

formulées contre Lucas et qui laisse croire que la complexité de notre algorithme est insurmontable.

La riposte de Lucas est directe : comment peut-on être plus assuré de l'ignorance de notre algorithme si ce n'est en n'*étant* pas un algorithme? Cette riposte assume implicitement que l'ignorance n'est aucunement garantie et qu'il existe une méthode effective permettant de savoir si un algorithme est bien le nôtre. Or il s'avère que les objections formelles qui ne pouvaient être formulées dans le cadre informel de l'IA faible sont maintenant pertinentes. L'une d'entre elles montre qu'il n'existe pas de méthode effective (en admettant la thèse de Church) qui permet de savoir si un système formel est bien le nôtre. Encore une fois, le fait de ne jamais considérer les implications complètes de l'hypothèse que l'esprit est une machine empêche Lucas de bien évaluer la conclusion de Benacerraf.

Une tentative récente et exhaustive menée par Roger Penrose récupère les efforts de Lucas. Mais nous avons vu que les considérations nouvelles qu'il met de l'avant n'ont que des conséquences marginales. Bien qu'il soit tout à fait conscient de la distinction entre l'IA forte et l'IA faible, Penrose donne l'impression que son platonisme en mathématiques l'empêche de bien saisir les implications de chaque version. Il riposte (à son insu, semble-t-il) à l'argument de Benacerraf de la même manière que Lucas, négligeant l'objection formelle que nous venons de rappeler et qu'il connaît, mais qu'il met plutôt à contribution dans un contexte d'IA faible. Aussi, il explique comment un algorithme d'une machine pensante pourrait être conçu, mais cela a pour but de montrer qu'en principe cet algorithme doit être connaissable et considéré sûr. Il s'agit toujours de l'opposition principe/pratique, et les objections restent les mêmes. Enfin, nous réalisons que l'argument de Penrose/Lucas dans le cadre de l'IA faible ne peut se permettre de postuler que les capacités des machines pensantes pourront excéder celles des humains sous peine de ne plus pouvoir juger de la sûreté des machines.

En terminant, nous réalisons que la version forte du Mécanisme, loin d'être réfutée par les arguments précédents, semble plus plausible que jamais. S'il existe une limitation imposée par le théorème de Gödel et partagée par la machine pensante et l'esprit humain, elle concerne une certaine ignorance inaliénable de notre nature algorithmique profonde, ce qui correspond bien à l'impression que l'on peut avoir en

constatant le nombre et la complexité du fonctionnement des neurones de notre cerveau humain, siège matériel de notre esprit.

Bibliographie

Benacerraf, Paul, 1967, « God, The Devil and Gödel », *The Monist*, 51, pp. 9-32

Boolos, George, 1990, « On Seeing the Truth of the Gödel Sentence », *Behavioral and Brain Sciences*, 13 (4), pp. 655-656 (Dans ce numéro de *Behavioral and Brain Sciences*, on trouve une série d'articles commentant le livre *The Emperor's New Mind* de Penrose.)

Chalmers, David J., 1995, « Minds, Machines, And Mathematics », *Psyche*, vol. 2, no 9, http://psyche.cs.monash.edu.au/v2/psyche-2-09-chalmers.html

Chihara, Charles S., 1972, « On Alleged Refutations of Mechanism Using Gödel's Incompleteness Results », *The Journal of Philosophy*, 69 (17), pp. 507-526

Eisenstadt, Stuart A. et Simon, Herbert A., 1998, « Human and Machine Interpretation of Expression in Formal Systems » *Synthese*, 116 (3), pp. 439-461

Feferman, Solomon, 1962, « Transfinite Recursive Progressions of Axiomatic Theories », *Journal of Symbolic Logic*, vol. 27, no 3, pp. 259-316 (Cet article poursuit le travail entrepris par Turing 1939)

Feferman, Solomon, 1996, « Penrose's Gödelian Argument », *Psyche*, vol. 2, no 9, http://psyche.cs.monash.edu.au/v2/psyche-2-07-feferman.html

Gentzen, Gerhard, 1969, *The Collected Papers of Gerhard Gentzen*, Szabo M. E. (éd.), Amsterdam, North-Holland Publ., 338 p.

Herken, Rolf, 1988, *The Universal Turing Machine : a Half-century Survey*, Oxford, Oxford University Press, 681 p.

Hofstadter, Douglas, 1985, *Gödel, Escher, Bach. Les brins d'une guirlande éternelle*, Paris, InterEditions, 884 p.

King, David, 1996, « Is the Human Mind a Turing Machine », *Synthese*, 108, no 3, pp. 379-389

Kripke, Saul A., 1982, *Wittgenstein on Rules and Private Language*, Cambridge, Harvard Universitty Press, 150 p.

LaForte, Geoffrey, Hayes, Patrick J., et Ford, Kenneth M., 1998 « Why Gödel's theorem cannot refute computationalism », *Artificial Intelligence*, 104, pp. 265-86

Lucas, John Randolf., 1961, « Minds, Machines and Gödel », *Philosophy*, XXXVI, pp.112-127; réimprimé dans Anderson, Alan R. (éd.), *Minds and Machine*, Englewood Cliffs, Prentice-Hall, 1964, pp. 43-59 (Les références renvoient à la réimpression.)

Lucas, J. R., 1968, « Satan Stultified : A Rejoinder to Paul Benacerraf », *The Monist*, 52 (1), pp. 145-58

Lucas, J. R., 1970, *The Freedom of the Will*, Oxford, Clarendon Press, 181 p.

Maudlin, Tim, 1996, « Between the Motion and the Act... », *Psyche*, vol. 2, no 9, http://psyche.cs.monash.edu.au/v2/psyche-2-02-maudlin.html

McCullough, Daryl, 1996, « Can Humans escape Gödel », *Psyche*, vol. 2, no 9, http://psyche.cs.monash.edu.au/v2/psyche-2-04-mccullough.html

Mendelson, Elliott, 1997, *Introduction to Mathematical Logic*, New York, Chapman & Hall, 4ième édition, 440 p.

Nagel, E., J. Newman, K. Gödel et J.-Y. Girard, 1989, *Le théorème de Gödel*, Paris, Édition du Seuil, 185 p.

Penrose, Roger, 1989, *The Emperor's New Mind*, New York, Oxford University Press, 466 p.

Penrose, Roger, 1995, *Les ombres de l'esprit. À la recherche d'une science de la conscience*, traduit par Christian Jeanmougin, Paris, InterEditions, 461 p.

Penrose, Roger, 1996, « Beyond the Doubting of a Shadow », *Psyche*, vol. 2, no 9, http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html

Putnam, Hilary, 1964, « Minds and Machines », in Anderson, Alan R. (éd.), *Minds and Machine*, Englewood Cliffs, Prentice-Hall, pp.72-97. Publication originale dans Sidney Hook (éd.), *Dimensions of the Mind : A Symposium*, New York, New York University Press, 1960

Shanker, S.G. (Ed), 1988, *Gödel's Theorem in Focus*, London and New York, Routledge, 261 p.

Smullyan, Raymond M., 1992, *Gödel's Incompleteness Theorems*, New York, Oxford University Press, 139 p.

Tieszen, Richard, 1994, « Mathematical Realism and Gödel's Incompleteness Theorems », *Philosophia Mathematica*, (3) Vol. 2, pp. 177-201

Turing, Alan, 1939, « Systems of Logics Based on Ordinals », *Proceedings of the London Mathematical Society*, 45, pp. 544-546

Turing, Alan, 1950, « Computing Machinery and Intelligence », *Mind*, Vol. LIX, No 236, pp.433-460 réimprimé dans Anderson, Alan R. (éd.), *Minds and Machine*, Englewood Cliffs, Prentice-Hall, 1964, pp. 4-30

Wang, Hao, 1974, *From Mathematics to Philosophy*, New York, Humanities Press, 418 p.

Wang, Hao, 1990, *Kurt Gödel*, Paris, Colin, 337 p.

Wang, Hao, 1996, *A Logical Journey : From Gödel to Philosophy*, Cambridge (Mass), The MIT Press, 391 p.

Webb, Judson, 1968, « Metamathematics and the Philosophy of Mind », *Philosophy of Science*, 35 (2), pp. 156-178

Whiteley, C. H., 1962 « Minds, Machines and Gödel : A Reply to Mr Lucas », *Philosophy*, 37, pp. 61-62