NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, tests publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30.

Canada

L'intelligence artificielle : reproduction ou simulation de l'esprit?

THESE DE MAITRISE

Département de Philosophie,
Université d'Ottawa.

Jean-Noël Ringuet.
Juillet 1987.

ISBN 0-315-46848-3

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

AVANT-PROPOS

De nombreux livres et articles de revue ont été publiés ces dernières années sur l'**intelligence artificielle**, ce récent produit des techniques informatiques. Les fabricants d'ordinateurs et les concepteurs de logiciels vantent de plus en plus fréquemment l'*intelligence* de leurs produits, le Japon, serré de près par les Etats-Unis et l'Europe, s'est engagé dans un ensemble de recherches ayant pour objet la construction d'ordinateurs capables de *comprendre*, et les futurologues nous invitent à nous préparer à la vie dans un *environnement intelligent*.

L'absence de définition des termes utilisés, quand ce n'est pas l'usage délibéré de notions équivoques, a pour effet de semer une confusion importante sur les réalisations actuelles et possibles de la technologie informatique. La méprise, à cause de sa rentabilité, est encore malheureusement entretenue par l'avidité de certains chercheurs en intelligence artificielle profitant des rêves souvent naïfs des politiciens et militaires derrière la très notoire Initiative de Défense Stratégique (*Star Wars*) aux Etats-Unis.

On connaît moins les interventions de philosophes sur le sujet qui ont favorisé, sinon un consensus sur le sens des termes, du moins une plus grande prudence dans leur application à ce champ de recherche relativement nouveau. Cette thèse a pour principal objectif de donner une vue d'ensemble des contributions les plus marquantes de la philosophie sur la question de l'éventualité, au moyen d'ordinateurs, d'une "intelligence" synthétique.

La plupart des documents recensés sont de langue anglaise. A des fins d'harmonie, nous avons traduit en français les citations qui apparaissent dans le texte. Le lecteur pourra se référer à l'appendice pour le texte original des traductions dont nous sommes responsable.

Nos remerciements les plus sincères au professeur Jerzy A. Wojciechowski du département de philosophie de l'Université d'Ottawa qui a bien voulu diriger cette thèse, ainsi qu'à madame Christiane Désilets qui a participé à la revision de nos textes et nous a assuré d'un support constant tout au long de cette recherche.

TABLE DES MATIERES

INTRODUCTION

Chaque pas en avant de la technologie entraîne une redéfinition des rapports de l'espèce humaine avec la nature, tout en modifiant l'image que celle-ci a d'elle-même. C'est par l'entremise de ses outils que l'être humain se construit une image non seulement physique, mais métaphysique de son univers et de la place qu'il y occupe.

L'apparition de la machine à vapeur, tout en bouleversant l'organisation du travail, a permis l'éclosion d'idéologies conquérantes fortement marquées par l'anthropocentrisme: l'homme devenait maître d'un destin régi auparavant par les caprices de divinités orgueilleuses et susceptibles. L'épanouissement des sciences rendait possible le décodage non seulement des lois qui régissent la nature physique, mais aussi (en partie, du moins) de celles de la vie et permettait même de jeter un éclairage nouveau et troublant sur le monde jusque-là obscur de la psyché.

Certains n'hésitent pas à voir dans l'expansion extrêmement rapide de l'informatique un événement comparable à cette Révolution Industrielle qui, en quelques siècles, a modifié complètement le visage de la planète; car il ne s'agit plus uniquement de mettre au point des outils qui prolongent les membres humains, mais d'utiliser le pouvoir

des machines pour prolonger et amplifier l'**intelligence.** Tel Pygmalion qui épousa sa création, Galatée, l'ordinateur représente l'intégration ultime entre l'artisan et l'artéfact. Ainsi, il est de plus en plus courant d'entendre parler du cerveau en psychologie comme d'un "processeur d'information" où l'on étudie des choses comme "l'algorithme de stockage de la mémoire à long terme"; d'un autre côté, des chercheurs en intelligence artificielle se font fort d'avoir mis au point des programmes qui rendent l'ordinateur capable de "percevoir, comprendre, apprendre, parler, etc." comme l'être humain.

Ces développements sont soutenus par un nouveau type de discours qui concerne directement la philosophie non seulement à cause de ses conséquences éthiques, mais par ses postulats épistémologiques: dans quelle mesure peut-on parler d'**intelligence** de la machine? Peut-on dire d'un ordinateur qui exécute une série d'opérations logiques comparables à celles réalisées par un esprit humain qu'il **pense**?

En reproduisant dans la machine la capacité de calculer, de raisonner, d'apprendre, l'être humain a l'enivrante impression de se recréer lui-même: peut-être sommes-nous en train de préparer une prochaine étape de l'évolution, ou l'acier remplacera la chair, les micro-circuits le système nerveux, la machine l'homme...

Les questions fusent de toute part, les réponses aussi, certaines prudentes, d'autres extrêmement audacieuses. Le débat, qui met aux prises psychologues, informaticiens, linguistes, neurophysiologues, philosophes n'est pas toujours dénué d'émotivité et de subjectivité, comme toute question d'ailleurs qui met en jeu des valeurs et des conceptions relatives à l'être humain et à sa finalité. Néammoins, il permet de lever progressivement l'équivoque dans l'usage de notions qui jusqu'ici n'étaient réservées qu'à l'esprit humain.

Au chapitre premier, nous ferons voir l'émergence des premières interrogations soulevées par l'apparition des ordinateurs ainsi que la pertinence, pour la philosophie, d'une réflexion particulière sur l'intelligence artificielle. Le chapitre II brossera un tableau des origines de cette nouvelle discipline, des espoirs et des échecs auxquels elle a donné lieu, ainsi que de ses perspectives à moyen et long terme. Les troisième et quatrième chapitres traiteront de façon toute particulière des divers aspects du débat philosophique sur la question de l'intelligence des machines: au chapitre III, nous aborderons plus spécialement certains problèmes soulevés par les limites des systèmes logiques (selon le théorème de Gödel), et nous présenterons une synthèse de la polémique lancée par John R. Searle, philosophe américain, sur la capacité d'**intentionnalité**

inhérente à la nature biologique de l'esprit humain: le quatrième chapitre sera consacré aux thèses du philosophe américain Hubert L. Dreyfus dont la perspective résolument phénoménologique a causé de sérieux remous dans les milieux de l'intelligence artificielle aux Etats-Unis. En guise de conclusion, le dernier chapitre proposera une appréciation personnelle des présupposés à caractère philosophique qui sous-tendent les projets de réalisation d'une **intelligence de synthèse.**

CHAPITRE I

Philosophie et Intelligence Artificielle

Les premières interrogations

Beaucoup d'attention a été portée depuis quelques années au développement de l'informatique, ainsi qu'à ses implications économiques, sociales, et scientifiques. Bien que cette technologie existe depuis bientôt une quarantaine d'années, elle n'a longtemps été accessible qu'aux milieux d'affaires ou scientifiques et à peu près hors de portée du commun des mortels. Au cours des années '70, la micro-informatique, sans toutefois bouleverser les concepts de base de l'ordinateur, en a généralisé l'usage: tant la rapidité de calcul que la polyvalence de ces machines a suscité à la fois émerveillement et inquiétude.

On dit en effet de l'ordinateur que c'est une machine "universelle" qui peut calculer, classer des informations, les comparer, exécuter des simulations dont certaines présentent beaucoup d'analogie avec des propriétés de l'esprit humain comme traduire un texte, tenir une conversation limitée avec un interlocuteur, battre à leur propre jeu même d'excellents joueurs d'échecs, emmagasiner et utiliser

le savoir d'experts en différents domaines, composer une pièce musicale, un poème, etc. Ces performances spectaculaires n'ont pas manqué de provoquer des réactions parfois marquées d'un certain animisme: certains ont vu dans l'ordinateur la réalisation du vieux rêve prométhéen de création par l'homme d'une entité à son image; d'autres se sont inquiétés de la place de l'humain dans un environnement "intelligent" où la machine assumerait la plupart des travaux jusqu'ici spécifiques à notre espèce. Le "complexe de Frankenstein", i.e. la crainte du créateur de voir sa créature se retourner contre lui, a été considérablement exploité dans la littérature de science-fiction[1] traitant de robots et de super-ordinateurs, alors que certains prophètisent, au contraire, un nouvel âge d'or. On a même vu la prestigieuse revue TIME proclamer le micro-ordinateur Homme de l'année 1982.

Ce mélange de crainte et d'enthousiasme à l'égard de l'ordinateur n'est toutefois pas apparu avec la micro-informatique. Malgré l'allure monstrueuse de leurs machines composées de centaines de kilomètres de câbles, de milliers de relais électro-mécaniques et de lampes à vide, les pionniers partagèrent les mêmes sentiments devant les

---

[1] Et au cinéma: dans le célèbre film de Kubrick, 2001: l'Odyssée de l'Espace, le robot HAL se révolte contre son équipage humain.

prouesses de leurs créatures; par exemple, John von Neumann, auteur de la théorie des jeux et père de l'EDVAC (Electronic Discret Variable Computer) en 1945, s'interrogeait[2] sur les possibilités de l'ordinateur de reproduire les fonctions du cerveau humain. A la même époque, Norbert Wiener, dont la théorie "cybernétique" mettait en évidence l'analogie des mécanismes de transmission d'information chez les machines et les humains, craignait que ceux-ci n'abdiquent leur pouvoir de décision au profit de celles-là[3]; il s'inquiétait du flou de la frontière entre ces deux entités et affirmait même que si un jour les humains parvenaient à créer un esprit artificiel, ils ne réussiraient pas mieux que Dieu ne l'a fait avec l'homme et pourraient fort bien connaître les déboires du créateur légendaire du Golem[4].

---

[2] Sans véritablement y croire, toutefois. Voir J. von Neumann, The Computer and the Brain. Yale University Press, N.H., Conn., 1958.

[3] N. Wiener, The Human Use of Human Beings, Houghton Mifflin, Boston, 1958. Voir aussi God and Golem, Inc., The MIT Press, Cambridge, Mass., 1964.

[4] Légende de la tradition juive fréquemment citée dans la littérature portant sur l'intelligence artificielle: un rabbin vivant à Prague au XVIème siècle aurait créé une statue d'argile qui s'est animée lorsque ce dernier lui inscrivit le nom de Dieu sur le front. Utilisée comme espion pour prévenir les juifs de Prague des pogroms avant que ceux-ci ne surviennent, la créature, appelée Joseph Golem, devint de plus en plus violente pour finalement se tourner contre son auteur. Ce dernier effaça alors le nom de Dieu du front du Golen, qui se désintégra aussitôt en poussière.

Bref, très tôt dans l'histoire de l'informatique surgissent les comparaisons entre l'esprit humain et la machine. Vers la fin des années cinquante, certains chercheurs[5] font porter leurs efforts vers la réalisation de programmes reproduisant des comportements intelligents. Les premières tentatives ont lieu dans le domaine des jeux et de la traduction: ce champ nouveau d'investigation de l'informatique se voit baptiser bientôt du nom un peu provocateur d'**Intelligence Artificielle (I.A.)**. Les progrès réalisés sont suffisamment spectaculaires pour que soit sérieusement soulevée, ailleurs que dans le domaine de la fiction, la question: **"Les machines peuvent-elles penser?"**: interrogation qui rejoint naturellement la philosophie et sa recherche sur la nature de l'esprit. Il faudra cependant attendre les années soixante avant que des philosophes reprennent à leur compte cette problématique.

Philosophie et outils "intelligents"

La réflexion philosophique sur la technologie n'est pas nouvelle, bien que depuis Platon la philosophie se soit généralement peu souciée de l'évolution des savoirs

---

[5] Les mieux connus étant Claude E. Shannon, Allan Newell, H.A. Simon, Anthony Oettinger. Voir l'historique de l'IA au chapitre II.

pratiques; ce fait s'explique en bonne partie par le peu d'égard, sinon le mépris ouvert, des classes dominantes (dont étaient issus la plupart des philosophes) à l'égard du labeur manuel. Il aura fallu beaucoup de temps avant que la philosophie n'intègre pleinement à sa réflexion l'**outil**, cet objet qui pourtant dès l'aube de l'humanité a permis à l'homme non seulement de dominer son environnement mais d'extérioriser, dans la pierre et sur les parois de cavernes, son imaginaire.

Les outils comme extensions fonctionnelles des membres humains remontent à bien longtemps avant l'arrivée d'*Homo Sapiens*. Toutefois, tout indique que ce n'est qu'avec notre espèce que l'outil s'est agrémenté de motifs artistiques témoignant des représentations intellectuelles de leurs auteurs. A partir des premières sculptures et peintures rupestres, une part croissante de la compétence technique de l'être humain est consacrée à l'extériorisation de son activité imaginaire. Cependant, l'évolution de l'outil comme support et prolongement d'une activité intellectuelle plus abstraite ne connaît son élan qu'avec l'apparition de l'agriculture et de l'élevage il y a environ une dizaine de milliers d'années, et sa progression est rigoureusement parallèle au développement de la technologie "pratique". On fait usage d'abord des calculi (petits cailloux utilisés pour

compter) qui accélèrent considérablement la capacité de calcul arithmétique puis l'invention de l'écriture symbolique vient procurer un support permanent à la mémoire: il y a évidemment une relation entre ces innovations et les besoins d'une communauté où l'accumulation a multiplié les échanges et rendu la tenue de comptes nécessaire. Peu après, on assiste littéralement à l'explosion de la production d'outils spécialisés dans la transmission et le traitement de l'information: la fonction de l'écriture, d'abord réduite à la comptabilité, s'étend à la communication des idées. Puis apparaissent le boulier et divers outils de mesure du temps comme de l'espace, tels le cadran solaire, la boussole, le sablier, le compas, la règle à calcul... Enfin, l'invention de l'imprimerie au XVème siècle amplifie de façon décisive la médiatisation de l'activité intellectuelle.

Les développements technologiques de l'Antiquité et du Moyen-Age n'ont pas été le produit de sciences théoriques, ce qui explique sans doute en bonne partie l'intérêt médiocre des philosophes du temps à leur égard. Ce n'est que lorsque la technologie développa un objet comparable dans ses mouvements et sa capacité de précision à la représentation Ptoléméenne des sphères célestes que des philosophes baissèrent les yeux: enfin, ciel et terre se rejoignaient et

la mise au point de l'horloge[6] rendait accessible un pouvoir divin, celui de reproduire l'ordre et l'harmonie, jusque là réservé au monde extra-terrestre et à l'abstraction mathématique. Ce fut le triomphe du mécanisme, auquel peu de philosophes de la Renaissance restèrent indifférents: le jeune Pascal réalisait la première machine à additionner, Descartes comparait tous les phénomènes de la nature, y compris ceux des corps vivants, aux délicats engrenages de l'horloge (conception reprise et élargie à l'esprit humain par La Mettrie dans L'Homme-Machine) et Leibniz faisait appel aux pendules dans une métaphore célèbre pour rendre compte de l'harmonie du corps et de l'esprit.

C'est ainsi que le XVIIIème siècle vit la mise au point des premières machines automatiques qui non seulement prolongeaient les membres humains dans l'exécution d'un travail, mais surtout incorporaient, dans leur conception même, du **savoir** sur la séquence des opérations à réaliser: objets d'amusement de l'aristocratie, comme les fameux automates de Vaucanson ou précurseurs de la société industrielle, tel le métier à tisser de Jacquard. Mais on atteint vite les limites de la mécanique: la réalisation d'outils capables de

---

[6] Sur l'importance de l'invention de l'horloge pour la pensée humaine, voir J. Daniel Bolter, Turing's Man: Western Culture in the Computer Age, The University of North Carolina Press, Chapell Hill, 1985, chap. 2.

traitement d'information est nécessairément complexe, et la multiplicité des pièces mobiles rendait aléatoire la transmission de commandes d'une partie de la machine à l'autre. Charles Babbage en est l'illustration, lui qui dès la première moitié du XIXème siècle établit les plans du tout premier ordinateur, entièrement mécanique: la "machine analytique", avec processeur, mémoire, programmation et cartes perforées. Elle aurait fonctionné si on avait pu la construire, mais la fabrication et l'assemblage de ses 50,000 pièces représentaient un défi insurmontable pour la technologie de l'époque.

Au moment où le progrès avait désormais réconcilié les sciences théoriques et la technologie, les philosophes du XIXème siècle délaissèrent les machines et se préoccupèrent surtout des conséquences sociales et politiques de l'explosion des sciences et des techniques (Proudhon, Fourier, Marx, etc.). Ce n'est qu'avec le développement de la technologie électronique, et la création de machines capables non seulement de calculer, mais de manipuler des symboles et de réaliser des opérations logiques, que technologie et philosophie se rencontrèrent à nouveau. Les processus mécaniques permettant la réalisation de calculs arithmétiques, si ingénieux aient-ils été, étaient perçus à l'ère pré-électronique plutôt comme des artifices et nul ne

songeait à établir un parallèle sérieux entre ces bruyants assemblages métalliques et l'esprit humain. Mentionnons toutefois que Lady Lovelace, disciple et collaboratrice de Charles Babbage, observa que la "machine analytique" pourrait non seulement solutionner des problèmes mathématiques, mais à l'instar du cerveau humain, manipuler des symboles[7]. Cependant, dès la mise au point des premiers ordinateurs, l'analogie avec l'intellect s'imposa avec la même force que l'évocation des mouvements célestes suite à l'invention de l'horloge, suscitant la formulation d'une nouvelle problématique philosophique.

"Les machines peuvent-elles penser?"

Les paramètres du débat actuel furent vraiment posés par Alan Mathison Turing[8], dans un texte aujourd'hui considéré

---

[7] "[The Analytical Engine] can arrange and combine its numerical quantities exactly as if they were letters or other general symbols; and in fact it might bring out its results in algebraical notation, were provisions made accordingly." Lady A. Lovelace, Notes upon the Memoir "Sketch of the Analytical Engine Invented by Charles Babbage" by L.F. Menabrea, (Genèves, 1842), cité dans Bolter, The Turing's Man:..., p. 66.

[8] Mathématicien et logicien anglais (1912-1954), créateur d'une machine à décrypter les messages allemands lors de la dernière guerre, qui développa en 1936 (de façon purement théorique) le schéma fonctionnel d'une machine universelle - dite **machine de Turing** - qui servira de modèle aux ordinateurs actuels.

comme un classique[9] de l'I.A.. L'auteur y envisage d'abord la formulation "Les machines peuvent-elles penser?", pour conclure que la difficulté de définir le sens des mots "machine" et "penser" en fait une mauvaise question. Il suggère plutôt l'approche suivante:

> "Ce nouveau type de problème peut être présenté sous forme d'un jeu, que nous appelerons "jeu d'imitation". Il nécessite trois joueurs, un homme (A), une femme (B), et un troisième joueur qui posera les questions (C), et qui peut appartenir à l'un ou l'autre sexe. Ce dernier restera dans une pièce séparée de celle où se trouve le couple. Le but du jeu, pour ce troisième joueur, est de déterminer, chez les deux autres joueurs, qui est l'homme et qui est la femme. Il les connaît sous le nom de X et Y, et doit dire, au terme du jeu, soit "X est A et Y est B", soit "X est B et Y est A". Pour ce faire, il a droit à des questions du type: "X peut-il me dire de quelle longueur sont ses cheveux?"
>
> Supposons que X est en fait A, c'est à X de répondre. La règle du jeu, pour A, c'est de tenter d'induire C en erreur. Par conséquent, il pourra répondre, par exemple:
>
> "J'ai les cheveux coupés au carré, les plus longues mèches atteignent environ vingt-deux centimètres."
>
> Pour éviter, naturellement, que les voix n'aident C, les réponses devront être écrites, si possible, à la machine. Le mieux serait un téléimprimeur entre les deux pièces; ou encore, autre solution, les questions et les réponses peuvent être transmises par un intermédiaire. Le but du jeu pour la joueuse (B) est d'aider le joueur (C). La meilleure tactique en ce qui la concerne, est peut-être encore de ne donner que des

---

[9] Computing Machinery and Intelligence, Mind, Vol. LIX, No. 236, 1950, reproduit dans Daniel C. Dennett et Douglas R. Hofstadter, The Mind's I: Fantasies and Reflections on Self and Soul, Bantam Books, N.Y., 1982, chap. 4.

> réponses justes. Elle peut ajouter des déclarations telles que "C'est moi la femme, ne l'écoutez pas!", mais comme le joueur A peut tout aussi bien émettre pareil discours, cela n'avancera pas à grand-chose.
>
> Et nous posons maintenant cette question: "Que se passera-t-il si l'on fait tenir par une machine le rôle de (A) dans ce jeu?" Le joueur (C), en pareil cas, échouera-t-il aussi souvent que lorsque le jeu est joué avec un homme et une femme pour partenaires? Ces questions prennent la place de notre interrogation première, "Ces machines sont-elles capables de penser [10]?"

Une machine capable de se prêter à ce jeu est réalisable, conclut Turing, et vers l'an 2000 les ordinateurs seront suffisamment perfectionnés pour qu'un interrogateur n'ait pas plus de soixante-dix pour cent de chances de procéder à une bonne identification en cinq minutes de dialogue; alors plus personne, espère-t-il, ne s'étonnera du fait qu'on parle de "machines pensantes".

Ce test - bien connu sous le nom du **test de Turing** - devait relancer, sur de nouvelles bases toutefois, un débat presqu'aussi vieux que la philosophie elle-même sur la nature de l'intelligence: la pensée existe-t-elle indépendamment de ses manifestations extérieures?

---

[10] Turing, loc. cit., p. 7, (cité dans Hubert L. Dreyfus, Intelligence Artificielle: mythes et limites. Paris, Flammarion, 1984, p. 10. (Traduction de la deuxième édition américaine de What Computers Can't Do: the limits of artificial intelligence. New York, Harper & Row Publishers, 1979)

Mais par quelles particularités ce nouveau genre de machines évoque-t-il aussi fortement l'esprit humain? Il ne peut s'agir de leur configuration matérielle (externe ou interne): car sauf lors de quelques expériences en cybernétique dans les années '50[11], il n'y a pas de ressemblances vagues ou même lointaines entre les agencements de circuits et de transistors sur un microprocesseur et la disposition des neurones du cerveau. Quant à leur mode de fonctionnement, malgré certaines similitudes entre le processus électro-chimique des neurones et celui électronique des microprocesseurs, l'incroyable complexité des premiers et leur capacité de traiter "en parallèle" (simultanément et de façon intégrée) des millions de bits d'information suggère plus de différences que de similitudes avec ces derniers. Sans compter que la conception des ordinateurs modernes utilisés par l'intelligence artificielle n'exige pas de ses auteurs de connaissances spéciales sur la physiologie du cerveau.

La situation est cependant différente du côté des processus logiques (ou programmes) grâce auxquels fonctionnent les ordinateurs. Bien qu'on ne connaisse qu'approximativement certains processus du cerveau humain, on ne peut nier la ressemblance du cheminement suivi par les

---

[11] Voir chapitre 2, p. 34, le PERCEPTRON.

machines dans certains types de travaux avec celui de l'esprit humain dans les mêmes circonstances. Même dans les cas où les processus sont évidemment très différents - nous verrons au prochain chapitre le cas du jeu d'échecs - les séquences d'opérations ont un caractère logique dans lequel tout être humain normal se reconnaît puisque, ne l'oublions pas, ces programmes sont le fruit d'une intelligence humaine. Enfin, et surtout, il y a la performance fort particulière des ordinateurs dans une infinité de domaines intellectuels qui se traduit par la production d'objets qui seraient considérés comme étant le résultat d'une démarche intelligente s'ils étaient le produit d'êtres humains.

Mais est-ce suffisant pour qu'on prête à une machine ces propriétés liées à la pensée humaine qu'on appelle la conscience, la subjectivité, l'intentionnalité? Ces états psychiques auxquels prétend chaque humain normal et que nous nous concédons généralement volontiers les uns les autres ont la fâcheuse caractéristique de ne se manifester qu'au coeur de l'expérience subjective. Ainsi, même si un jour une quelconque machine électronique s'en réclamait, qu'est-ce qui nous permettrait de distinguer vraiment les réactions d'un ordinateur, programmé de façon très sophistiquée et pourvu d'une base de connaissances analogue à celle d'un être humain moyen, des réponses d'une entité consciente, dotée de

liberté, capable de projets? Bref, comment reconnaître une machine qui **simule** de celle qui **reproduit** non seulement l'aspect externe de la pensée, mais ses états internes?

La définition qu'on donne de la "pensée" est donc d'une extrême importance dans la façon d'envisager les réalisations de l'I.A.. Or, en existe-t-il une définition universellement acceptable? Selon le philosophe américain Mortimer J. Adler, qui s'interroge sur la pertinence de la question "Une machine peut-elle penser?":

> "La littérature concernant les ordinateurs et les robots considère cette question comme tellement vague et ambiguë qu'il ne peut y avoir de bonne réponse. L'usage que font du mot "penser" les psychologues, les neurologues, les informaticiens et les philosophes est tellement varié selon qu'on l'applique à des animaux, à des humains ou à des machines que si on demandait "Les animaux peuvent-ils penser?" ou "Les machines peuvent-elles penser?", la réponse serait OUI dans certains sens du mot, et NON dans d'autres. Personne n'a pu produire une définition acceptable de la pensée humaine qui pourrait rendre compte de toute sa diversité et servir de norme pour mesurer l'aptitude d'un robot à simuler l'éventail entier des activités de l'esprit humain[12]."

En conséquence, Adler propose de circonscrire le test de Turing à **une** performance bien spécifique à l'esprit humain,

---

[12] MORTIMER J. ADLER, "The Challenge to the Computer.", Proceedings of the American Catholic Association XLII, 1968, p. 24. (Extrait du volume The Difference of Man and the Difference It Makes, N.Y., Holt, Rinehart and Winston, Inc., 1967, chap. 14 "From Descartes to Turing".)

soit la capacité de converser couramment en langue naturelle; en effet, le langage, qui d'après Descartes[12] distingue l'homme de l'animal, démontre sans équivoque la présence de **pensée conceptuelle**. Donc, selon Adler, une machine qui relèverait avec succès ce qu'il appelle le "défi cartésien", même si elle échouait dans d'autres domaines, mériterait le qualificatif de "pensante".

Mais jusqu'à quel point la pensée correspond-elle à l'envers du langage? Cela signifie-t-il que tout ce qui ne parle pas est dépourvu de vie psychique digne de ce nom? A ce sujet, le défi cartésien qu'envisage Adler comporte encore une certaine ambiguité. Or, il s'agit du coeur de la problématique philosophique concernant l'intelligence artificielle, et pour qu'un progrès appréciable soit accompli sur le sujet, on doit attendre de la réflexion théorique de l'I.A. un certain nombre de précisions sur les difficultés qui viennent d'être évoquées. Par exemple, comment l'I.A. définit-elle l'intelligence? comment identifier, derrière un comportement apparemment intelligent, une **présence psychique**? et enfin, qu'est-ce qui peut permettre de distinguer une **simulation** d'une **reproduction** de la pensée humaine?

---

[12] Discours de la Méthode, 5ème partie, *in fine*.

Pertinence d'une réflexion philosophique sur l'I.A.

Tous les chercheurs en I.A. ne reconnaissent pas la pertinence de questions philosophiques à propos de leurs activités. Pour la tendance technicienne de l'I.A., la machine "intelligente" sera celle qui pourra faire croire à tout le monde en tout temps qu'elle l'est et l'I.A. doit consacrer l'essentiel de ses énergies au développement des techniques qui permettront d'atteindre cet objectif. Mais le pragmatisme d'une telle position ne liquide pas ses implications philosophiques sous-jacentes: en effet, comment déterminer que la réponse d'une machine est **intelligente**? Parce qu'elle correspond à un comportement "normal" chez l'humain? Parce qu'elle s'exprime symboliquement? Parce qu'elle manifeste de la sensibilité? Parce qu'elle est capable d'apprendre? Au fait, c'est quoi, l'intelligence? Bref, toute référence à ce concept renvoie l'I.A. "pratique" à la question fondamentale de la philosophie. Comment une expérience quelconque conduit à une représentation, comment elle est organisée et emmagasinée, en quoi elle correspond au "monde extérieur", sont des questions parfaitement semblables à celles de la philosophie depuis ses origines. La tentative la plus humble de reproduire sur une machine un processus de pensée s'appuie nécessairement sur une certaine conception de l'esprit humain et de ses opérations; le nom même

d'Intelligence Artificielle implique une conception épistémologique qui mérite d'être élucidée dans l'intérêt même de la science que cette discipline souhaite constituer.

La philosophie est la seule science, avec la psychologie, qui a développé des théories générales sur l'esprit et la pensée; l'I.A. est un développement plus récent dans ce même effort de décrire ce qui se passe dans notre esprit, une sorte d'épistémologie des temps modernes, à cette différence près, comme aime le rappeler Roger Schank[14], qu'elle vérifie ses hypothèses sur l'ordinateur... La plupart des chercheurs en I.A. reconnaissent que pour rendre leurs machines plus intelligentes, ils ont besoin d'une théorie générale du fonctionnement de l'intelligence chez les humains. D'où l'émergence, particulièrement aux Etats-Unis, des **sciences cognitives**[15], disciplines au carrefour de la linguistique, de la psychologie, de la philosophie, de l'anthropologie, de la neurologie et de l'informatique. Le principe directeur de cet effort multidisciplinaire est à l'effet que l'esprit,

---

[14] Théoricien américain de l'IA, et expert en systèmes de reconnaissance de langage. Auteur, entre autres, de The Cognitive Computer: on Language, Learning and Artificial Intelligence. Addison-Wesley Publishing Co., 1985.

[15] Le principal théoricien de la science cognitive est Jerry Fodor, auteur de The Language of Thought, New York, Crowell, 1975.

comme l'ordinateur, constitue un système de manipulation de symboles dont il reste à découvrir le mode de fonctionnement.

Dans ce contexte, l'I.A. reprend à son compte les questions millénaires de la philosophie sur la nature de l'intelligence humaine: "Nous essayons de définir avec précision ce qui permet aux humains de comprendre, d'apprendre, de penser et de changer avec le temps[16]."

A quel champ de la philosophie devrait revenir le privilège de cette collaboration? Il est sans doute trop tôt pour distribuer de façon aussi précise des rôles hiérarchisés. Pour la philosophie des sciences, par exemple, la démarche très empirique de l'I.A., sans théorie générale ni mode de développement très structuré, ne lui permet peut-être pas encore de revendiquer ce titre de science; du côté de la psychologie philosophique, beaucoup d'intérêt a été porté à la mise en parallèle entre les réalisations de l'I.A. et les processus mentaux (particulièrement dans les domaines de la perception et du langage), et cet échange entre les deux disciplines semble avoir été jusqu'à maintenant des plus féconds; pour sa part, l'épistémologie philosophique se doit de vérifier si l'homme peut réussir à reproduire ce qui a

---

[16] Schank, The Cognitive Computer:..., *op. cit.*, p. 30.

traditionnellement été considéré comme son essence, ou si l'I.A. est victime dans sa conception même d'une gigantesque illusion dont la source serait une conception tragiquement étroite de l'esprit. Bref, comme le fait remarquer Daniel Andler[17], puisque l'I.A. constitue une pensée de la pensée, et qu'il n'y a pas d'autre discipline susceptible de penser l'I.A., le défi se pose à toutes les branches de la philosophie. Enfin, selon Hubert Dreyfus, un des premiers philosophes américains à s'intéresser à l'I.A., ce qu'il faut n'est rien de moins qu'une "critique de la raison artificielle":

> "Les ordinateurs ont déjà provoqué une révolution technologique comparable à la révolution industrielle [...] Si réellement nous sommes à la veille de créer une intelligence de synthèse, nous sommes près d'assister au triomphe d'une conception très particulière de la raison. De fait, si l'on peut effectivement douer de raison un ordinateur, voilà qui viendra confirmer cette conception de l'homme que depuis deux mille ans les penseurs de l'Occident cherchent vraiment à étayer, mais sans avoir pu, jusqu'à présent, la formuler ou la soumettre à l'expérimentation, faute de l'outil dont ils disposent désormais: l'homme perçu comme un objet. [...]
>
> Nous devons tenter de définir le champ d'action et les limites de cette forme de raison qui s'est déployée en force depuis l'essor et le perfectionnement de la "machine analytique". Nous devons essayer de comprendre jusqu'à quel point il peut exister une véritable intelligence artificielle et, s'il se trouve des limites à la possibilité de simuler sur ordinateur une conduite

---

[17] Daniel Andler, Avant-propos du livre de Hubert L. Dreyfus, Intelligence Artificielle: Mythes..., p. XIII.

intelligente, il nous faut déterminer ces limites et rechercher ce qu'elles signifient. Ce que nous apprendrons sur les limites de l'intelligence chez les ordinateurs nous fournira des enseignements sur la nature et le champ d'action de l'intelligence humaine[18]."

---

[18] Dreyfus, *op. cit.*, pp. 17-18.

- CHAPITRE II

Espoirs, difficultés et perspectives de l'I.A.

L'intelligence artificielle existe maintenant depuis plus de trente ans; son histoire est mouvementée, marquée de progrès extrêmement spectaculaires et rapides suivis de longues périodes de stagnation. Les espoirs de ses fondateurs se sont heurtés à des difficultés que l'état des connaissances des années cinquante sur la pensée humaine ne permettait pas de prévoir; ces échecs ont toutefois permis de mieux comprendre certains processus cognitifs chez l'être humain et même d'en reproduire quelques-uns sur ordinateur.

Ce chapitre traitera, après avoir proposé une définition de l'intelligence artificielle, de l'histoire récente de cette discipline, de ses ambitieux projets dans le domaine des jeux, de la compréhension du langage naturel, de la solution de problèmes et plus succinctement; de la reconnaissance de formes (vision); de cette étude des réalisations actuelles, nous dégagerons la problématique majeure de l'intelligence artificielle à ce jour, ainsi que les perspectives à moyen et long terme pour la création de

machines vraiment "intelligentes". Quant aux inévitables descriptions techniques des projets qui seront traités, nous n'en présenterons que l'essentiel nécessaire à la compréhension des enjeux philosophiques dont les prochains chapitres feront l'objet.

Définitions de l'intelligence artificielle

Bien qu'il n'y ait pas encore de définition universellement acceptée de l'intelligence artificielle, la plupart des chercheurs semblent s'entendre sur les deux conceptions suivantes: la première, plus restreinte, définit l'intelligence artificielle comme une discipline qui, à l'aide de techniques de programmation, a pour objectif de faire reproduire par l'ordinateur des comportements analogues à ceux de l'intelligence humaine (capacité d'apprentissage, de raisonnement, de compréhension du langage naturel, de résolution de problèmes, etc); la deuxième définition, plus théorique, considère l'intelligence artificielle comme l'étude, à partir de simulations sur ordinateur, de l'esprit humain. Ainsi, selon Seymour Papert,

> "Prise dans son sens le plus étroit, l'I.A. désigne la discipline qui se donne pour but d'accroître la capacité des machines à accomplir des performances que l'on considérerait comme marque d'intelligence si elles étaient le fait d'êtres humains. Son objectif étant de concevoir des machines, on pourrait dire qu'elle est une

> branche avancée de l'ingénierie. Seulement, pour mettre au point de telles machines, il faut d'ordinaire réfléchir non seulement sur la nature des machines, mais encore sur la nature des fonctions intelligentes que l'on veut leur voir remplir [...] Et c'est dans ce genre de recherches qu'il faut voir la plus large définition de l'intelligence artificielle: il s'agit en fait d'une science cognitive, une science qui s'intéresse aux sources du savoir.[1]"

Margaret Boden qui s'intéresse à l'intelligence artificielle du point de vue de la psychologie voit dans cette discipline "l'utilisation de programmes d'ordinateur et de techniques de programmation pour éclairer les principes de l'intelligence en général et de la pensée humaine en particulier"[2]. Marvin Minsky, pionnier de l'intelligence artificielle, affirme que "le scientifique qui oeuvre dans ce domaine appelé I.A. essaie de mettre au point des théories sur les connaissances spécifiques nécessaires à la résolution de problèmes et sur les processus de connaissance plus généraux utilisés dans la mise en oeuvre de ces connaissances spécifiques", partageant ainsi sés buts "avec la philosophie dans l'étude de problèmes touchant l'esprit, la pensée, la

---

[1] PAPERT, SEYMOUR, Jaillissement de l'esprit: ordinateurs et apprentissage. Flammarion, Paris, 1981, p.195. (Traduction de Mindstorms, Basic Books, N.Y., 1980.)

[2] BODEN, Margaret A., Artificial Intelligence and Natural Man. New York, Basic Books, 1977, p.5.

raison et les sentiments[3]". Enfin, Schank distingue dans l'intelligence artificielle l'approche technologique, axée sur le produit, de l'approche scientifique, dont le principal souci est de développer une théorie générale de l'esprit humain[4].

Pendant longtemps, toutefois, c'est la définition pratique qui a dominé dans le milieu de l'intelligence artificielle; même si encore aujourd'hui, chacun privilégie le sens correspondant le mieux aux perspectives dans lesquelles il se trouve engagé, on peut affirmer que de plus en plus la plupart des chercheurs de cette discipline reconnaissent et acceptent cette double définition. Les présupposés simplistes de certaines expériences de modélisation de l'intelligence et les déceptions encourues ont fait réaliser

---

[3] MINSKY, MARVIN L., "Computer Science and the Representation of Knowledge", The Computer Age: A Twenty-Year View, edited by Michael L. Dertouzos and Joel Moses, MIT Press, Cambridge, 1979, (pp.392-421), p. 400.

[4] "Today, AI researchers have started to ask questions about the real nature of human intelligence (...) This seemingly simple shift in approach actually makes a profound difference for the definition of AI and for the course of AI research. Rather than concentrating on a specific task, such as getting a computer to play chess, we are addressing the essential theoretical and philosophical questions of human intelligence. This attitude reflects the two basic approaches to Artificial Intelligence that have evolved so far: the product-directed or technological approach, and the theory-directed or scientific approach.", Schank, *op. cit.*, pp. 30-31.

l'impérative nécessité d'une réflexion théorique de haut niveau sur les processus intellectuels; d'où également les relations très étroites entretenues par la communauté de l'intelligence artificielle avec la linguistique, la psychologie, la philosophie, et évidemment l'informatique.

Les origines de l'intelligence artificielle

On croit souvent, à tort, que l'intelligence artificielle est un rejeton de l'informatique; il faut plutôt en situer l'origine à partir des espoirs de la cybernétique des années 1940 , dont l'objectif était de reproduire les mécanismes d'échanges d'informations à l'oeuvre dans les communications et dans la régulation des mouvements chez les êtres vivants. A la même époque, les premiers informaticiens, qui baptisèrent d'ailleurs leurs machines "computers" (pour "calculateurs"), ne voyaient en effet pas autre chose dans ces objets que de grosses calculatrices prodigieusement rapides, mais limitées aux opérations mathématiques élémentaires. Or, le caractère programmable de ces machines suggéra à certains cybernéticiens l'idée de leur faire simuler des opérations analogues à celles de l'esprit humain[5].

---

[5] Sur cette période, voir l'article de JACQUES PITRAT, "La naissance de l'intelligence artificielle", La Recherche, no 170, octobre 1985, pp. 1130-1141.

C'est ainsi qu'au début des années cinquante, grâce aux nouvelles techniques de programmation, on parvint à faire réaliser à l'ordinateur des programmes de calcul formel, comme par exemple, trouver la fonction dérivée d'une fonction donnée; les succès dans ce domaine encore très proche des mathématiques firent alors réaliser le potentiel de l'ordinateur dans le traitement non plus seulement de calculs arithmétiques, mais de **symboles** non numériques tels des variables. A partir de ce moment, toute situation formalisable devenait "traitable" par l'ordinateur.

1956 est considérée comme une année-clé de l'intelligence artificielle. Suite à l'observation du comportement d'élèves confrontés à des problèmes de logique, Allen Newell, H.A. Simon de l'Université Carnegie-Mellon et J.C. Shaw de la Rand Corporation firent la constatation que leurs sujets ne procédaient pas à une vérification systématique de toutes les possibilités de solution; au contraire ceux-ci avaient recours à des méthodes empiriques, mélange de règles "sur le pouce" (d'intuitions) et de raccourcis faisant appel à l'expérience antérieure, à l'analogie, à des "trucs" que toute personne expérimentée dans un domaine donné développe avec le temps. On baptisa ces méthodes imprécises **heuristiques** (du grec *heuriskein*, "trouver"), par opposition aux méthodes "algorithmiques".

L'algorithme, le procédé dominant en informatique, consiste à prévoir de façon détaillée et à explorer toutes les étapes conduisant à la solution d'un problème. On tenta donc d'incorporer des heuristiques à un programme appelé "Logic Theorist", système permettant de démontrer des théorèmes élémentaires en logique des propositions:

> "...ce programme a été conçu afin de déterminer comment il est possible de résoudre des problèmes ardus, tels que démontrer des théorèmes mathématiques, extraire de certaines données des lois scientifiques, jouer aux échecs, ou saisir le sens d'un texte anglais en prose.
>
> Les recherches dont il est question ici se donnent pour but de démontrer les processus complexes (l'heuristique) qui se révèlent efficaces dans la solution de problèmes. Autrement dit, nous ne nous intéressons pas à des méthodes qui garantiraient une solution juste, mais mettraient en oeuvre une somme impressionnante de calculs; car notre but est tout autre: il est de tâcher de comprendre comment un mathématicien, par exemple, est capable de démontrer un théorème, lors même qu'il ne sait pas encore, quand il se lance dans ce travail, comment il va procéder, ni même s'il va réussir[6]".

Au cours d'un colloque au Collège de Dartmouth (New Hampshire) la même année avec d'autres spécialistes de diverses disciplines - cybernéticiens, mathématiciens, psychologues, ingénieurs - réunis à primé abord autour de

---

[6] Allen Newell, J.C. Shaw et H.A. Simon, "Empirical Explorations with the Logic Theory Machine: A Case Study in Heuristics", Computers and Thought, Edward A. Feigenbaum et Julian Feldman, eds. (New York, McGraw-Hill, 1963), p. 109, cités dans Dreyfus, *op. cit.*, p. 14.

préoccupations reliées à la cybernétique, ces trois chercheurs firent une démonstration de leur programme et des possibilités de traitement de symboles par l'ordinateur. C'est à cette occasion que fut proposée et adoptée la désignation d'"intelligence artificielle" pour ce nouveau domaine de recherche. C'est aussi lors de cette rencontre historique que l'on conçut un projet ayant pour hypothèse de base que "chaque aspect de l'apprentissage ou de toute autre faculté associée à l'intelligence peut être décrit assez précisément pour qu'on puisse concevoir une machine capable de le reproduire[7]".

Les autres projets élaborés à cette occasion témoignent des aspirations des pionniers de l'intelligence artificielle: citons la création d'un système de neurones artificiels, d'un robot capable de se représenter son environnement, d'un programme de jeu d'échecs, d'un système pouvant démontrer des théorèmes mathématiques, etc. Les mois suivants furent féconds: suite à la réalisation du premier dictionnaire automatique (anglais-russe) doté d'une syntaxe primitive élaboré par Anthony Oettinger, on consacra des efforts

---

[7] Pamela McCorduck, Machines Who Think, San Francisco, W.H. Freeman & Co., 1979. p. 93, citée dans Le grand dérangement: à l'aube de la société d'information, par Arthur J. Cordell, Conseil des Sciences du Canada, Ottawa, 1985, p. 114.

considérables à la recherche dans le domaine de la traduction automatique. De leur côté, Simon et Newell mirent au point, en 1958, leur célèbre General Problem Solver (G.P.S.) capable, à partir d'une même stratégie générale de réduction des différences entre un état initial (les données) et l'état final (le but à atteindre), de résoudre une gamme variée de problèmes: énigmes comme celle visant à faire franchir un fleuve à des cannibales et des missionnaires, en évitant à ces derniers d'être dévorés par les premiers, puzzles, problèmes d'échecs ou d'intégration symbolique. C'est suite à ce succès pour le moins spectaculaire que Simon et Newell affirmèrent que "l'intuition, l'inspiration, la perspicacité, la faculté d'apprendre ne sont désormais plus l'apanage des humains: n'importe quel gros ordinateur puissant et rapide peut également en faire preuve lui aussi[8]"; c'est aussi à la même occasion qu'ils firent cette prédiction célèbre dans les milieux de l'intelligence artificielle à l'effet que tout au plus dans dix ans (i.e. en 1968), il y aurait un programme à la fois champion d'échecs et capable de démontrer un important théorème mathématique[9]. 1958 fut aussi l'année de la mise au point, par ces deux mêmes chercheurs, du premier

---

[8] Herbert A. Simon et Allen Newell, "Heuristic Problem Solving: the Next Advance in Operations Research", Operations Research, vol. 6 (janv.-fév. 1958), p. 6 cités par Dreyfus, *op. cit.*, p. 16.

[9] *Ibid.*, p. 22.

programme d'échecs (bien que présenté comme un bon joueur, notons que ce programme fut battu par un débutant de 10 ans). Parallèlement, du côté de la cybernétique, Frank Rosenblatt créait le PERCEPTRON: à la suite d'observations sur les agencements des cellules nerveuses, le fonctionnement de cette machine dépendait de "neuristnors" (cellules électroniques reproduisant grossièrement les neurones) organisés en réseaux et capables de percevoir des images. Les résultats limités de cette expérience compte tenu de ses coûts et du gigantisme de l'appareillage nécessaire en fit la dernière expérience notable d'intelligence artificielle dans le domaine de la cybernétique; l'intelligence artificielle obtenait de bien meilleurs résultats à partir des ordinateurs, ce qui remit en cause l'orientation héritée de la cybernétique tendant à reproduire des modèles physiques du cerveau. Les performances du Logic Theorist et du General Problem Solver avaient fini par démontrer que ce qui compte, ce n'est pas tant comment le cerveau fonctionne physiologiquement, mais ce qu'il en résulte.

Où en est l'intelligence artificielle trente ans plus tard? Nous allons procéder à un survol des réalisations les plus représentatives dans divers domaines (jeux, reconnaissance du langage naturel, résolution de problèmes et

perception), pour ensuite en proposer un bilan et dégager certaines perspectives d'avenir.

Les jeux

Le jeu compte parmi les toutes premières applications envisagées pour l'ordinateur: ainsi, dès 1945, K. Zuse, pionnier des premiers ordinateurs, développa un programme incorporant les règles du jeu d'échecs; Claude Shannon, un des pères de la théorie de l'information, conçut en 1949 une méthode informatique pour jouer aux échecs et Alan Turing en 1950 simula sur papier un programme capable de jouer[10]. L'intérêt pour le jeu s'explique par le fait que cette activité fait appel au raisonnement abstrait, à l'élaboration de statégies en fonction d'un but, selon des règles précises qui se prêtent bien à une mise en forme informatique.

Nous avons vu plus haut les résultats plutôt limités obtenus par Simon et Newell avec le premier programme d'échecs à être mis en oeuvre sur ordinateur. Reconnaissons à leur décharge que l'enjeu était de taille. De nombreux jeux sont facilement réalisables sur ordinateur (par exemple le TICTACTO, le jeu de NIM, OTHELLO) dans la mesure où le nombre d'états que doit parcourir le programme est limité; il existe même dans certains cas des algorithmes permettant au

---

[10] A ce sujet, lire Pitrat, *loc. cit.*, p. 1132.

programme de gagner à tout coup, ou tout au moins de faire match nul. Mais pour le jeu d'échecs, la méthode utilisée était de type combinatoire (envisageant tous les coups possibles à partir d'une position, puis toutes les ripostes, et ensuite toutes les réponses possibles à chacune des ripostes, et ainsi de suite), ce qui avait pour effet de provoquer après seulement les quelques coups d'ouverture une "explosion combinatoire", le nombre de coups possibles à considérer dépassant largement les capacités des ordinateurs les plus puissants de l'époque.

Des résultats plus intéressants furent obtenus en 1959 avec CHECKER, premier jeu de dames électronique mis au point par A.L. Samuel et comportant deux innovations importantes pour les développements futurs de l'intelligence artificielle: d'une part, ce programme comportait une fonction d'évaluation, attribuant une "valeur" à chaque coup envisagé (les possibilités étant incomparablement plus restreintes qu'aux échecs); d'autre part, ce programme **apprenait**, dans la mesure où il retenait en mémoire les configurations gagnantes. Capable de jouer contre lui-même, CHECKER fut bientôt en mesure de battre son créateur et même des maîtres aux dames[11]. Une étude approfondie du programme

---

[11] L'essentiel des procédures de ce programme est décrit dans l'article cité de Pitrat, p. 1134.

et de ses performances révéla toutefois que l'"intelligence" qu'il manifestait dépendait beaucoup plus des critères découverts et incorporés par Samuel (par exemple, valorisant toute position favorisant le contrôle du centre de l'échiquier) que de ses capacités d'apprentissage proprement dites; CHECKER était prisonnier des stratégies de son concepteur, et n'aurait jamais pu en développer de nouvelles par ses propres moyens.

Néammoins, il devait contribuer à l'élaboration de techniques nouvelles qui ont permis d'introduire une certaine heuristique dans les jeux d'échecs, réduisant les possibilités de coups à explorer aux plus intéressants, grâce aux fonctions d'évaluation: lorsque l'arbre des possibilités est restreint, le programme peut envisager plusieurs "profondeurs" (coup-riposte-coup, etc.). C'est ainsi que vers 1967 l'ordinateur atteint aux échecs le niveau d'un joueur moyen.

Aujourd'hui, malgré l'utilisation d'appareils sophistiqués pouvant envisager des centaines de milliers de positions par seconde, les programmes fondés sur les techniques décrites ci-haut plafonnent encore loin derrière les grands maîtres internationaux; pourtant, ces spécialistes n'envisagent qu'un nombre limité, tout au plus quelques dizaines, de coups à la fois. D'où vient la supériorité des

humains sur la puissance pourtant colossale des ordinateurs? Deux facteurs, selon des chercheurs[12], semblent jouer un rôle déterminant: alors que les programmes jouent coup par coup, réagissant au jeu de l'adversaire, le jeu humain applique une stratégie où chaque mouvement est relié à un BUT particulier; d'autre part, et il semble que les réalisations futures de programmes d'échecs intelligents en dépendent, sa capacité d'apprentissage permet au joueur humain expérimenté de développer une quantité phénoménale de connaissances heuristiques qui lui permettent, d'un coup d'oeil, d'aller directement aux possibilités les plus intéressantes. En dépit de ressources restreintes[13], les recherches actuelles dans le domaine du jeu tendent à développer des programmes capables d'apprentissage, de planification, et surtout, dotés d'une base de connaissances puisées auprès des grands maîtres.

---

[12] Voir à ce sujet de BENOIT FALLER, "L'ordinateur et les jeux de l'esprit", La Recherche, no 170, octobre 1985, pp. 1164-1174, et aussi d'ALAIN BONNET, L'intelligence artificielle: promesses et réalités, InterEditions, Paris, 1984, chap. 14.

[13] Selon Faller, tout au plus une centaine de chercheurs dans le monde travaillent dans le domaine du jeu. Notons aussi que des succès non négligeables ont été atteints dans d'autres domaines que les échecs, par exemple, au BACKGAMMON où le champion du monde a été battu par un programme, au POKER ainsi qu'au BRIDGE, où on a mis au point des systèmes capables de raisonner par hypothèses (*loc. cit.*, pp. 1168-1170).

La compréhension du langage naturel[14]

Les premiers efforts dans le domaine de la traduction automatique (le programme d'Oettinger de traduction du russe vers l'anglais) se contentaient de proposer pour chaque mot de la langue à traduire un répertoire des mots possibles dans la langue objet, sans rien interpréter du sens du texte: il s'agissait beaucoup plus d'un dictionnaire électronique que d'un système de traduction proprement dit, tout au plus capable de comparer une chaîne de caractères donnée avec une liste en mémoire, et de produire automatiquement une suite de mots traités de façon purement formelle. Afin de réussir la traduction de phrases complètes, dont l'agencement dépend de règles de syntaxe variables d'une langue à l'autre, on tenta alors d'améliorer ces systèmes en leur introduisant des règles grammaticales: les programmes ainsi réalisés analysaient dans un premier temps la structure syntaxique de la phrase à traduire, pour ensuite remplacer chaque mot par un mot correspondant, et enfin reconstruisaient la phrase en appliquant la syntaxe de la langue cible. La limite de cette technique fut vite atteinte: comment en effet choisir la bonne structure grammaticale et surtout les bons mots si on en ignore le sens?

---

[14] Sur ce sujet, voir l'article de Daniel Kayser, "Des machines qui comprennent notre langue", La Recherche, no. 170, pp. 1198-1209.

Les performances extrêmement limitées des programmes de traduction mécanique provoquèrent vers le milieu des années soixante l'abandon de ce champ de recherche au profit d'un domaine nouveau, soit la **compréhension** du langage. Diverses techniques furent utilisées pour développer des programmes capables de converser avec l'homme en langue naturelle; le plus célèbre est sans doute ELIZA[15], inventé par Joseph Weizenbaum vers 1964, qui simulait avec une vraisemblance remarquable, le discours d'un psychothérapeute non directif de type rogérien. Or, la technique utilisée était relativement simple, s'appuyant sur un programme en deux niveaux, le premier analysant la phrase de l'interlocuteur, et le deuxième fournissant un scénario qui, à partir de mots-clés, générait une réponse à peu près pertinente, en réutilisant les expressions utilisées par l'usager ou en formulant des questions et observations passe-partout. Mais comme dans le cas des programmes de traduction, il ne s'agissait que d'un traitement purement formel de chaînes de caractères sans "compréhension" de la part de l'ordinateur: en conséquence, en dépit de performances attrayantes pour un

---

[15] Au grand désarroi de son auteur, d'ailleurs, qui vit avec stupéfaction des experts en psychiatrie proclamer pour bientôt la psychiatrie presque complètement automatisée. Voir Weizenbaum, Joseph, Puissance de l'Ordinateur et Raison de l'Homme. Paris, Editions d'Informatique, 1981, pp. 7-13. (Traduction de Computer Power and Human Reason. W.H. Freeman and Co., 1976.)

non-initié[16], la technique utilisée ne pouvait, de par sa nature même, conduire à la compréhension du langage. Boden en fait l'évaluation suivante:

> "Ce genre de programmes répond au langage sans le comprendre [...] La plupart sont limités à un seul domaine du discours - contrairement aux hommes et aux femmes qui peuvent parler de baseball et d'algèbre, de psychologie et de généalogie [...] Au lieu d'avoir intelligemment recours à un système de connaissances recouvrant des domaines sémantiques nombreux et variés, ces programmes utilisent des règles relativement isolées et rigides pour produire une réponse à une entrée linguistique. Bref, une "machine parlante" du genre ne se comporte pas comme quelqu'un conversant dans sa langue maternelle, mais ressemble plutôt à une personne ayant recours à divers trucs et formules passe-partout pour cacher son incompréhension d'une langue étrangère[17]".

Une nouvelle technique fut donc élaborée vers la fin des années soixante, intégrant l'analyse grammaticale à

---

[16] Weizenbaum (*op. cit.*, pp. 5-6) donne l'exemple suivante d'un échange avec ELIZA:
Usager: Les hommes sont tous les mêmes.
ELIZA: DE QUELLE FACON
Usager: Ils sont toujours en train de nous embêter avec leurs histoires.
ELIZA: PENSEZ-VOUS A UN EXEMPLE PRECIS
Usager: Eh bien, mon ami m'a fait venir ici.
ELIZA: VOTRE AMI VOUS A FAIT VENIR ICI
Usager: Il dit que je suis souvent déprimée.
ELIZA: JE SUIS NAVRE D'ENTENDRE QUE VOUS ETES DEPRIMEE
Usager: C'est vrai, je suis malheureuse.
ELIZA: PENSEZ-VOUS QUE VENIR ICI VOUS AIDERA A NE PAS ETRE MALHEUREUSE (etc.)

[17] BODEN, Margaret A., Artificial Intelligence and Natural Man. *op. cit.*, p. 102.

de nouveaux procédés d'analyse **sémantique** et manifestant le début d'une véritable compréhension du langage. Le programme SHRDLU, présenté par Terry Winograd du M.I.T. en 1970 est un modèle du genre: il s'agissait de la simulation d'un robot doté d'un bras manipulateur dans un monde formé d'objets colorés et d'une boîte. Ce programme était capable de comprendre des ordres donnés en anglais lui demandant de manipuler les objets de son "monde" en les interprétant, et non seulement en répondant mécaniquement comme ELIZA, grâce à une **base de connaissances** sur certaines propriétés des objets composant son univers (par exemple, une pyramide peut tenir sur un cube, mais non l'inverse)[10].

Ce projet représente un jalon important pour le monde de l'intelligence artificielle dans la voie de la représentation des connaissances; car les limites des projets antérieurs venaient de leur incapacité de faire correspondre le monde extérieur (auquel réfèrent les mots) avec un système de représentations accessibles à l'ordinateur. En distinguant désormais l'analyse grammaticale de l'analyse sémantique, les travaux actuels dans le domaine de la compréhension du langage se concentrent sur l'élaboration de systèmes de représentations conceptuelles indépendants de la langue

---

[10] Pour une description complète de SHRDLU, voir Boden, *op. cit.*, chap. 6.

analysée: ces systèmes, reposant sur des techniques variées, procèdent soit par classification des mots lus dans des réseaux sémantiques reliant les concepts les uns aux autres selon des taxinomies, ou à partir de scénarios (scripts) fournissant au programme des connaissances de base sur les situations abordées, ou encore selon des prototypes (angl. *frame*), représentant un univers décomposé en classes et sous-classes d'objets comportant des attributs.

Ces techniques permettant à l'ordinateur de procéder à des inférences dans certains domaines bien délimités ont donné lieu à des réalisations fort intéressantes au cours des années 70: soulignons entre autres, sous la direction de Schank, le projet SAM, capable de comprendre un texte sur un sujet spécifique et d'en fournir un résumé pertinent (en diverses langues), le programme CYRUS simulant les réponses d'un sénateur américain de droite interrogé sur des questions de politique étrangère, et le programme IPP capable de généralisations (parfois un peu outrancières...) et d'apprentissage à partir de comptes rendus d'événements particuliers.

Il faut cependant des efforts considérables pour incorporer les connaissances nécessaires à ce genre de programmes, et leurs performances restent limitées à des domaines forcément très restreints; ainsi, SAM ne peut

interpréter que les histoires relatives à des accidents d'autos, alors qu'IPP ne peut procéder à des inférences qu'à partir d'attentats terroristes, et CYRUS et SHRDLU ne pourront jamais échanger entre eux sur leurs univers respectifs. En conséquence, pour autant de sujets, les chercheurs devraient construire de toutes pièces autant de modèles du monde, ce qui représente un défi insurmontable; il devient donc nécessaire de parfaire la capacité d'apprentissage des programmes pour qu'ils puissent élaborer eux-mêmes leurs propres modèles. De plus, l'esprit humain ne se contente pas d'enregistrer passivement les faits nouveaux en les juxtaposant aux connaissances acquises; il apprend à partir de ce qu'il connaît déjà et de plus, chacun sait combien fréquemment une nouvelle expérience ou connaissance nous oblige à réévaluer des notions apprises de longue date. En conséquence, les programmes capables d'apprendre devront être dotés d'une mémoire dynamique capable de s'auto-modifier en fonction de ses nouveaux acquis[19].

Quant à la reconnaissance vocale, les obstacles sont encore plus sérieux; car aux difficultés sémantiques

---

[19] "We are trying to develop a system that is capable of building up an increasing number of memories about different situations it has experienced, which it then can use in understanding even more stories and experiences [...] We must develop systems with flexible, changeable knowledge structures that can learn while doing." Schank, *op. cit.*, p. 164.

identiques à celles de la langue écrite s'ajoutent les sérieux problèmes de compréhension des mots individuels dans un débit oral continu et de perception des différences de prononciation: les projets les plus récents ne permettraient que l'identification de quelques centaines de mots[20].

La résolution de problèmes

Nous avons fait référence plus tôt dans ce chapitre au "General Problem Solver" de Newell et Simon; cette première période d'expérimentation dans le domaine de la résolution de problèmes se caractérise par la recherche d'une **logique universelle** applicable à tout problème.

La déclaration suivante résume bien les postulats des chercheurs engagés dans cette voie:

> "Tous les êtres humains sont des systèmes de traitement de l'information et possèdent donc certaines caractéristiques organisationnelles fondamentales communes; tous les êtres humains ont en commun quelques caractéristiques structurelles universelles, telles que des paramètres de mémoire pratiquement identiques. Ces caractéristiques communes entraînent des comportements semblables pour tous les résolveurs humains de problèmes[21]".

---

[20] Sur le sujet, cf. Richard Parent, Point de vue québécois sur l'intelligence artificielle, Ministère des Communications du Québec, 1984, p. 52.

[21] Simon et Newell, *op. cit.*, p. 864, cités dans Weizenbaum, *op. cit.*, p. 116.

Mais le G.P.S. avait ses limites: il exigeait une quantité astronomique d'informations lorsqu'il passait, par exemple, de la démonstration d'un théorème à la solution d'un problème aux échecs: alors que les mathématiques se prêtent bien à la déduction pure, à partir de principes logiques, on ne peut bien jouer aux échecs en ne connaissant que les règles du jeu. Cette approche, ignorant les connaissances spécifiques au profit d'une méthode universelle de résolution de problèmes, fit donc long feu.

Les recherches dans le domaine de la résolution de problèmes se tournèrent vers des applications plus limitées, intégrant des connaissances particulières au domaine d'application. On s'attarda aussi à l'étude des divers processus d'inférence chez l'humain; ainsi, en 1964, Thomas G. Evans réalisa un système (ANALOGY), doté de connaissances en géométrie et capable de résoudre des problèmes d'analogie géométrique semblables à ceux utilisés dans les tests d'intelligence[22].

Il faudra attendre toutefois les années 1970 pour qu'une percée majeure comportant des applications pratiques soit

---

[22] Des descriptions de ce programme sont données par Boden, *op. cit.*, pp.319-322 et par Dreyfus, *op. cit.*, pp. 100-106; pour les principes de programmation utilisés pour la résolution de problèmes analogiques, voir Alain Bonnet, *op. cit.*, pp. 160-162.

réalisée dans le domaine de la résolution de problèmes, grâce aux **systèmes experts.** Il s'agit de programmes informatiques qui ont cette caractéristique d'incorporer le savoir et l'expérience d'un spécialiste dans un domaine donné. Newell et Simon en furent les premiers instigateurs, après avoir réalisé qu'une grande partie des connaissances d'un expert pouvaient se représenter sous forme de règles informatiques (appelées aujourd'hui "règles de production") du genre "Si-alors-sinon..."[29]. Le tout premier système expert réalisé en 1974, MYCIN, est capable de diagnostiquer différents types d'infections sanguines, en indiquant le degré de fiabilité de son diagnostic, et de proposer la thérapie en conséquence, presque aussi bien qu'un spécialiste humain. De nombreuses autres applications sont apparues dans d'autres domaines, tels la génétique (GENESIS), l'analyse spectrométrique en chimie (DENDRAL), la prospection minière (PROSPECTOR), l'armement, l'éducation, le droit, la gestion, etc.

Ces systèmes se démarquent des programmes traditionnels en informatique dans la mesure où ils effectuent une séparation entre les éléments de connaissance susceptibles de servir (la base de connaissances) et le programme qui détermine l'utilisation de ces connaissances (l'interpréteur,

---

[29] Un exemple hypothétique: SI fièvre ET éruptions OU rougeurs ALORS rougeole; SI fièvre ET NON boutons ET NON rougeurs ALORS infection OU inflammation, etc.

ou le moteur d'inférence). Cette séparation a l'avantage de permettre l'accumulation en vrac de connaissances spécifiques, et la conception d'interpréteurs pouvant idéalement oeuvrer sur plusieurs bases de connaissances dans des domaines différents[24]. Le développement d'un système expert suppose un travail très complexe: il faut d'abord "extraire" les connaissances d'un spécialiste, non seulement les notions factuelles qu'il a accumulées, mais ses procédés (souvent inconscients) y compris ses méthodes "heuristiques"; il faut ensuite mettre en forme (modéliser) les faits, concepts, relations et procédures obtenus à l'étape précédente. La mise au point de ces techniques a donné lieu au développement d'une nouvelle discipline, l'"ingénierie de la connaissance". Mais malgré leur grande popularité, les systèmes experts comportent de sérieuses limites: d'une part, ils sont confinés à leur domaine d'expertise, et même à l'intérieur de celui-ci, les représentations des concepts utilisés sont superficielles (bien que MYCIN, par exemple, analyse de façon fort compétente les infections sanguines, il ne saurait répondre à des questions sur la nature du sang, la

---

[24] En théorie, car jusqu'à maintenant les interpréteurs sont trop spécifiques pour passer d'un domaine de connaissances à l'autre sans d'importantes modifications. Sur le sujet, voir HERVE GALLAIRE, "La représentation des connaissances", La Recherche, no 170, octobre 1985, pp. 1240-1248.

fonction de la circulation, du coeur ou des artères, etc.). D'autre part, certains chercheurs, comme Schank, mettent en doute la prétention des systèmes experts à faire partie du domaine de l'intelligence artificielle:

> "Les systèmes experts, bien que potentiellement utiles, ne nous ont pas fait progresser dans notre démarche de création d'une machine intelligente. L'intelligence véritable exige la capacité d'apprendre; elle peut raisonner à partir de ses expériences, TIRER AU JUGE (shoot from the hip), faire appel à des connaissances générales et inférer à partir de simples intuitions. Aucune de ces aptitudes ne se retrouve dans les systèmes experts. Ils ne s'améliorent pas avec l'expérience, ils passent simplement à la règle <si/alors> suivante[25]".

La perception des formes (vision)

Bien qu'il s'agisse d'un domaine important de recherches en intelligence artificielle, particulièrement au chapitre de la robotique, nous ne nous y attarderons pas très longuement. Les réalisations sont encore infiniment primitives, comparées à celles de l'oeil animal (ou humain) et les succès dans ce domaine particulier vont vraisemblablement dépendre des progrès qui seront réalisés dans les autres secteurs de l'intelligence artificielle dont nous traitons à la fin de ce chapitre.

---

[25] Schank, *op. cit.*, p. 34.

Les premiers programmes dignes de mention en vision artificielle ont été conçus par L.G. Roberts et Adolfo Guzman à la fin des années soixante[26]: à partir de polyèdres représentant des objets solides en trois dimensions dans une disposition où certains étaient partiellement cachés par d'autres, ces deux programmes avaient pour objectif de distinguer les formes les unes des autres.

La grande difficulté de ce genre d'exercice résulte du fait que le système visuel (humain ou artificiel) ne perçoit que deux dimensions, et que la troisième doit être déduite: ainsi, l'information nécessaire peut être obtenue par l'humain par exemple en se déplaçant parmi les objets lorsqu'ils lui sont inconnus, ou à partir de représentations analogues en mémoire. Il faut donc trouver divers moyens de fournir aux programmes de reconnaissance de formes les renseignements nécessaires à l'interprétation des images fragmentaires qu'ils perçoivent.

C'est ainsi que le programme de Roberts pouvait non seulement identifier un certain nombre des objets présentés

---

[26] Pour une description plus complète de ces programmes, voir Margaret A. Boden, *op. cit.*, chapitres 8 et 9. Aussi, l'article de Maurice Briot et Robert de Saint Vincent, "La vision des robots", La Recherche, no 170, octobre 1985, pp. 1264-1273.

(même ceux partiellement cachés), mais aussi calculer leurs dimensions et leur position relative grâce à des connaissances très détaillées qu'on lui avait incorporées sur, entre autres, la géométrie tridimensionnelle, les angles de prises de vue de la caméra par rapport au sol, et certaines classes fondamentales d'objets sous formes de représentations géométriques abstraites. L'image, même partielle, d'un objet était alors comparée aux représentations internes dans la mémoire du système qui, par de multiples rapprochements et transformations, la reconstruisait en entier. Guzman est allé encore plus loin avec son programme SEE en lui fournissant non pas des connaissances détaillées sur les formes géométriques complètes susceptibles d'être rencontrées, mais des représentations beaucoup plus simples et abstraites (des "vertex" tels un angle en L, une forme en pointe de flèche, une fourche, etc.) à partir desquelles le système pouvait reconnaître des objets de forme inconnue les uns des autres et les distinguer de leur arrière-plan.

D'autres programmes plus sophistiqués suivront dans les années '70, capables d'analyser les contrastes d'intensité lumineuse sur des objets de formes diverses pour en dégager la silhouette, d'autres pour identifier l'ombre changeante projetée par les objets, ou encore pour percevoir le

mouvement, les contours de visages, etc. Mais les handicaps techniques encore aujourd'hui sont colossaux:

> "Pourquoi est-il si difficile d'élaborer un système de vision artificielle? L'un des principaux obstacles à cette mise au point est la formidable puissance de calcul qu'il faut mettre en oeuvre, et que l'on n'a d'ailleurs pas encore réussi à estimer correctement. On sait maintenant qu'une douzaine de centres cérébraux - constituant 60% du cortex humain - prennent part à l'activité de la vision; la rétine, pourvue de cent millions de cônes et de bâtonnets et de quatre couches de neurones supplémentaires, effectue dix milliards d'opérations par seconde avant même que la "représentation" de la scène observée n'atteigne le nerf optique[27]".

Les difficultés sont semblables à celles de la reconnaissance du langage naturel, du jeu d'échecs et du développement de systèmes-experts; dans tous les cas on est confronté après quelques succès de départ au problème de l'explosion combinatoire du nombre de connaissances nécessaires pour un fonctionnement efficace même dans en environnement banal (une cuisine, par exemple). Il semble bien alors que sans expérience du monde, sans connaissances de sens commun ni capacité d'apprentissage, la perception des ordinateurs sera limitée pour longtemps encore à la reconnaissance de formes géométriques simples dans des mondes-jouets; il faudra des progrès révolutionnaires dans

---

[27] BRIOT MAURICE ET ARNAUD ROBERT DE SAINT VINCENT, *loc. cit.*, p. 1264.

ces divers domaines pour permettre la création d'un robot non spécialisé capable de se déplacer dans un environnement inconnu.

Perspectives

> "La question fondamentale n'est donc plus de savoir si certains aspects de l'activité de compréhension peuvent être reproduits en machine, mais plutôt si les capacités encore très limitées des systèmes actuels peuvent être développées ou bien si le comportement intelligent est par essence irréductible à l'approche informatique. Dans ce dernier cas, les résultats actuels, pour spectaculaires qu'ils puissent parfois paraître, seraient voués à plafonner à un niveau tellement bas que l'emploi de mots comme "compréhension" ou "intelligence" artificielles constitueraient un abus inadmissible[20]".

Or, du point de vue technique, il semble bien que nous soyons encore loin de la réalisation d'ordinateurs, qui à l'instar de HAL dans le film **2001: Odyssée de l'espace**, seraient capables simultanément de jouer aux échecs, de voir, de comprendre et parler un langage naturel, d'apprendre par eux-mêmes, de modifier leur programme d'opération, et de faire preuve de conscience de leurs actes. Bien que l'on parvienne à reproduire sur l'ordinateur certains comportements jusqu'ici spécifiques à l'être humain, ceux-ci ne peuvent s'exercer que dans des domaines extrêmement restreints; de plus, comme nous l'avons vu avec les jeux

---

[20] KAYSER, DANIEL, *loc. cit.*, p. 1198.

d'échecs, le plus souvent les processus utilisés n'ont rien de comparable à ceux, généralement très mal connus, des humains. Enfin, la difficulté majeure de l'intelligence artificielle réside présentement dans la compréhension de la langue, laquelle se heurte à des obstacles insurmontables dans l'état des techniques actuelles: comment faire partager à la machine la connaissance du monde nécessaire à la communication, au raisonnement de sens commun, et même à l'empathie qui joue un rôle non négligeable dans les échanges entre humains? Le milieu de l'intelligence artificielle est dramatiquement conscient de ses limites, et se garde bien désormais des prédictions optimistes de ses débuts: mais des efforts considérables sont présentement déployés au Japon, en Europe et aux Etats-Unis pour le développement d'ordinateurs dits de 5ème génération[29] dotés de multiprocesseurs capables de traiter simultanément (grâce à une nouvelle architecture utilisant non plus des circuits **séquentiels** mais **parallèles**) des milliards d'informations, et qui pourront, du moins le prétend-on, voir, comprendre le langage naturel et traduire

---

[29] Voir MOTO-OKA, TOHRU, "Les ordinateurs de cinquième génération", La Recherche, no 154, avril 1984, pp. 516-525. Aussi, sur le même sujet, EDWARD FEIGENBAUM et PAMELA McCORDUCK, La Cinquième Génération: le pari de l'intelligence artificielle à l'aube du 21ème siècle, InterEditions, Paris, 1984. (Traduction de The Fifth Generation, Addison-Wesley Publishing Co., Reading, Mass., 1983.)

un texte avec une fiabilité de 90%; ces super-machines comporteront plusieurs niveaux de connaissance leur permettant tout comme les humains d'établir des interrelations, des analogies entre leurs vastes bases de connaissances, et devraient être capables d'apprendre et de s'auto-améliorer avec l'expérience. Par ailleurs, on est à mettre au point, entre autres à l'Institut Armand-Frappier de Montréal, des bio-puces à partir de la plus petite structure complexe connue, la protéine, permettant, outre une miniaturisation fantastique de l'ordinateur, le développement d'une interface (jonction) directe avec le système nerveux humain.

Ces objectifs de recherche sont riches en postulats de toutes sortes sur la nature de la pensée; ils soulèvent d'importantes interrogations sur le rôle de l'expérience, de la sensibilité, de l'intuition et de la subjectivité dans l'activité cognitive. Il va de soi que la philosophie se sente interpellée par ces questions nouvelles.

CHAPITRE III

Machines, logique et intentionnalité

Dans Computing Machinery and Intelligence[1], Turing avait rassemblé un certain nombre d'objections à ses théories sur la possibilité de réaliser une intelligence de synthèse. On peut les résumer comme suit:

> l'argument théologique, à l'effet que la pensée est une fonction de l'âme immortelle, donnée à l'homme par Dieu, et dont sont privés animaux et machines.
>
> l'objection mathématique, qui invoque particulièrement le théorème de Gödel sur l'*incomplétude* des systèmes logiques pour l'étendre aux ordinateurs.
>
> l'argument de la conscience, faculté essentielle aux pensées et aux émotions, sans laquelle la machine ne peut se comparer au cerveau.
>
> l'énumération des diverses incapacités des ordinateurs au chapitre du sens moral, du langage, du sens de l'humour, de la capacité d'aimer, etc. (dérivé de l'objection précédente).
>
> l'argument dit de Lady Lovelace, à l'effet qu'une machine est dépourvue d'originalité et de créativité, parce que prisonnière de sa programmation.
>
> l'imprévisibilité du comportement humain, qui ne peut être enfermé dans un ensemble de règles détaillées prévoyant toutes les circonstances possibles.

---

[1] The Mind's I, *op. cit.*, pp. 56-57.

> la continuité du système nerveux biologique, par opposition au fonctionnement discontinu (arrêt-marche) de l'ordinateur.[1]

Ces objections ne sont pas toutefois du même ordre: ainsi, l'argument théologique qui forme une classe à part, part du principe que Dieu n'a créé l'âme que pour l'être humain et qu'il serait impie pour ce dernier de s'arroger un privilège divin. Pourtant, répond Turing, injurions-nous Dieu lorsque nous engendrons des enfants? Quelque chose interdit-il au Tout-Puissant de créer des âmes pour d'autres créatures que l'être humain? En conséquence, qu'il s'agisse d'enfants ou de machines, "ne sommes-nous pas plutôt, dans l'un et l'autre cas, les instruments de Sa volonté en fournissant l'enveloppe matérielle des âmes que Dieu crée[2]?"

Les autres arguments relevés par Turing se chevauchent parfois; certains répètent à peu près le même raisonnement sous un aspect légèrement différent. Pour en simplifier l'analyse, nous les regrouperons sous les trois catégories suivantes:

---

[1] Nous omettons volontairement l'argument à l'effet que l'esprit humain est doté de pouvoirs métapsychiques; Turing n'a pas échappé aux croyances très répandues à ce sujet au début des années cinquante même dans les milieux scientifiques.

[2] Turing, *loc. cit.*, p. 58.

1) Ceux qui invoquent les limites des systèmes logiques (notamment à partir du théorème de Gödel).

2) Ceux selon qui la conscience - ou en termes plus actuels, l'**intentionnalité** - est une faculté sinon spécifiquement humaine, du moins propre aux êtres biologiques.

3) Enfin, les arguments à l'effet que le cerveau et la pensée de l'être humain ne peuvent se réduire à des systèmes de relais électroniques ni à des règles de procédure logiquement formalisables.

Nous procéderons dans les pages qui suivent à une présentation plus détaillée de ces thèses et de leurs principales réfutations. Le chapitre actuel sera notamment consacré aux conséquences pour l'intelligence artificielle de la théorie de l'incomplétude des systèmes axiomatiques, particulièrement bien illustrée dans un article célèbre d'un philosophe d'Oxford J.R. Lucas[2], et à l'argument sur l'intentionnalité développé par le philosophe américain, John R.

---

[2] LUCAS, J.R., "Minds, Machines, and Gödel", Philosophy, 36, (1961), pp. 112-127. (Reproduit dans la collection d'articles Minds and Machines, par Alan Ross Anderson, Engelwood Cliffs, N.J., Prentice-Hall, 1964.)

Searle[4]. Nous réserverons pour le chapitre suivant l'exposé critique du point de vue phénoménologique sur les processus cognitifs en intelligence artificielle dont Hubert Dreyfus a été jusqu'ici le principal protagoniste.

Lucas et le théorème de Gödel

Depuis que l'homme a réalisé que c'est sa raison qui le distingue de l'animal, il a tenté d'en codifier le fonctionnement: ainsi fit Aristote avec le raisonnement déductif, Euclide avec la géométrie, Descartes avec "la Méthode". Mais l'effort ultime s'est produit au XXème siècle, lorsque Russell et Whitehead entreprirent l'ambitieux projet de dégager les fondements de la logique et des mathématiques en les unifiant dans un système cohérent et complet (i.e. dont les théorèmes sont démontrables de l'intérieur même du système): cet exercice produisit le monumental Principia Mathematica en 1910. Cependant, jusqu'à quel point les démonstrations de cet ouvrage pouvaient se suffire à elles-mêmes, et ne faire jamais l'objet d'aucune contradiction, c'est ce que s'efforcèrent de découvrir les

---

[4] SEARLE, JOHN R., "Minds, Brains and Programs", The Behavioral and Brain Sciences, Vol. 3, Cambridge University Press, 1980, pp. 416-423, accompagné de 27 réponses de chercheurs de différentes disciplines. (L'article de Searle est reproduit dans The Mind's I de Dennett et Hofstadter, au chapitre 4.)

mathématiciens sans succès au cours des vingt années suivantes.

Or, en 1931 parut un ouvrage[5] du mathématicien autrichien Kurt Gödel révélant non seulement les failles du système axiomatique de Russell et Whitehead, mais de façon encore plus générale démontrant que tout système axiomatique comporte un certain nombre de propositions "indécidables" dont on ne peut prouver ni la vérité ni la fausseté dans le cadre même du système. Toutefois, la vérité de ces propositions peut être constatée de l'extérieur par un esprit humain, ce qui entraîne le paradoxe suivant: pour qu'un système puisse démontrer la vérité de certaines de ses propositions, il devrait être inconsistant, i.e. ne pas respecter ses propres axiomes[6]. Du point de vue des mathématiques, cette découverte eut le même effet que celle du principe d'incertitude de Heisenberg en physique.

On n'allait pas tarder à en vérifier les conséquences pour l'informatique, puisque les ordinateurs, nécessitant une

---

[5] Gödel, Kurt. On Formally Undecidable Propositions. New York: Basic Books, 1962. Original publié sous le titre "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I." Monatshefte für Mathematik und Physik, 38 (1931), 173-198.

[6] Pour un exposé fort bien étoffé du théorème de Gödel et de ses conséquences, voir D.R. Hofstadter, Gödel, Escher, Bach...

programmation logique, devraient être soumis aux mêmes limites que les systèmes axiomatiques: Turing avait relevé cette objection qui sera reprise et approfondie par Lucas en 1961 afin de démontrer la supériorité de l'esprit humain, capable d'inconséquence, sur la machine:

> "Le théorème de Gödel doit s'appliquer aux machines cybernétiques car il est de l'essence même d'une machine d'actualiser de façon concrète un système formel. Il s'ensuit donc que pour toute machine conséquente et capable d'opérations arithmétiques simples, il y a une formule dont elle est incapable de démontrer la vérité - i.e. qui est improuvable à l'intérieur même du système - mais dont la vérité peut être constatée. Il s'ensuit donc qu'aucune machine ne peut reproduire de façon complètement adéquate l'esprit: ce dernier est essentiellement différent de la machine[7]."

La machine, selon Lucas, est entièrement déterminée dans sa conception même: à telle entrée, tel type de construction doit produire de façon rigoureusement prévisible telle réponse. Même en y introduisant une fonction aléatoire, cette fonction, selon Lucas, ne permettra des choix qu'entre des possibilités conséquentes avec le système. C'est pour cette raison que dans la mesure où tout modèle mécanique de l'esprit devra incorporer la capacité d'énoncer des vérités arithmétiques, il se heurtera inévitablement à une "formule gödelienne" qu'il ne pourra prouver; pourtant, un esprit

---

[7] Lucas, J.R. "Minds, Machines...", *op. cit.*, p. 113.

humain, au même moment, pourra constater que cette formule est vraie.

Ce que Lucas remet ainsi en cause, ce n'est pas le fait qu'une machine puisse reproduire telle opération particulière de l'esprit humain: il conteste plutôt la possibilité qu'une même machine puisse en reproduire **toutes** les fonctions, à cause justement de cette capacité de l'esprit, grâce à la conscience, de pouvoir "sortir" de lui-même:

> "[...]le concept de créature consciente est implicitement différent de celui d'un objet inconscient. Lorsque nous affirmons qu'un être conscient sait quelque chose, non seulement nous affirmons qu'il sait cette chose, mais qu'il sait qu'il sait cette chose, et qu'il sait qu'il sait qu'il sait, et ainsi de suite, tant que nous voulons bien poser la question; il faut bien reconnaître là une régression à l'infini [...]. Bien qu'un être conscient puisse continuer ainsi indéfiniment, nous ne faisons pas état de cela pour simplement faire valoir le nombre de tâches qu'il peut accomplir; ni ne prétendons-nous que l'esprit soit une séquence infinie formée d'un moi, d'un sur-moi et d'un sur-sur-moi. Nous tenons plutôt à faire valoir qu'un être conscient forme un tout, et que nous ne référons aux parties de son esprit que dans un sens métaphorique.
>
> Les paradoxes de la conscience surgissent parce qu'un tel être peut être conscient de lui-même et des autres objets de son environnement sans pour autant perdre son unité constitutive. Ce qui signifie qu'un être conscient peut faire face à des questions gödeliennes devant lesquelles les ordinateurs sont impuissants parce qu'il peut à la fois s'observer lui-même et envisager ses propres actions sans avoir à devenir autre que celui qui a agi. On peut fabriquer une machine capable, disons, de s'observer elle-même, mais il lui faudra alors devenir autre qu'elle-même, c'est-à-dire la vieille machine avec une "pièce neuve"; l'esprit par contre

> a pour propriété inhérente de pouvoir à la fois réfléchir sur lui-même et critiquer ses actes sans "pièce" supplémentaire[8]."

Aucune machine, donc, peu importe son degré de complexité, ne pourra franchir les limites radicales constatées par Gödel dans tout système formel: l'ordinateur dans son essence même est ainsi voué à l'incomplétude, mais pas l'esprit humain: "La machine ne pourra jamais être un modèle satisfaisant de l'esprit. Les prototypes que nous produisons sont des mécaniques **mortes** alors que l'esprit, lui, est **vivant** et demeure toujours un pas en avant de tout système formel ossifié. Grâce au théorème de Gödel, l'esprit aura toujours le dernier mot[9]". Autre conséquence d'une importance extrême pour la philosophie, selon Lucas: l'application du théorème de Gödel aux sciences cybernétiques porte un coup fatal aux conceptions mécanistes de l'esprit.

Cet argument n'a pas manqué de susciter de nombreuses réactions. Turing avait jugé présomptueux ce type d'argumentation qui affirme, sans preuve d'aucune sorte, que les limites d'une machine particulière ne s'appliquent pas à l'esprit humain. Cette réfutation a été reprise par

---

[8] *Ibid.*, p. 125.

[9] *Ibid.*, p. 116.

Hofstadter[10] qui souligne que l'affirmation de Lucas à l'effet que l'esprit humain échappe aux limites des systèmes axiomatiques n'est pas démontrée de façon explicite: étant donné qu'il n'existe pas d'algorithme capable de rendre compte de la façon dont l'esprit humain dépasse l'ordinateur, rien ne peut garantir que face à un problème d'une très grande complexité l'humain ne se heurtera pas éventuellement aux mêmes limites.

Au sujet de l'argumentation de Lucas, Margaret Boden[11] fait remarquer que le théorème de Gödel ne s'applique qu'aux systèmes clos, dont les axiomes et les règles d'inférence sont établis de façon rigide; selon elle, si un programme est capable d'apprentissage, donc d'intégrer de nouvelles règles et axiomes, ce qui était "indécidable" la veille pourrait devenir "décidable" le lendemain. Il est alors erronné d'étendre aux machines la démonstration de Gödel, laquelle ne prouve pas que des propositions connues comme vraies par des humains ne pourraient l'être également par des machines programmées adéquatement. C'est dans ce même ordre d'idée

---

[10] Hofstadter, *op. cit.*, pp. 475-476.

[11] Margaret A. Boden, Artificial Intelligence and Natural Man, *op. cit.*, pp. 434-435.

que, suite à la parution de l'article de Lucas, F.H. George[12] avait remis en cause l'analogie postulée entre systèmes formels et ordinateurs; selon lui, l'argumentation aurait été valable si les "machines cybernétiques" n'étaient que des machines déductives, ce qu'elles ne sont pas: selon George, ces machines, contrairement aux appareils automatiques, peuvent se programmer elles-mêmes et réorganiser leurs instructions en fonction de changements dans leur environnement, échappant ainsi aux contraintes des systèmes formels.

Marvin Minsky[13] fait valoir que ce qui peut être désastreux en logique mathématique, par exemple le paradoxe du menteur[14] (que Gödel a reproduit en termes mathématiques), ne pose aucun problème en informatique: les paradoxes auto-réfléchissants proviennent du fait qu'on tente de décrire dans un seul et même moment figé un fait qui change et se transforme dans le temps. Selon ce chercheur, un programme pourrait accomplir ce que nous, humains, faisons devant une

---

[12] F. H. George, "Minds, Machines, and Gödel: another reply to Mr Lucas". Philosophy, no 37, (1962), pp. 62-63.

[13] Marvin Minsky, "Computer Science and the Representation of Knowledge", *loc. cit.*, p. 399.

[14] Attribué à Epiménide, le Crétois: "Tous les Crétois sont des menteurs".

formule paradoxale du genre "je mens toujours": d'abord, envisager la proposition comme vraie, ensuite comme fausse, puis comme vraie pour enfin, s'il s'agit d'une programme raffiné, en remettre la solution à plus tard...

Par ailleurs, la thèse de Lucas présuppose que l'ordinateur est incapable d'irrationalité, qu'il est prisonnier d'un fonctionnement logique sans failles, ce qui ouvre la porte à une réfutation supplémentaire de la part de Hofstadter[15]: tout programmeur peut facilement faire produire à l'ordinateur un ensemble de propositions inconsistant. Il est en effet possible, tout en respectant le fonctionnement de circuits logiques (au niveau du *hardware*), avec un ensemble d'instructions orientées vers autre chose que le calcul logique ou mathématique, de générer à un autre niveau de l'ordinateur des résultats erronnés (mais sans fautes du point de vue du programme exécuté)[16]. C'est de cette même façon, selon Hofstadter, que les neurones du cerveau (qui ne semblent pas échapper aux lois connues de la physique et des mathématiques) supportent différents niveaux de manipulation

---

[15] Hofstadter, *op. cit.*, pp. 577-578.

[16] Un petit exemple bien simple en BASIC d'un programme qui donne immanquablement "2+2=5":

```
10 INPUT "Tapez un chiffre:";A
20 INPUT "Tapez un autre chiffre:";B
30 IF A=2 AND B=2 THEN C=5: GOTO 50
40 C=A+B
50 PRINT "La somme de ces chiffres est de:";C
```

de symboles susceptibles d'illogisme, de confusion et d'appréciations subjectives. L'erreur de Lucas, conséquemment, en serait une de simplification et de réduction des opérations fortement hiérarchisées de l'esprit humain à celles de son niveau à la fois le plus rigide et le plus élémentaire.

Même si à notre avis, la démonstration de Lucas n'atteint pas l'objectif recherché, elle n'en est pas moins fort intéressante et a le mérite de développer pleinement une objection à peine esquissée par Turing. Mais, et c'est sa grande faiblesse, elle postule une identité parfaite entre la machine et un système axiomatique quelconque. Comme le soulignait tout particulièrement Hofstadter plus haut, l'esprit humain peut, à différents niveaux, opérer selon des logiques différentes (ce qui explique sans doute les tensions et contradictions éprouvées par tout esprit lucide); ce type de structure, formée de niveaux d'interactions superposés, est éminemment facile à reproduire dans un programme informatique (dans sa forme, évidemment, pas dans son contenu). Jusqu'à nouvel ordre, donc, il ne semble pas que les limites des systèmes axiomatiques démontrées par Gödel aient les répercussions envisagées par Lucas pour l'intelligence artificielle.

Searle et l'intentionnalité[17]

John R. Searle, philosophe à l'Université de Californie à Berkeley, distingue deux sortes d'intelligences artificielles: la "faible" et la "forte". Pour l'intelligence artificielle faible, l'intérêt de l'ordinateur réside dans son utilisation comme outil pour l'étude de l'esprit; l'intelligence artificielle forte considère plutôt l'ordinateur comme une véritable intelligence capable de **comprendre** et de reproduire les processus cognitifs humains: les travaux de Terry Winograd (SHRDLU - voir le chapitre 2) et surtout de Roger Schank sont nommément visés par la critique de Searle, qui nous propose comme moyen de vérification de nous faire simuler (métaphoriquement) le fonctionnement d'un programme "intelligent" typique.

Laissons-nous donc enfermer dans une pièce avec une liasse de papiers sur lesquels sont inscrits des caractères chinois (en supposant que nous ignorons tout de cette langue); à ce matériel on ajoute un deuxième lot de caractères chinois avec en plus un ensemble de règles rédigées en français sur la façon d'associer les caractères du premier groupe avec ceux du deuxième; puis supposons qu'enfin l'on nous confie un troisième lot de caractères

---

[17] Searle, "Minds, Brains, and Programs...", *loc. cit.*

chinois avec un mode d'emploi, en français, sur la façon de les assortir aux deux autres et de produire des symboles chinois particuliers en réponse à certaines configurations de ce troisième groupe de caractères. On nous révèle alors que cette dernière livraison s'appelle "questions", la première "scénario", et la deuxième "histoire", alors que les caractères que nous produirons vont s'appeler "réponses aux questions"; enfin, les règles qu'on nous a fournies en français porteront le nom de "programme"[18]. Searle suppose qu'avec un peu de pratique, dans l'hypothèse où les règles qu'on nous a fournies sont adéquates, ce système nous permettra de produire en chinois des réponses plausibles aux questions qui nous auront été posées (en chinois) aussi efficacement (du point de vue d'un interlocuteur comprenant ces deux langues) que lorsque nous conversons en français. A cette différence près qu'en français, nous **comprenons** le sens des questions et des réponses, mais pas en chinois.

L'ordinateur, qui opère de la même façon que nous venons de le faire avec les symboles chinois, ne comprend donc certainement pas le **sens** des histoires qu'il traite. En

---

[18] La métaphore de Searle simule une des méthodes de programmation conçues par Schank pour la reconnaissance du langage naturel entre autres dans le programme SAM (voir chapitre 2): le "scénario" (ou script) sert à décoder "l'histoire" et à répondre de façon pertinente, selon des "règles" (i.e. le programme), aux "questions" posées.

conséquence, il n'est pas non plus justifié de la part de l'I.A. "forte" de prétendre que l'examen de ce genre de programme peut aider à l'explication des mécanismes de compréhension chez l'humain. D'où la conclusion à l'effet qu'attribuer la capacité de **comprendre** aux ordinateurs autrement que par métaphore ou analogie est un abus de langage: c'est, selon Searle, se laisser prendre à la tendance naturelle que nous avons d'étendre aux artéfacts notre propre intentionnalité[19] lorsque nous prétendons, par exemple, que la calculatrice **sait** additionner, que le radar **perçoit** un obstacle, et ainsi de suite.

On ne peut ainsi, affirme Searle, comparer la démarche de l'ordinateur et celle de l'esprit: lorsque, dans son exemple, nous manipulons de façon purement formelle les symboles chinois - selon la procédure informatique - nous n'y comprenons absolument rien. Ce mode d'opération n'a rien à voir avec ce que nous faisons lorsque nous conversons dans notre propre langue.

> "Lorsque je dis que la porte **comprend** les instructions qu'elle reçoit d'une cellule photo-électrique, j'énonce quelque chose de tout à fait différent de l'affirmation que je **comprends** le français. Si Schank se contentait d'affirmer que

---

[19] John Searle, *loc. cit.*, p. 418. L'auteur utilise la notion d'intentionnalité dans son sens husserlien, i.e. en référence à des états de conscience orientés vers le monde extérieur.

> l'ordinateur comprend de la même façon que la porte, et non comme moi, le débat cesserait là. Mais Newell et Simon (1963) prétendent que l'activité cognitive des ordinateurs est parfaitement semblable à celle des humains. Cette prétention a tout au moins le mérite d'être franche, et c'est à elle que je vais m'attarder. Je défendrai donc que l'ordinateur ne peut comprendre littéralement que ce qu'une auto ou une calculatrice comprend, c'est-à-dire proprement rien[20]".

Searle tente dans la suite de son article de prévoir les objections à sa démonstration et d'y répondre de la façon suivante:

1) L'objection du "système": selon certains chercheurs, tout en reconnaissant que la personne enfermée dans une pièce ne comprend rien, elle fait partie d'un tout, et c'est **le système pris dans son ensemble** qui comprend le chinois.

Searle réplique en supposant que cette même personne assimile tous les éléments du système: caractères chinois, règles de procédure, etc. Une fois le système bien mémorisé, sera-t-elle en mesure de mieux comprendre le sens des mots chinois utilisés? Et pourtant, elle aura assimilé le système dans sa totalité... Prétendre qu'un tel bricolage comprend, c'est confondre **manipulation formelle d'informations** et **compréhension** proprement dite: en ce sens, ajoute Searle, sous un certain angle l'estomac est un système qui traite de

---

[20] *Ibid.*

l'information, mais qui oserait affirmer que cet organe y comprend quelque chose? Enfin, Searle pose la question suivante à l'I.A. "forte": puisque vous vous prétendez une branche de la psychologie, à moins que vous affirmiez que l'esprit est partout, quelle distinction faites-vous entre les phénomènes psychiques et ceux qui ne le sont pas?

2) L'objection du "robot": si toutefois nous placions un ordinateur avec un programme différent de celui décrit ci-haut à l'intérieur d'un robot capable de voir, de se déplacer, de manger, et le reste: ce robot serait alors dirigé par l'ordinateur qu'il contient, au même titre que le cerveau dirige le corps. Il ne s'agirait alors plus d'un système manipulant strictement des symboles, mais d'une entité véritablement reliée à son environnement et donc capable de comprendre.

A ceci Searle fait remarquer que l'ajout de facultés motrices ou perceptives ne fournirait pas à l'ordinateur au sein du robot le don de compréhension ou d'intentionnalité, car les informations sûr le monde "extérieur" transmises par les "organes" sensoriels du robot seraient traitées de façon toute aussi formelle que dans l'exemple original; c'est le programme de l'ordinateur, et non pas un état intentionnel quelconque, qui ferait réagir le robot.

3) L'objection du simulateur de cerveau: imaginons un programme qui reproduit l'activité synaptique du cerveau d'un chinois au moment où ce dernier interprète des questions et y répond. S'il y avait identité au niveau synaptique, quelle différence pourrait-alors exister entre le programme du cerveau chinois et le programme de l'ordinateur?

Searle répond qu'encore une fois, nous n'aurions pas simulé le bon processus en nous bornant à la structure formelle de l'activité neuronale: nous aurions négligé ce qui compte vraiment dans le cerveau, soient ses propriétés causales (causal properties) et sa capacité de produire des états intentionnels (intentional states).

4) La combinaison des trois premières objections: supposons une entité qui amalgame les caractéristiques mentionnées dans les trois objections précédentes, soit un robot avec à l'intérieur un ordinateur reproduisant toute l'activité synaptique du cerveau humain, dont le comportement est parfaitement semblable à celui d'un être humain...

Evidemment, si j'ignore tout des processus internes de cette machine, reconnaît Searle, je serai fortement tenté de lui attribuer de l'intentionnalité; mais ceci change-t-il quelque chose au fond du problème? Qu'une entité ressemble à un être humain même dans ses comportements n'est pas

suffisant pour que nous lui attribuions les mêmes états psychiques. Bien sûr, nous reconnaissons de l'intentionnalité à certains mammifères supérieurs, comme les chiens ou les primates, mais pour de bonnes raisons: les comportements de ces animaux ne peuvent être expliqués sans référence à une forme quelconque d'intentionnalité, et ils sont comme nous faits de chair et de sang. Cependant, une fois informés de la composition et du mode de fonctionnement de ce robot, nous laisserions tomber bien vite la présomption d'intentionnalité que les apparences nous auraient fait antérieurement reconnaître.

5) L'objection de la reconnaissance d'autres esprits: comment faites-vous pour reconnaître que les autres comprennent quelque chose si ce n'est par l'observation de leur comportement extérieur? A ceci, Searle répond:

> "Ce qui importe n'est pas tant *comment* je fais pour reconnaître à d'autres personnes une vie psychique, mais *ce que* je leur attribue à cette occasion. Or, il ne se peut pas que je leur attribue des processus informatiques et leurs résultats puisque ces phénomènes peuvent exister sans vie psychique. Ne nous laissons pas endormir par ces arguments. Les "sciences cognitives" présupposent le caractère réel et connaissable de la vie psychique au même titre que la physique le fait pour les phénomènes de la nature.[21]"

---

[21] *Ibid.*, pp. 420-421.

C'est d'ailleurs cet argument d'inspiration behavioriste qui est à la base du test de Turing dont Searle conteste la validité: en vertu de sa métaphore, en effet, deux systèmes pourraient réussir le test alors qu'un seul des deux serait doté de la capacité de comprendre.

Le point de vue défendu par Searle a ceci de particulier qu'il n'a recours à aucun argument de type spiritualiste. Searle affirme même plus loin n'avoir aucune raison de principe de croire qu'un jour une machine n'aura pas la capacité de comprendre puisque selon lui, notre corps et notre cerveau constituent ce genre de machine; dans la mesure où les mêmes causes peuvent produire les mêmes effets, une entité reproduisant notre structure biologique pourrait fort bien percevoir, apprendre, comprendre, agir, donc faire preuve d'intentionnalité. C'est grâce à notre nature biologique, et non à cause de propriétés purement formelles de notre activité cérébrale, que nous jouissons du privilève d'états de conscience intentionnels. Donc, une machine digitale, peu importe son niveau de sophistication matérielle ou la complexité de son programme, ne pourrait avoir d'autre propriété intentionnelle que celle que lui transmet son concepteur ou son usager. L'intelligence artificielle "forte" est coupable de behaviorisme. Pis encore, ajoute Searle, elle fait preuve de dualisme en prétendant reproduire des

processus mentaux humains sans leur substrat physiologique: quoi de plus cartésien que de couper la pensée des propriétés matérielles du cerveau? Searle conclut en ces termes:

> "Une machine peut-elle penser?" A mon point de vue **seulement** une machine peut penser, mais une espèce spéciale de machine, c'est-à-dire le cerveau et tout appareil doté de la même puissance causale que le cerveau [...]. Bien sûr le cerveau est un ordinateur digital. Puisque tout est ordinateur digital, pourquoi pas le cerveau? Cependant, ce qu'il faut retenir, c'est que le potentiel du cerveau de produire de l'intentionnalité n'est en rien comparable avec l'actualisation d'un programme informatique, puisque de tels programmes peuvent être réalisés sans posséder quelque état psychique que ce soit. Quoi que fasse le cerveau lorsqu'il génère un processus intentionnel, cela ne peut se comparer à l'actualisation d'un programme car aucun programme n'a d'intentionnalité qui lui soit propre[22]."

L'argumentation présentée par Searle est magistrale tant par sa vigueur que par sa clarté, bien que son caractère polémique lui fasse privilégier un point de départ (l'affirmation de Newell et Simon en 1963) qui n'est pas nécessairement le plus représentatif des ambitions présentes des milieux de l'intelligence artificielle[23]. De plus, ce qui

---

[22] *Ibid.*, p. 423.

[23] Schank, en réponse à Searle, reconnaît: "On ne peut dire d'aucun programme que nous avons réalisé jusqu'à ce jour qu'il comprend vraiment. En conséquence, aucun de nos programmes n'explique la faculté de compréhension chez l'homme." Schank, "Understanding Searle", The Behavioral and Brain Sciences, *loc. cit.*, p. 446.

est particulièrement rare chez les philosophes qui contestent la possibilité qu'un ordinateur puisse un jour penser, Searle construit sa critique sur des postulats résolument mécanistes (le cerveau étant une machine) pour dénoncer l'idéalisme de certains théoriciens de l'I.A.: il s'attaque tout particulièrement à cette conception fort répandue dans les milieux de l'intelligence artificielle que le programme est à l'ordinateur ce que l'esprit est au cerveau.

Les réponses à l'article de Searle, provenant pour une grande part de tenants de l'intelligence artificielle "forte", contestent vigoureusement bon nombre de ses affirmations. Ainsi, pour n'en citer que quelques-unes, Bruce Bridgeman[24] répond ce qui suit à l'accusation de dualisme:

> "L'accusation de dualisme portée contre l'I.A. par Searle ne tient pas car les mécanistes, loin de mettre l'accent sur une fonction particulière de l'organisme, affirment plutôt que les processus psychiques ne sont représentés dans un système physique que lorsque celui-ci fonctionne. L'enregistrement d'un programme déposé sur une tablette n'est pas plus conscient qu'un cerveau conservé dans un bocal: affirmer que ce programme, lorsqu'exécuté par un ordinateur, est capable d'intentionnalité revient à dire que la machine adéquate résulte d'une organisation imposée à un substrat matériel. Cette organisation n'est pas plus "psychique" que son substrat[25]."

---

[24] Bruce Bridgeman, "Brains + Programs= minds", The Behavioral and Brain Sciences, *loc. cit.*, p. 427.

[25] *Ibid.*

D'autres reprennent à leur compte l'argument du système critiqué par Searle: Daniel Dennett fait observer qu'en cherchant à capturer le point de vue intérieur d'un agent conscient, Searle cherche trop profondément et lui fait penser à quelqu'un qui voudrait situer l'emplacement de la conscience au niveau des synapses. Or, ce n'est pas du tout à ce niveau de description que se trouve le sujet conscient, affirme Dennett, car la conscience est une propriété du système[26]. De la même façon, Schank critique le niveau d'interprétation adopté par Searle:

> "Je trouve difficile de croire que tout ce que les philosophes recherchent, depuis des siècles, ce sont des explications en termes chimiques des phénomènes qui gouvernent notre vie.
> Car c'est la position à laquelle aboutit Searle. En effet, mises à part les explications chimiques, que reste-t-il? [...] Le cerveau est-il capable de comprendre? Nous, les humains, pouvons sûrement comprendre, mais cette masse de matière appelée cerveau le peut-elle? Tout ce qui s'y passe, ce sont des réactions chimiques et des impulsions électriques, du même ordre que les caractères chinois[27]."

D'autres critiques vont surtout pointer les faiblesses des concepts d'**intentionnalité** et de **puissances causales** utilisés par Searle. Entre autres, Martin Ringle fait

---

[26] Daniel Dennett, "The Milk of Human Intentionality", The Behavioral and Brain Sciences, *loc. cit.*, pp. 428-430.

[27] Schank, *loc. cit.*, pp. 446-447.

observer que malgré ses nombreuses assertions sur le caractère biologique de l'intentionnalité, nulle part Searle ne l'explique: or, c'est ce qui aurait permis de déterminer la nature de ces phénomènes du système nerveux qui ne peuvent en principe être reproduits par une machine[28]. On reproche aussi à Searle sa trop grande discrétion sur la nature des "puissances causales" du cerveau génératrices d'intentionnalité; sans plus de définition, rien n'interdirait d'y inclure, pourquoi pas, les forces occultes ou d'autres entités dont la présence est plutôt inhabituelle au sein de théories mécanistes. C'est ce que tente de démontrer Zenon Pylyshyn dans sa réponse à Searle:

> "Si de façon progressive les cellules de votre cerveau étaient remplacées par des microprocesseurs programmés de façon telle qu'ils accompliraient les mêmes opérations d'entrée-sortie que les cellules qu'ils remplacent, vous pourriez sans doute poursuivre la conversation comme si de rien n'était, tout en cessant éventuellement de **signifier** quelque chose par là. Ce qui, pour un observateur, continuerait d'être des mots, pour vous deviendrait un bruit quelconque provoqué par vos circuits[29]".

---

[28] Martin Ringle, "Mysticism as a Philosophy of Artificial Intelligence", The Behavioral and Brain Sciences, *loc. cit.*, pp. 444-445

[29] Pylyshyn, Zenon W., "The "causal power" of machines", The Behavioral and Brain Sciences, *loc. cit.*, p. 442.

Selon Pylyshyn, on ne peut réduire ainsi l'intentionnalité au genre de matériau (stuff) qui lui sert de substrat sans tomber dans le piège d'une réduction absurde: "Quel est le bon matériau [à la base de l'intentionnalité]? Les agencements de cellules, les neurones individuels, le protoplasme, les molécules protéiques, les atomes de carbone et d'hydrogène, les particules élémentaires? Peu importe le niveau que choisirait Searle, on peut le simuler parfaitement avec la "mauvaise sorte de matériau[30]".

John Haugeland, pour sa part, oppose la métaphore suivante à celle de Searle: supposons un personnage dont le cerveau est défectueux parce que ses neurotransmetteurs ne fonctionnent plus. Heureusement, un "démon de Searle", incroyablement petit et rapide, assume le rôle de messager entre chaque neurone, ce qui permet au cerveau de fonctionner comme s'il était sain. Peut-on, comme devrait le faire logiquement Searle, nier à ce cerveau la capacité d'intentionnalité sous prétexte que des puissances causales naturelles en sont absentes? Pourtant, note Haugeland, aucun neurone n'a la puissance causale appropriée, celle-ci étant assumée par le "démon"; du point de vue de cet auteur, l'erreur principale de Searle est de n'avoir pas distingué **l'intentionnalité originale** de l'intentionnalité **dérivée**: la

---

[30] *Ibid.*

première étant celle qui se manifeste "de l'intérieur" d'un système, à partir des interactions entre ses éléments, alors que l'intentionnalité dérivée désigne ce qui est "emprunté", transmis de l'extérieur. Ainsi, pensée et perception relèvent de l'intentionnalité originale alors que les signes linguistiques (par exemple des mots écrits, un enregistrement de paroles) n'ont que l'intentionnalité que leur prêtent les utilisateurs du langage. C'est cette intentionnalité qui convient aux systèmes actuels d'intelligence artificielle, dans la mesure où les relations entre les différents symboles du système sont purement formelles, "sémantiquement inertes"[21]. Mais jusqu'à quel point les interactions "sémantiquement actives" de l'intentionnalité originale, propres aux relations qu'entretiennent les pensées les unes avec les autres et avec le monde, ne pourraient être reproduites un jour par une machine? Haugeland avoue partager le point de vue négatif de Searle sur cette question, mais pour des raisons bien différentes.

Malgré toutes ses imprécisions, il faut reconnaître que l'argumentation de Searle marque des points contre les théories fonctionnalistes et behavioristes de l'esprit

---

[21] John Haugeland, "Programs, Causal Powers, and Intentionality", The Behavioral and Brain Sciences, *loc. cit.*, pp. 432-3.

humain, qui se contentent généralement d'adopter le point de vue d'un observateur extérieur sans considérer les propriétés intrinsèques des phénomènes observés. Dans sa réponse à une critique, Searle affirme:

> "J'éprouve une difficulté supplémentaire avec le behaviorisme et le fonctionnalisme dans la mesure où je ne peux croire que quelqu'un ait véritablement foi dans ces thèses. Je sais que certains affirment y croire, mais que voulez-vous que je pense lorsque Rachlin [psychologue behavioriste] dit qu'il n'y a pas "d'états psychiques sous-jacents...au comportement" [...]? Il n'y aurait donc pas de douleur derrière les manifestations de souffrance de Rachlin[22]?"

Bien sûr, l'intentionnalité dont Searle fait état n'est qu'un des éléments de la subjectivité: la conscience, l'intuition, l'expérience, les émotions en sont aussi des manifestations qui ne peuvent être ignorées. Nous reviendrons, au chapitre suivant, à l'étude de la problématique soulevée par Searle, mais dans le cadre plus vaste des thèses critiques de Hubert Dreyfus sur les processus cognitifs de l'ordinateur et de l'homme.

---

[22] Searle, en réponse aux objections, *loc. cit.*, p. 454.

CHAPITRE IV

Un point de vue phénoménologique: Hubert L. Dreyfus

Hubert L. Dreyfus, professeur de philosophie à l'Université de Californie (Berkeley), représente sans aucun doute la tendance critique la plus radicale des projets originant du milieu de l'intelligence artificielle. Dès 1961, il publiait une note de discussion[1] dans un compte-rendu de conférences données par des experts du milieu de l'I.A. au Massachussett Institute of Technology; en 1964, Dreyfus réalisait une étude pour le compte de la RAND Corporation, publiée sous le titre Alchemy and Artificial Intelligence[2]: cette recherche, ainsi qu'un article ultérieur, "Why Computers Must Have Bodies in Order to Be Intelligent[3]" serviront de noyau à son principal ouvrage What Computers Can't Do: A Critique of Artificial Reason[4] dont la première

---

[1] Citée par Pamela McCorduck dans Machines Who Think, *op. cit.*, chap. 9.

[2] Cité par P. McCorduck, *op. cit.* Publié sous forme miméographiée en 1964 et imprimé en 1967 dans Rand Paper, p. 3244.

[3] Dreyfus, H.L., "Why Computers Must Have Bodies in Order to Be Intelligent", Review of Metaphysics, 21, 1967, pp. 13-32.

[4] Publié en français sous le titre L'intelligence artificielle: mythes et limites., *op. cit.*

édition a été publiée en 1972. Dreyfus envisage la question de l'intelligence artificielle du point de vue de la phénoménologie mais ne se situe pas toutefois entièrement dans la foulée de Husserl et de la phénoménologie transcendantale, à laquelle il reproche son postulat d'objectivité[5]: il préfère le point de vue de la phénoménologie existentielle (tel que développé par Martin Heidegger, Maurice Merleau-Ponty et Michael Polanyi) dont la réflexion a mis en évidence les phénomènes subjectifs qui accompagnent - et sans doute rendent possible - le comportement intelligent.

La première partie de What Computers Can't Do... recense les principales réalisations de l'I.A. dans les domaines de la traduction, des jeux, de la résolution de problèmes et de la reconnaissance de formes, pour les comparer aux prédictions de Newell et Simon en 1958 (cf. chapitre II). Cette analyse, essentiellement empirique, conduit Dreyfus à identifier des limites radicales aux programmes de l'I.A.; en comparant les performances

---

5 "...in criticizing the assumption that everything essential to intelligent behavior can be understood in terms of the combination of isolable determinate elements, I will be able to adapt arguments used by Husserl and Gurwitsch, who call themselves transcendental phenomenologists. These transcendental phenomenologists, however, as their name is meant to indicate, share the first assumption with the workers in AI: they believe that everything can be understood from the point of view of a detached objective thinker. To claim that this assumption, too, is unjustified is therefore to criticize not only the workers in AI but one branch of phenomenology as well." *Ibid.*, pp. 15-16.

respectives de l'ordinateur et de l'humain, Dreyfus constate chez ce dernier l'existence de capacités virtuellement inaccessibles à l'ordinateur qu'il identifie comme suit: l'aptitude à distinguer l'assentiel de l'accessoire, de tenir compte du contexte, de s'appuyer sur des indices "aux marges du conscient", et finalement de percevoir le type (le paradigme) auquel appartient un être individuel. L'existence de ces facultés exclusivement humaines explique, selon le philosophe, qu'après des débuts souvent spectaculaires, les différentes recherches et expérimentations en I.A. n'ont que fort peu évolué au cours des dernières années; mais alors que les experts en simulation cognitive tendent à réduire ces difficultés à des problèmes d'ordre purement techniques (lenteur et mémoires limitées des ordinateurs), Dreyfus attribue les résultats décevants de l'I.A. à certains présupposés dont ses chercheurs sont prisonniers.

Dans sa recherche pour le compte de la RAND Corporation, dès 1964, Dreyfus identifiait deux de ces présupposés: les postulats épistémologiques et ontologiques, auxquels s'ajouteront dans What Computers Can't Do les postulats biologiques et psychologiques.

> "...affirmer que l'homme fonctionne comme un système de manipulations de symboles revient à admettre quatre postulats distincts:
> 1) Un postulat biologique, d'après lequel [...] le cerveau traite l'information selon des opérations

> discrètes, grâce à quelque équivalent biologique de commutateurs de type oui/non.
> 2) Un postulat psychologique, d'après lequel l'esprit peut être envisagé comme un système opérant sur des éléments binaires d'information, selon des règles formelles [...]
> 3) Un postulat épistémologique, d'après lequel tout savoir peut être explicitement formulé [...]
> 4) [...] le postulat ontologique, d'après lequel tout ce qui existe est un ensemble de faits, dont chacun est logiquement indépendant de tous les autres.[6]".

Or, selon Dreyfus, le bien-fondé de ces postulats n'est en aucune façon corroboré par l'expérience, pas plus que par les arguments *a priori* invoqués par les chercheurs en I.A.

Le postulat biologique:

Les découvertes de la technologie ont souvent servi de point de comparaison pour appréhender les phénomènes du cerveau humain. Ainsi, souligne Dreyfus, "durant la période comprise entre l'invention du relais téléphonique et son apothéose dans le calculateur numérique, le cerveau humain - qui a toujours été appréhendé à la lumière des dernières découvertes de la technique - était généralement représenté sous la forme d'un immense standard téléphonique; après quoi il revêtit les traits d'un calculateur électronique[7]". Cette analogie était rendue possible par l'observation que les

---

[6] *Idem*, L'intelligence artificielle: mythes et limites, pp. 192-193.

[7] *Ibid.*, p. 195.

neurones opèrent de façon semblable aux relais d'ordinateurs, soit à partir de signaux électriques du type tout ou rien (on/off). Cette découverte du caractère binaire des neurones devait renforcer les croyances dans les milieux scientifiques et autres en la similitude du fonctionnement du cerveau et celui des ordinateurs numériques.

Pourtant, affirme Dreyfus, cette croyance n'a pas véritablement d'assises empiriques car malgré la présence d'impulsions électriques dans ses opérations, il n'est pas du tout prouvé que le cerveau traite l'information de façon *numérique*. Citant les recherches de von Neumann sur les ressemblances entre cerveau et ordinateur, et celles de chercheurs du M.I.T. sur la physiologie neurale, Dreyfus tente de démontrer que les processus cérébraux sont plutôt d'ordre *analogique*:

> "Ce n'est pas parce que l'impulsion nerveuse est du type tout ou rien qu'un processus numérique a lieu. La distinction entre numérique et analogique est une distinction d'ordre logique, elle ne prend appui ni sur la configuration de la machine ni sur le genre d'impulsions électriques observées dans le système nerveux. La différence essentielle, en matière d'information, entre un traitement numérique et un traitement analogique, c'est que dans le cas du traitement numérique un élément unique représente un symbole du langage descriptif adopté, et que par conséquent il achemine spécifiquement : un fragment donné d'information; alors que dans un système qui fonctionne à la manière d'un calculateur analogique ce sont des variables physiques continues qui représentent l'information. Le cerveau, qui opère par rafales d'impulsions, ne peut être un

> authentique calculateur numérique que si chaque impulsion correspond à quelque symbole spécifique dans une séquence de traitement de l'information; si en revanche c'est la *cadence* à laquelle les impulsions sont transmises qui se révèle être l'unité de base dans le fonctionnement du système nerveux - et c'est là ce que von Neumann semble penser - alors le cerveau procède à la façon d'un calculateur analogique[8]".

Conséquemment, ces observations tendent à démontrer qu'une analogie trop étroite entre la physiologie neuronale et les processus électriques de l'ordinateur "est une hypothèse empirique qui a fait son temps [...]. Au contraire, étant donné la **différence** observée entre l'organisation cérébrale, fortement interactive, et celle de la machine, à caractère non interactif, il est clair que les faits d'expérience - dans la mesure où sont pertinents les arguments tirés de la biologie - n'incitent guère à penser qu'à partir des calculateurs numériques on produira de l'intelligence[9]".

Le postulat psychologique

Le "postulat psychologique", selon Dreyfus, conçoit le cerveau comme un système qui traite l'information de la même façon que le calculateur numérique, sous forme d'opérations discrètes (élémentaires, discontinues) exécutées conformément

---

[8] *Ibid.*, pp. 197-198.

[9] *Ibid.*, p. 199.

à des programmes heuristiques. Le philosophe rejette évidemment cette conception, et la critique qu'il en fait s'étend aux fondements mêmes de la psychologie scientifique; ainsi, tout en partageant avec la psychologie la conviction que le cerveau, ses entrées, et ses sorties sont matériels, Dreyfus estime que cette science n'est pas parvenue à se définir un objet qui lui soit propre:

> "Et si la psychologie veut se démarquer de la biologie, le psychologue doit être en mesure de décrire le fonctionnement du cerveau humain à quelque autre niveau que celui des réactions physico-chimiques.
>
> La thèse [du postulat psychologique] que nous allons examiner affirme qu'en effet pareil niveau existe (c'est le niveau "traitement de l'information"), et qu'à ce niveau l'esprit humain fait appel à des processus semblables à ceux d'un ordinateur - comparaison, classification, recherches à l'aide de listes, etc. - quand il lui faut adopter une conduite intelligente[10]".

Il est en effet devenu courant, en psychologie, de parler du cerveau comme d'un système de traitement de l'information; la psychologie a même emprunté de nombreux concepts à l'informatique pour formuler de nouvelles hypothèses sur la perception, la mémoire, etc. Mais affirme Dreyfus, cet échange entre les deux disciplines repose sur une équivoque grossière sur la notion d'information: ce terme, qui désigne couramment les messages signifiants émis et reçus par l'humain, est utilisé en cybernétique pour

---

[10] *Ibid.*, p. 201.

nommer tout signal transmis, indépendamment de sa signification. Or, la psychologie, qui reprend les conceptions et le vocabulaire de la théorie cybernétique de l'information, ne tient pas compte de cette importante différence de sens: elle suggère que l'information "signifiante", au même titre que l'information "signal", peut se réduire en éléments isolés et indépendants les uns des autres. Toutefois, en informatique, rappelle Dreyfus, pour qu'une unité d'information (ou *bit*[11]), au sens de la cybernétique, devienne signifiante pour l'humain, il faut l'intervention du programmeur:

> "C'est précisément le rôle du programmeur que d'effectuer la transition entre des propos chargés de sens (qui contiennent de l'information au sens habituel du terme) et ces chaînes binaires sur lesquelles opère un ordinateur, ces suites de bits sans signification qui sont l'information au sens technique du terme. Et rien ne prouve que l'on puisse jamais se passer de ce traducteur humain[12]".

Dreyfus relève une autre confusion entretenue par la psychologie sur le type de "traitement de l'information" effectué par le cerveau; par exemple, une façon pour l'ordinateur d'évaluer la profondeur de champ est d'avoir recours à divers calculs sur les gradients de texture. Or, en utilisant la notion d'information comme moyen terme sans

---

[11] Contraction de l'anglais BInary DigiT.

[12] *Ibid.*, p. 205.

en relever l'ambiguïté, la psychologie commet alors le sophisme suivant: le cerveau qui traite une information "signifiante" fonctionne à la manière d'une calculatrice numérique qui traite un signal. Un exemple de ce genre de raisonnement est relevé chez le psychologue Jerry Fodor lorsque celui-ci affirme que puisque "l'évaluation des distances et de la profondeur de champ [faite par le système nerveux] est causalement déterminée par les gradients de texture", il doit exister dans le cerveau une "procédure de calcul du gradient de texture", et en conséquence, "chaque opération du système nerveux correspond à une certaine suite d'opérations élémentaires[13]". Toutefois, questionne Dreyfus, sur quoi s'appuie-t-on pour ainsi affirmer que le système nerveux, parce qu'il réagit aux variations de gradients de texture, le fait sous forme de calculs mathématiques? Pareil énoncé revient à dire que les planètes se maintiennent en orbite grâce à leur capacité de résoudre des équations différentielles: ce n'est pas parce que des phénomènes peuvent être mis en formules traitables par un calculateur numérique que ces phénomènes sont nécessairement des processus discrets. L'erreur de Fodor et d'autres psychologues, poursuit-il, résulte d'une confusion entre **simulation** et **représentation**; l'ordinateur qui, à partir de

---

[13] Jerry A. Fodor, "The Appeal to Tacit Knowledge in Psychological Explanation", The Journal of Philosophy, vol. 20 (24 oct. 1968), p.629, cité par Dreyfus, *ibid.*, p. 206.

mêmes données que le cerveau, produit les mêmes résultats exécute une **simulation** alors qu'une **représentation** implique des processus comparables. Quel intérêt ces simulations peuvent-elles avoir pour la psychologie puisque l'important n'est pas dans la similitude des résultats, mais dans celle des processus? Pour apporter quelque chose à la psychologie, il faudrait que l'ordinateur puisse exécuter des processus **de type psychologique** analogues à ceux de l'humain. Or, rien ne démontre l'existence de ce genre de processus dans l'ordinateur: au contraire,

> "Aucun fait d'expérience, aucun témoignage extérieur ne viennent confirmer le postulat psychologique. En fait, les expériences mêmes avancées pour tenter de démontrer que l'esprit humain fonctionne comme un calculateur numérique auraient plutôt tendance à démontrer, dès lors qu'on les considère sans poser par avance que ce postulat est indiscutable, qu'empiriquement au moins l'hypothèse n'est pas soutenable[14]".

Cette erreur n'est d'ailleurs pas le propre de la psychologie: elle est "l'héritière d'une vénérable tradition" que Dreyfus fait remonter jusqu'à Platon, soit la conviction présente dès les débuts de la philosophie grecque que l'analyse peut permettre la compréhension du fonctionnement de l'esprit:

> "Kant analysait explicitement toute expérience, y compris la perception, à la lumière

---

[14] *Ibid.*, pp. 215-216.

> d'un ensemble de règles [...] dans le cas de comportement raisonné, pour la démonstration de théorèmes, par exemple, ou pour tout acte moral, Platon était d'avis que, même si les sujets n'étaient pas conscients d'agir selon des règles, leur action n'en avait pas moins une structure rationnelle, qui pouvait être explicitée par le philosophe [...]. Par conséquent, pour Platon, une théorie du comportement rationnel, qui nous permet de comprendre *ce que* donnera ce comportement dans tel ou tel cas, nous *explique* par là également *comment* est produit ce comportement[15]".

Evidemment, tel que mentionné plus haut, Dreyfus ne conteste pas la nature objective de l'être humain et des processus cérébraux: mais ceux-ci, selon lui, ne peuvent être décrits objectivement qu'au niveau neurophysiologique, sous forme de transformations d'énergie. Or, ce genre d'étude ne comporte aucune explication valable pour la psychologie qui s'intéresse plutôt à des phénomènes tels que l'esprit, les intentions, la perception, les souvenirs, etc. Il y a bien la phénoménologie qui leur donne son attention, mais la psychologie refuse d'en reconnaître les descriptions parce qu'elles ne suivent pas une démarche de type algorithmique; elle préfère continuer sa recherche, entre la physiologie et la phénoménologie, d'un niveau d'explication spécifiquement psychologique... en pure perte, estime Dreyfus, car même si la physiologie était un jour en mesure de décrire complètement les transformations effectuées dans le système nerveux au contact de fréquences lors, par exemple, de

---

[15] Ibid., pp. 218-219.

l'audition d'une pièce musicale, celles-ci sont et resteront toujours d'une nature radicalement différente de la mélodie entendue par un sujet. Une bonne partie de la psychologie moderne, tout particulièrement la psychologie cognitive, se voit ainsi remise en cause:

> "Que l'homme soit un objet matériel, transformant des apports d'ordre matériel conformément aux lois physico-chimiques, c'est certain. Mais qu'il soit possible de rendre compte de son comportement comme s'il s'agissait d'un mécanisme de traitement de l'information, transformant des données représentant les traits du monde extérieur, c'est ce dont il est permis de douter quand on considère les difficultés évoquées dans ce chapitre. Rien ne permet de penser - pas plus au niveau de l'expérience qu'à celui de nos connaissances scientifiques actuelles - que l'on pourrait expliquer de la sorte la conduite humaine, puisqu'au niveau physique nous sommes confrontés à des combinaisons sans cesse fluctuantes d'énergie, et au niveau phénoménologique à des objets, qui se situent dans le champ déjà structuré de l'expérience[16]".

Le postulat épistémologique

Moins radical que le postulat psychologique, le présupposé épistémologique se présente ainsi: dans la mesure où les facteurs qui régissent un comportement peuvent s'exprimer sous forme de règles formelles, ces règles peuvent servir à la reproduction de ce même comportement sur ordinateur (même si elles sont différentes des règles effectivement mises en oeuvre par le cerveau). Reprenant un

---

[16] *Ibid.*, p. 234.

exemple du phénoménologue Michael Polanyi, Dreyfus explique que le comportement d'un individu qui roule à bicyclette peut s'exprimer sous forme de lois physiques: ces lois, toutefois, n'expriment que sa **compétence**, i.e. ce qu'il peut faire, et non sa **performance**[17], i.e. la manière dont son cerveau procède pour lui permettre de se maintenir sur le vélo.

C'est ainsi que malgré son caractère modéré, ce postulat partage, avec le postulat psychologique, la même conviction platonicienne que "comprendre une chose, c'est lui donner une expression formelle"; Dreyfus y décèle "la double affirmation: a) que toute conduite délibérée peut recevoir une expression formelle, et b), que cette expression formelle peut servir à reproduire la conduite en question[18]".

Marvin Minsky s'est fait le défenseur de la première partie de cette affirmation en s'appuyant sur l'argumentation de Turing selon laquelle même s'il n'est pas possible de formuler l'ensemble des règles permettant de prévoir ce qu'une personne, en toutes circonstances, **devrait** faire, rien n'interdit de découvrir, derrière ses comportements effectifs, les règles réellement mises en oeuvre. De la même façon, Minsky estime que tout comportement humain peut, du moins en théorie, s'exprimer sous forme d'un ensemble de

---

[17] Traduction de l'anglais **performance**.

[18] *Ibid.*, p. 238.

règles reproductibles par l'ordinateur; mais cette croyance, rétorque Dreyfus, n'est qu'une "généralisation abusive" tirée des succès de la physique et résulte (encore une fois) de l'ambiguïté de l'expression "traitement de l'information".

> "Un calculateur numérique qui résout des équations décrivant les processus d'un système analogique, et qui par là en simule la *fonction*, n'en simule pas pour autant les processus de "traitement de l'information". Car il ne traite pas, en fait, l'information que traiterait pour sa part le système analogique simulé, mais *une information totalement différente*, concernant les propriétés physiques ou chimiques du système analogique en question[19]".

L'erreur de Minsky et de bien d'autres chercheurs de l'I.A. s'exprime dans la conviction qu'un ordinateur est capable de reproduire le comportement humain "non pas en résolvant des équations représentant des processus physiques, mais bien en *traitant des données représentant des faits concernant le monde extérieur, à l'aide d'opérations logiques*[20]". Dreyfus cite Minsky, pour qui:

> "Les processus mentaux ressemblent...aux types de processus que l'on rencontre dans les programmes d'ordinateur: des associations de symboles arbitraires, des schémas de stockage à structure

---

[19] *Ibid.*, p. 244.

[20] *Ibid.*, pp. 244-245.

arborescente, des 'branchements conditionnels, et ainsi de suite[21]".

Cette doctrine, qui selon Dreyfus conduit au "mentalisme", revient à croire en l'existence d'un "niveau mental où s'opèrent des descriptions symboliques susceptibles d'une formalisation numérique[22]". Or, la physique ne permet de décrire rien d'autre - sous forme numérique - que des entrées d'énergie et les opérations neurophysiologiques mises en oeuvre pour leur transformation; de plus, même à ce niveau, toute simulation du cerveau se heurte à un nombre de données tel qu'il est virtuellement impossible d'envisager un système de traitement capable de procéder aux calculs astronomiques nécessaires à pareille entreprise. Une fois levée la confusion entre les lois physiques du comportement humain et les processus de traitement de l'information qui accompagnent ce comportement, Dreyfus conclut que la première affirmation n'a plus guère de fondements.

La deuxième affirmation part du point de vue que l'expression formelle d'une conduite délibérée en permet la reproduction: elle s'appuie tout particulièrement sur les découvertes de Noam Chomsky à l'effet que l'acte

---

[21] Marvin Minsky, "Matter, Mind, and Models", in Semantic Information Processing, p. 429, cité par Dreyfus, *ibid.*, p. 245.

[22] Dreyfus, *op. cit.*, p. 245.

linguistique, apparemment dépourvu de règles précises, se prête cependant bel et bien à une formalisation reproductible sur ordinateur, et ce sans qu'il y ait eu réduction du phénomène au niveau physique. Reprenant sa distinction entre **compétence** et **accomplissement**, Dreyfus note que ce succès de la linguistique s'applique à la compétence de l'acte linguistique, soit à sa structure syntaxique, et non à son accomplissement, i.e. à son utilité (sa pragmatique). Rien ne permet d'affirmer qu'on pourra jamais élaborer une théorie formelle de la pragmatique, car alors que les formalismes de la syntaxe échappent au temps et à l'espace, l'**utilisation** du langage se fait dans le cadre de "situations de la vie réelle, et dans le temps présent - des situations dans lesquelles les objets ont une signification particulière, datée et située[23]". C'est ce qui explique la capacité de l'humain de traiter des phrases non conformes aux règles du langage tout en les comprenant parfaitement alors que l'ordinateur est impuissant dans ce genre de circonstances; pour réaliser une machine capable de s'orienter dans le monde concret,

> "...il faudrait aux chercheurs de l'I.A. pouvoir élaborer une théorie générale - donc universelle et intemporelle - de l'activité humaine vivante, toujours en situation et en perpétuel devenir [...]. Se persuader que l'on parviendra (sous le prétexte que les succès de la linguistique en font

---

[23] *Ibid.*, p. 254.

> foi) à permettre à la machine de comprendre et de parler nos langages naturels, ce n'est pas tant se faire une idée fausse du fonctionnement de notre conscience que mal percevoir combien sont liées intelligence théorique et intelligence pratique [...]. Bref, c'est affirmer, à la manière de Leibniz, qu'il peut exister une "théorie de la pratique[24]".

Mais évidemment comme aucune théorie du genre n'a pu être produite tant par les sciences physiques que comportementales, les machines sont vouées à demeurer encore longtemps "existentiellement stupides", incapables de traiter l'ambiguïté, les exceptions aux règles, et donc, tout particulièrement, de comprendre et de parler une langue naturelle intelligemment.

Le postulat ontologique

L'environnement humain peut être représenté sous la forme de données indépendantes, explicites et parfaitement déterminées: cette thèse de nombreux chercheurs en I.A. correspond à ce que Dreyfus nomme le postulat ontologique. Or, si on accepte cet énoncé à titre d'hypothèse, le problème suivant se pose:

> "...pour saisir un propos, structurer un problème, reconnaître une forme, l'ordinateur, nous l'avons vu, doit sélectionner et interpréter les données dont il dispose en fonction d'un contexte;

---

[24] *Ibid.*, pp. 254-255.

> mais ce contexte, comment allons-nous le transmettre à l'ordinateur[25]?"

Par exemple, de combien de connaissances devra disposer l'ordinateur pour reconnaître un contexte quelconque? Du point de vue du postulat ontologique, il ne pourra s'agir que "d'une énorme masse de données exprimées en menus faits indépendants et isolables[26]". Dreyfus cite à nouveau Minsky, qui, dans un estimé de la quantité de connaissances non spécialisées nécessaires pour qu'une machine fasse preuve de bon sens, parle d'une centaine de milliers de connaissances spécifiques. Tout en s'interrogeant sur les énormes problèmes de stockage et d'organisation des données que pose cette évaluation, Dreyfus va plus loin et en conteste même les fondements. Rien, selon lui, ne démontre que les humains puissent appréhender la réalité sous forme de connaissances discrètes:

> "Identifier un objet comme un siège, c'est saisir sa relation à d'autres objets et aux êtres humains. Ce qui implique tout un contexte de notions sur l'univers vivant des humains, dans lequel entrent notamment la forme de notre corps, cette institution qu'est le mobilier, le fait que nous nous fatiguons, etc., etc., pour n'en citer que quelques-unes. Et tous ces facteurs ne peuvent pas davantage être isolés que la notion de siège elle-même. Tous sont susceptibles de recevoir leur

---

[25] *Ibid.*, pp. 263-264.

[26] *Ibid.*, p. 264.

> signification propre dans le cadre du contexte vivant dont ils font partie[27]".

Une erreur du genre n'est toutefois pas particulière à l'I.A., car elle s'inscrit dans un courant de pensée beaucoup plus large incarné dans rien de moins que la tradition philosophique occidentale dont "le but [...] qui fait à présent corps avec notre culture, consiste à éliminer tout ce qui est source d'incertitude - que ce soit sur le plan moral, intellectuel ou politique[28]". Qu'il s'agisse en effet de la tradition intellectualiste - illustrée entre autres par les substances simples de Leibniz - ou empiriste - pensons à ces éléments ultimes de l'expérience que sont les impressions chez Hume - ou même de "l'atomisme logique" de Russell et Wittgenstein, toutes ces tendances reprennent à leur compte, selon Dreyfus, l'idéal platonicien d'un "monde dans lequel soient garanties la clarté, la certitude et la maîtrise de toutes choses: un monde de structures de données, de théories de la décision, et d'automatisation."[29]

Cette conviction - bien que non conforme à l'expérience - est entretenue grâce aux succès de la physique moderne qui conçoit le monde comme un ensemble d'éléments

---

[27] *Ibid.*, p. 267.

[28] *Ibid.*, p. 268.

[29] *Ibid.*, p. 269.

isolés en interaction. Or, tout en reconnaissant la fécondité du postulat ontologique dans le domaine de la physique, Dreyfus croit que l'appliquer au niveau de l'expérience humaine, c'est confondre *monde* et *univers*, *situation humaine* et *état d'un système physique*. L'I.A. oublie alors un fait capital:

> "...le même agencement concret de matière peut être perçu de quantité de manières, comme autant de situations différentes, selon les buts et les intentions des êtres humains concernés. Par conséquent, bien qu'à un instant donné l'univers ne présente qu'un seul et unique état physique, il peut y avoir autant de situations qu'il y a d'êtres humains[30]".

N'est-ce pas là, rappelle Dreyfus, l'obstacle auquel s'est heurtée l'entreprise de traduction automatique qui a dû reconnaître que les mots n'acquièrent leur sens que grâce au contexte, c'est-à-dire à une situation saisie **dans son ensemble**, et non pas à partir de données particulières prises une à une. Seule la situation permet de déterminer les données **significatives**: c'est en quelque sorte un "sens de la situation" qu'il faudrait incorporer à l'ordinateur pour permettre à celui-ci de lever l'ambiguïté des faits qu'il doit traiter. Comment, en effet, traduire une situation en termes formels? Dreyfus affirme que la difficulté est non seulement technique, mais qu'il s'agit d'une impossibilité de

---

[30] *Ibid.*, p. 271.

principe: la reconnaissance d'un contexte exige que le programmeur donne à la machine les moyens de le reconnaître et pour que les faits ainsi fournis à la machine soient à leur tour dépourvus d'ambiguïté, l'ordinateur devra faire appel à un contexte plus large, et ainsi de suite dans un mouvement *ascendant*. Mais pour éviter une régression à l'infini de contexte large en plus large, l'ordinateur devrait aussi pouvoir reconnaître quels sont les traits pertinents et inaltérables de tout contexte: ce qui l'obligerait à réaliser un mouvement *descendant* vers un "contexte ultime", soit un élément primitif et invariable. Ce dilemne constitue la contradiction de tout effort de créer une intelligence de synthèse.

On pourrait croire qu'il existe une autre voie, si on observe comment l'humain se tire de cette difficulté: celui-ci, au lieu de faire appel à des ensembles hiérarchisés de contextes, semble plutôt reconnaître une situation comme un prolongement de la situation précédente, selon les anticipations développées dans l'instant qui précède la nouvelle situation. Mais il semble que cette voie aussi soit un cul-de-sac pour l'ordinateur: de situation antérieure en situation précédente, on remonte chez l'humain jusqu'aux réflexes initiaux du bébé, attribués au "câblage" de celui-ci. Mais encore là, une nouvelle difficulté surgit du fait que "rien n'explique comment l'enfant, à partir de réactions

invariables à des traits invariables de son environnement, va passer à la faculté de donner un sens aux choses en fonction de leur contexte, faculté dont même les chercheurs de l'I.A. s'accordent à reconnaître qu'elle caractérise l'adulte[21]".

Enfin, le postulat ontologique partage avec les postulats psychologiques et épistémologiques l'idée que l'intelligence agit comme un mécanisme de calculs sur des données brutes d'un univers décomposable en faits élémentaires; c'est cette conception qui explique en fin de compte, selon Dreyfus, la stagnation des travaux en intelligence artificielle après les succès impressionnants de départ. Pour échapper à cette ornière, Dreyfus propose une autre façon d'envisager l'intelligence, qui n'a peut-être pas la force de l'explication scientifique (laquelle se heurte de toute façon à des difficultés insurmontables lorsqu'elle tente de réduire le comportement humain en règles formelles), mais qui comporte l'avantage de décrire les caractéristiques fondamentales de celui-ci: il s'agit de la phénoménologie.

L'I.A., prisonnière de l'objectivité

Conscients de l'impuissance de la démarche traditionnelle de la pensée occidentale, un certain nombre de philosophes - Dreyfus mentionne Wittgenstein, Heidegger,

---

[21] *Ibid.*, p. 285.

Merleau-Ponty, Michael Polanyi - ont recherché une nouvelle voie dans la description de la conduite intelligente; ils ont mis en évidence des éléments qui jouent un rôle-clé dans la détermination du comportement intelligent, et que l'I.A. aurait peut-être intérêt à considérer, soit la responsabilité du corps dans la structuration de l'expérience, l'importance de la situation comme arrière-plan du comportement et enfin, le rôle des intentions et des besoins humains dans l'interprétation d'une situation.

Sur le rôle du corps, Dreyfus fait l'observation suivante:

> "Ses meilleurs résultats, l'informatique les a obtenus dans la simulation des fonctions rationnelles dites supérieures - celles dont on supposait naguère qu'elles étaient l'apanage exclusif de l'homme. Les ordinateurs maîtrisent brillamment les langages artificiels et les relations logiques abstraites. Et ce qui se révèle le plus difficile à simuler sur machine, c'est précisément ce type d'intelligence que nous avons en commun avec le monde animal - la reconnaissance des formes, entre autres exemples...[32]".

Prisonniers d'une seule approche des phénomènes, les chercheurs de l'I.A. - tout en reconnaissant les limites actuelles de leurs expériences - croient qu'un mouvement simple pour l'animal ou pour l'homme, comme par exemple, descendre un escalier, n'est que le résultat d'une somme

---

[32] *Ibid.*, p. 303.

fantastique de calculs effectués simultanément dans différents centres nerveux sur entre autres la pente de l'escalier, la hauteur estimée de chaque marche, sa profondeur, la perspective, le type de mouvement approprié permettant de conserver l'équilibre, etc. Les difficultés dans la reproduction de comportements de ce genre ne seraient que d'ordre technique, à cause des limites des ordinateurs actuels. Or, observe Dreyfus, la seule force de persuasion de pareille hypothèse, c'est qu'elle est, malgré son caractère tout à fait invraisemblable, "la seule en lice[22]" dans les milieux de l'I.A. Pourtant, ajoute-t-il, si dans un geste dépourvu d'intention j'ai touché quelque chose qui m'est agréable, comment expliquer que je peux reproduire sans aucune difficulté ce geste, *intentionnellement* cette fois, et sans recours aux règles formelles qui seraient nécessaires pour décrire les coordonnées de mon premier mouvement?

L'**intelligence du corps** c'est, selon Dreyfus, cette faculté que nous avons de saisir la signification du tout avant celle de ses éléments: citant les travaux de Piaget et de Merleau-Ponty, Dreyfus démontre que la reconnaissance des formes met en jeu une attitude d'anticipation globale et partiellement indéterminée, une *attente*, qui se précise au fur et à mesure du déroulement de l'expérience anticipée,

---

[22] *Ibid.*, p. 323.

dans un mouvement qui va du *tout* vers la *partie*. Toute perception comporte en effet une forme globale (une *gestalt*) accompagnée de certaines indéterminations - telles l'arrière-plan ou fond (l'*horizon extérieur* de Husserl) qui fournit le contexte de l'objet perçu, et les aspects cachés de l'objet (l'*horizon intérieur*): cette capacité de formuler des anticipations globales est liée à l'habileté du corps de répondre comme un tout à son environnement à partir de schémas, parce que "actif, tissé d'interconnexions organiques, notre corps est équipé pour répondre à son environnement grâce au sentiment qu'il entretient en permanence de son propre fonctionnement et de ses objectifs[34]".

C'est ainsi que le monde des humains est partiellement structuré d'avance, selon leurs anticipations et leurs champs d'intérêts. Par conséquent, les données de notre monde, loin d'être brutes, isolées, sont tout au contraire chargées de signification et c'est même cette signification qui leur donne une configuration et une pertinence à l'intérieur d'un contexte quelconque. Là où l'ordinateur doit partir de faits déterminés et indépendants, pour construire dans un mouvement ascendant un ensemble dont le contexte est finalement défini par le programmeur, l'intelligence humaine à l'inverse va

---

[34] *Ibid.*, p. 321.

d'une situation globale plus ou moins déterminée par des anticipations vers des éléments particuliers dont le contexte va déterminer la pertinence et l'intérêt. Enfin, affirme Dreyfus, ce qui permet au corps de se faire une représentation générale de ce qu'il doit entreprendre, ce sont ses **besoins matériels;** le corps qui éprouve un besoin ne connaît pas d'avance ce dont il a besoin, c'est, au gré de sa recherche, ce qui lui apportera une satisfaction qui deviendra le modèle recherché, et cette découverte modifiera, de façon parfois révolutionnaire, la lecture de son environnement. Ce genre de transformation joue non seulement au niveau de la perception, ou de l'expérience existentielle, mais au niveau des "révolutions conceptuelles" dans le domaine des théories scientifiques; Dreyfus invoque les thèses de Thomas Kuhn[25] sur l'importance du *paradigme* dans la détermination des faits en science, qui jouerait un rôle semblable à celui de l'*attente* dans la perception, pour finalement conclure:

> "...c'est ici qu'éclate la différence fondamentale qui sépare l'intelligence de la machine de l'intelligence humaine. L'intelligence artificielle ne peut opérer qu'au niveau de l'objectivité, de la raison - or, à ce niveau-là, les faits ont déjà été produits. Elle extrait ces faits de la situation dans laquelle ils sont organisés, et tente d'utiliser le résultat obtenu pour simuler une conduite intelligente. Mais ces

---

[25] Thomas Kuhn, La structure des révolutions scientifiques, (Chicago, 1962) Flammarion, 1972.

> faits, retirés de leur contexte, ne sont plus qu'une lourde masse fort peu maniable de données neutres, avec laquelle se débattent encore les chercheurs de l'intelligence artificielle [...] ces faits absolus n'existent pas...étant donné que ce sont les humains qui produisent les faits, les faits eux-mêmes sont sujets à révision.
> Enfin, si le philosophe ou le chercheur en intelligence artificielle prétendent surmonter cet obstacle en proposant de donner également une expression formelle aux besoins humains, source de ces changements de contexte, ils se retrouvent confrontés à l'origine même de toutes ces difficultés. On ne peut simuler sur machine désirs, besoins et objectifs, pour la simple raison qu'ils sont indéterminés, alors que la machine ne connaît pour mode d'existence qu'une série d'états toujours parfaitement déterminés[96]."

C'est pourquoi si un jour une intelligence de synthèse doit être réalisée, ce sera d'une façon très différente que ne le suggèrent les expériences actuelles à partir de calculateurs numériques; Dreyfus compare les chercheurs de l'I.A. à ces alchimistes d'autrefois, qui à partir d'expériences parfois positives au point de départ, se sont enlisés par la suite dans des recherches futiles. Or, si on peut aujourd'hui transformer le plomb en or, c'est à partir de méthodes qui n'ont plus rien à voir avec l'alchimie: à moins que l'I.A. ne veuille connaître le même sort, il lui faudra remettre en cause les postulats responsables des résultats décevants obtenus en regard des prédictions optimistes de Simon et Newell en 1958.

---

[96] Dreyfus, *op. cit.*, pp. 364-365.

Objections aux thèses de Dreyfus

Les positions de Dreyfus remettent en cause à la fois la pertinence des travaux actuels de recherche en I.A., la capacité de la psychologie cognitive de se définir un objet d'étude qui lui soit propre, et même en dernière analyse la conviction profonde de la pensée scientifique moderne à l'effet que tout phénomène, incluant l'intelligence, peut être appréhendé objectivement.

Les milieux engagés dans la recherche concrète en intelligence artificielle ont tout d'abord réagi avec une violence explicite peu fréquente dans le monde scientifique. Par exemple, lors de la parution de "Alchemy and Artificial Intelligence", Seymour Papert publia une réplique[37] dont les quelques extraits suivants sont typiques:

> "J'ai de la sympathie pour les "humanistes" qui craignent les effets du développement technique sur notre structure sociale, notre image traditionnelle de nous-mêmes et les valeurs de notre culture [...] Mais l'invasion progressive de l'ordinateur doit être affrontée. C'est céder à la peur que de remplir les départements d'Humanités de "phénoménologues" qui nous assurent que le nombre fini d'états des ordinateurs les empêche à jamais d'avoir accès aux domaines d'activités proprement humains [...]
>
> Le tiers de cet ouvrage n'est composé que de commérages qui n'ont rien à voir avec l'I.A. Quant au reste, si vous prenez une citation quelconque de

---

[37] Seymour Papert, The Artificial Intelligence of Hubert L. Dreyfus: a Budget of Fallacies. AI Memo 154. Cambridge, Mass.: MIT AI Lab., 1968.

> Dreyfus et allez en vérifier la source et le contexte, c'est toujours faux. Et pas seulement parce que Dreyfus n'a pas compris le contexte, car il n'y a aucune exception; ou bien c'est un trait éloquent de Dreyfus, ou bien il est difficile de trouver des exemples de mauvaises choses affirmées par les principaux chercheurs en I.A.[88]".

Un autre expert, Edward Feigenbaum, de Stanford, affirme:

> "Ce dont l'intelligence artificielle a besoin, c'est d'un **bon** Dreyfus... nous avons des problèmes qui pourraient être éclairés par un bon philosophe. Mais Dreyfus nous assomme de matériel mal compris, et dépassé de toute façon; chaque fois que nous le confrontons avec un programme plus intelligent, il nous répond qu'il n'a jamais dit qu'un ordinateur ne pouvait faire cela. Et qu'est-ce qu'il nous offre à la place? De la phénoménologie! De la ouate! De la guimauve![89]".

Au moment où le milieu de l'intelligence artificielle éprouvait de grandes difficultés dix ans après son premier colloque, alors qu'il lui fallait justifier des fonds de recherche importants pour des résultats plutôt minces après les succès du début, la critique de Dreyfus, parue d'abord sous le couvert prestigieux d'un document de la RAND Corporation, faisait mal: ce qui explique que les premières réactions des milieux de l'I.A. ont pris la forme d'attaques contre la personne et de procès d'intention, et qu'encore

---

[88] *Idem*, cité par Pamela McCorduck dans Machines Who Think, pp. 185 et 202.

[89] Cité par McCorduck, *op. cit.*, p. 197.

aujourd'hui, on y manifeste une attitude délibérée de mépris à l'égard de toute référence aux arguments de Dreyfus.

Dreyfus fera toutefois l'objet de critiques plus étoffées de la part des milieux de la psychologie intéressés à la simulation cognitive. C'est ainsi que Margaret Boden[40], qui ne partage toutefois pas les vues du philosophe, fait valoir le fait, contre certaines objections de ce dernier, que certains programmes, bien que de façon encore rudimentaire, réussissent à faire preuve d'attention sélective et de perception globale de situations; concernant l'affirmation qu'il est impossible pour une machine numérique de reproduire la pensée, parce que cette dernière ne procède pas par opérations discrètes, Boden estime que Dreyfus confond le code avec l'information codée: le philosophe lui-même n'utilise-t-il pas les 26 lettres de l'alphabet pour exprimer de l'information "indéterminée et ambiguë"? Sans partager entièrement le scepticisme de Dreyfus, elle lui concède cependant que les enjeux épistémologiques sont encore trop obscurs pour qu'on puisse affirmer avec assurance que tous les aspects de la pensée humaine vont pouvoir être simulés à l'aide d'ordinateurs.

---

[40] Boden, Artificial Intelligence and Natural Man, chap. 11.

Le psychologue Zenon W. Pylyshyn a aussi fourni un ensemble de réponses fort bien argumentées aux critiques de Dreyfus, en s'attardant aux "postulats épistémologiques" du philosophe. Pylyshyn observe qu'en réduisant le monde à deux ordres de réalités, l'une physique et l'autre phénoménale, Dreyfus nie le caractère continu du monde humain et du monde scientifique: en fait, ce "monde humain" ne se livre pas aussi spontanément que Dreyfus aime le croire, il est lui-même rempli d'inférences et de rationalisations à un niveau pré-scientifique[41]. En conséquence, la description phénoménologique ne peut revendiquer plus d'authenticité que celle de la science. Selon Pylyshyn, les phénoménologues comprennent mal le rôle de la science: eux-mêmes cherchent à parvenir directement à l'essence des phénomènes, et les formules scientifiques leur semblent de pauvres simulacres de cette réalité.

> "Il s'agit là d'une incompréhension fondamentale du rôle de l'étude scientifique. Einstein a dit, paraît-il, que le rôle de la science n'est pas de reproduire le goût de la soupe! *Son rôle n'est pas de copier les phénomènes, mais de les rendre accessibles à l'intelligence* [...] Le scientifique doit substituer à la chose réelle un système élaboré à partir de principes qu'il peut comprendre. La "réalité ultime" ne peut être atteinte dans sa totalité ni par la science, ni par la révélation, ni par la poésie ou l'illumination mystique [...] la tâche sans fin du

---

[41] Comme dans le cas d'illusions d'optique dont nous prenons conscience et que nous "corrigeons" sans avoir recours à une interprétation scientifique.

> scientifique est de développer une description qui s'articule à la fois au phénomène (l'évidence de ses sens) et à sa capacité de saisir intellectuellement la description (i.e. aux pouvoir de sa raison)[42]".

Quant à l'analyse faite par Dreyfus du rôle du corps, Pylyshyn lui reproche de confondre le rôle du corps dans la *genèse* de l'intelligence et dans sa *mise en oeuvre effective*. Bien sûr, le corps et son environnement jouent un rôle essentiel dans l'élaboration de structures conceptuelles, mais une fois celles-ci acquises, l'adulte par exemple peut paralyser et conserver toute son intelligence: à ce stade, "l'intelligence d'une personne dépend de la possession d'un corps dans le sens où c'est son corps qui contient les mécanismes où se réalise cette intelligence en lui fournissant les moyens pour la perception, la locomotion, etc.[43]". Dans la mesure où rien en principe n'oblige, dans la simulation d'un état quelconque, qu'on en reproduise la genèse, rien n'interdit non plus en principe qu'on puisse simuler l'état d'un organisme intelligent sans égard aux étapes de son développement.

---

[42] Zenon W. Pylyshyn, "Minds, machines and phenomenology: Some reflections on Dreyfus' "What Computers Can't Do". Cognition, 3, (1974), p. 65.

[43] Ibid., p. 69.

Quant à la contradiction **tout-partie** relevée par Dreyfus dans le processus de la perception, il ne s'agit pas là d'une difficulté propre à la perception, ni à la psychologie. Selon Pylyshyn, ce problème est aussi vieux que la philosophie des sciences elle-même: la perception est déterminée à la fois par un cadre théorique et des catégories conceptuelles, alors que celles-ci simultanément semblent influencées par l'expérience. L'erreur des philosophes, qui se sont divisés entre rationalistes et empiristes à partir de ce problème, est sans doute d'y avoir vu une antinomie, des alternatives qui s'excluent mutuellement; or, note Pylyshyn, il est intéressant de constater qu'en acceptant les deux directions comme complémentaires, les sciences[44], elles, font tout de même des progrès.

Même s'il était vrai que la perception va du tout vers la partie, il ne serait pas possible de la décrire en termes "holistiques": "dans la mesure où un ensemble de percepts a une structure, de sorte qu'un percept ressemble à un autre sous certains aspects et à un troisième sous d'autres, nous ne pouvons appréhender cette qualité autrement que par une description des percepts individuels en termes de relations

---

[44] Y compris l'I.A.: Pylyshyn cite les travaux de Minsky et de Papert sur les structures "hétérarchiques" où le niveau inférieur soumet des éléments à l'examen des niveaux supérieurs qui en retour vont influencer des systèmes à des niveaux encore plus élevés, ou inférieurs, ou reliés, etc.

entre des qualités ou éléments plus simples[45]". Enfin, pour Pylyshyn, il n'est pas possible à partir des arguments de Dreyfus de conclure à l'impossibilité de principe d'une intelligence de synthèse: ses descriptions phénoménologiques peuvent toutes s'expliquer logiquement "de telle façon que leurs aspects fonctionnels sont retenus et que seul est perdu le style poétique[46]". Il n'en considère pas moins Dreyfus comme un "antidote nécessaire" à certaines complaisances de cette jeune discipline qu'est la simulation cognitive en intelligence artificielle.

Pour conclure, la critique de Dreyfus a la force et la faiblesse de toute perspective phénoménologique: au risque d'ignorer la dimension objective des faits, elle met en relief, lorsqu'elle traite de la connaissance, les dimensions proprement subjectives de toute **expérience** généralement méconnues, sinon carrément délaissées, par l'approche scientifique, y compris celle de l'I.A.. D'où la pertinence de son analyse des postulats biologique et psychologique de l'I.A., de même que celle du postulat ontologique dans la mesure où elle met en contraste avec justesse l'**univers** de la physique (et de toute science) avec le **monde** de l'expérience.

---

[45] *Ibid.*, p. 73.

[46] *Ibid.*, p. 76.

L'étude de Dreyfus du postulat épistémologique nous semble cependant plutôt décevante. Tout en contestant avec raison l'assimilation faite par l'I.A. entre les règles d'exécution de programmes informatiques et celles de l'esprit, Dreyfus glisse lui-même vers une position de principe selon laquelle tout dans l'activité de l'esprit n'est pas le fait de règles formelles: c'est ce qui permettrait à l'humain de s'orienter dans le monde concret de la "pragmatique".

Qu'il n'y ait pas de règles **a priori** pour chacun des comportements existentiels est un truisme; mais une étude plus poussés, même phénoménologique, aurait peut-être permis à Dreyfus de réaliser l'aptitude de l'esprit humain à se forger, généralement par analogie avec des expériences antérieures, des règles **ad hoc** très souples, modifiables par essais/erreurs, qui vont lui permettre d'orienter son comportement dans des situations inédites.

Enfin, tout en insistant sur le rôle du corps, Dreyfus aurait pu faire appel de façon plus systématique aux données de la science sur l'intime relation corps-esprit que néglige totalement l'I.A.: en ne privilégiant que l'approche phénoménologique, Dreyfus s'est privé d'un ensemble d'arguments qui tout en convergeant avec ses descriptions,

leur auraient donné une crédibilité susceptible d'ébranler l'assurance un peu orgueilleuse derrière laquelle s'est retranchée l'I.A..

CHAPITRE V

Conclusion

Les résultats actuels ou prévisibles des recherches en intelligence artificielle permettent-ils d'espérer la mise au point d'une entité intelligente de synthèse, dotée d'une vie psychique comparable à celle de l'humain? Les travaux tant théoriques que pratiques de l'I.A. ont-ils permis la réalisation de progrès significatifs sur la connaissance de l'esprit humain?

Au chapitre I, nous énoncions trois obstacles majeurs relativement aux espoirs de l'I.A.: la difficulté de dégager une définition de l'intelligence qui soit vraiment compréhensive, l'absence de critères objectifs permettant avec certitude d'identifier, derrière un comportement, une **présence psychique** et enfin la confusion qui règne autour des notions de **simulation** et de **reproduction** des processus de l'intelligence humaine. L'I.A. a-t-elle relevé ces défis, ou tout au moins s'apprête-t-elle à le faire dans un avenir prévisible?

Définir l'intelligence...

Il serait étonnant, dans le domaine des sciences de la nature, qu'un chercheur réussisse, si ce n'est par accident, la synthèse complète d'une substance naturelle dont il ne connaîtrait pas bien la nature chimique... en pareil cas, il ne pourrait prétendre, jusqu'à plus ample connaissance de la matière originale, qu'avoir produit un succédané imitant certaines propriétés de la substance originale. Evidemment, aucun chercheur en I.A. n'a encore soutenu avoir réalisé un "esprit synthétique": mais la plupart de ceux qui sont engagés dans ce domaine sont convaincus que ce n'est qu'une question de temps, dans la mesure où les recherches sectorielles conduisent à la reproduction de fonctions particulières du cerveau qui pourront un jour être réunies dans un super-ordinateur capable de rivaliser avec l'être humain en termes de créativité, d'intuition, etc.

De la somme des parties jaillira le tout, ou du moins on l'espère. Les milieux de l'I.A. parlent d'ordinateurs qui *voient*, sont dotés de *connaissances*, *raisonnent*, *comprennent* et même *apprennent*. Rarement nous met-on en garde sur le caractère métaphorique des termes utilisés: tout au plus admet-on qu'il n'y a pas nécessairement identité des processus, mais pourquoi faudrait-il qu'il y ait un seul modèle d'intelligence? Chez les humains, toutes les

intelligences utilisent-elles les mêmes processus dans la solution de problèmes?

Il ne se dégage pas, comme en témoignent les chapitres précédents de cette recherche, que l'I.A. ait mieux réussi que la philosophie ou la psychologie à cerner l'essence de l'intelligence: celle-ci, lorsqu'on veut la saisir, se dissocie comme du vif-argent et nous file entre les doigts. Enfin, l'I.A. est peu portée à la réflexion théorique qu'exigerait ce genre d'entreprise, mais surtout elle ne cache pas son scepticisme, considérant l'insuccès relatif des efforts de la philosophie et de la psychologie en ce sens.

Or, comme le fait observer si justement Dreyfus, l'orientation des travaux de l'I.A. révèle la même tendance très marquée de toute la culture occidentale depuis la Grèce antique à réduire la spécificité de l'esprit humain à ses opérations rationnelles. Les fonctions de la pensée qu'on a tenté de faire reproduire par les ordinateurs sont manifestement celles qui ont toujours fasciné les peuples de l'Occident, permis le développement des sciences, mais on doit aussi reconnaître qu'il s'agit des habiletés de l'intellect les plus désincarnées et mécanisables: calcul, opérations de comparaisons d'éléments et de déductions logiques fondées, par exemple, sur les tables de vérité

booléennes, itinéraires de décisions selon des structures arborescentes, structuration de "connaissances", etc. Bien sûr les besoins propres à une société de haute technologie imposaient ce genre d'applications: mais on doit convenir que le "rationnel", comme par hasard, se plie singulièrement mieux aux exigences de formalisation que l'"animal".

Cette conception de l'intelligence qui sous-tend les travaux de l'I.A. reproduit le dualisme corps-esprit auquel est encore confrontée la philosophie occidentale; on retrouve même, dans les réflexions théoriques de l'I.A., le débat éternel entre idéalistes et matérialistes: chez ces derniers, par exemple, il y a les **mécanistes** (Minsky, Papert, Hofstadter, etc.) selon qui une meilleure connaissance des processus physiologiques neuronaux ou encore la mise au point de programmes se concentrant sur l'analyse et la reproduction de processus particuliers (représentations des connaissances, procédures pour l'utilisation spécifique de ces connaissances) conduira tôt ou tard, par complexification progressive, à l'intelligence de synthèse; de leur côté, les **mentalistes** (Schank, Winograd, Wilensky...) méfiants à l'égard du réductionnisme[1] de leurs collègues et selon qui

---

[1] Ainsi, selon Terry Winograd, bien que l'esprit humain soit un système **physique** de manipulation de symboles, "it is both possible and revealing to study the properties of physical symbol systems at a level of analysis abstracted from the physical details of how individual symbols and structures are embodied, and the physical mechanisms by which

l'essence de l'esprit est dans sa programmation, tentent plutôt de formaliser en langage informatique la connaissance humaine, de reproduire, non de mimer, les processus mentaux.

Si on compare d'ailleurs le champ d'expérimentation de l'I.A. avec celui de son ancêtre, la cybernétique, on ne peut faire autrement que réaliser qu'il y a eu un formidable rétrécissement de perspective: alors que cette dernière recherchait un paradigme commun à tout système général, non-vivant, vivant ou pensant, l'I.A. concentre ses efforts exclusivement sur les aspects les plus formels et procéduraux de l'intelligence, laissant presqu'entièrement le soin à la neurophysiologie d'établir la liaison entre ces procédures et les mécanismes physiologiques qui sont à leur origine.

Compte tenu de ses orientations présentes, rien donc ne laisse présager que l'I.A. contribue, du moins dans sa réflexion théorique, à produire une nouvelle définition unifiant les aspects fort variés de la pensée humaine mis en évidence jusqu'ici par la philosophie, la psychologie et la neurophysiologie.

---

the processes operate on them." ("Towards a Procedural Understanding of Semantics", Revue Internationale de Philosophie, #117-118, 1976, p. 264.)

Présence psychique et intelligence

Lorsque Pascal mit au point la première calculatrice mécanique, il cessa de considérer le calcul comme propriété spécifique de l'esprit humain, réservant à celui-ci la volonté . Depuis la mise au point d'ordinateurs qui "perçoivent" leur environnement, capables d'induction et de déduction, d'apprentissage, de résolution de problèmes, de dialogue avec un être humain (dans un registre li ité), de simulation de sentiments, de découverte de nouveaux théorèmes, de choix, etc., que nous reste-t-il si, comme Pascal, nous éliminons d'une définition de l'esprit tout ce qui peut être reproduit mécaniquement? Il nous reste alors le "noyau" résiduel jusqu'ici irréductible à l'entreprise scientifique, le monde de la **subjectivité** auquel renvoient les notions d'âme, de conscience, d'intentionnalité. A ce propos, les visions qu'entretient l'I.A. sur cette réalité fantômatique sont partagées, selon qu'elles accordent ou non une reconnaissance au caractère "objectif" de la conscience.

Ceux qui lui reconnaissent un caractère réel penchent vers une conception mécaniste de la conscience: ainsi, pour Hofstadter, cette faculté est comparable en termes d'informatique à un sous-système en constante interaction

avec d'autres sous-systèmes du cerveau et le monde extérieur[2], dont la fonction de "réflection" d'une multiplicité de sous-niveaux engendre une sorte de "résonnance" appelée la conscience de soi. C'est dans des termes semblables, mais plus près de la physiologie, que le neurophysiologue Ernest Kent[3] envisage l'existence de la conscience comme **propriété émergente** d'un système complexe: il serait concevable qu'une machine construite selon les standards d'aujourd'hui puisse avoir un comportement intelligent et dialoguer raisonnablement, mais selon lui, cela ne garantit en rien une présence psychique, des états mentaux, une subjectivité, dont l'existence exige sans doute un système d'une complexité qui défie encore l'imagination.

Mais le plus grand nombre préfère, dans la foulée du behaviorisme et du test de Turing, considérer la question comme non pertinente, comme en témoignent les quelques extraits suivants:

> "La pensée n'est pas un principe abstrait existant par lui-même en dehors des manifestations que nous pouvons appréhender [...] Il n'est jamais donné de pouvoir observer directement ce que le langage commun nomme la pensée [...] Ce concept de pensée apparaît, dès l'abord, mal adapté aux

---

[2] A ce sujet, voir Hofstadter, Godel,Escher..., pp.383-388, et pp. 708-710.

[3] KENT, ERNEST W., The Brains of Men and Machines, McGraw-Hill Publications Co., N.Y., 1981, dernier chapitre.

phénomènes qu'il est censé recouvrir ou impliquer. Issu de la philosophie traditionnelle, nourri de l'intimité des pulsions psychiques, fortifié d'habitudes invétérées d'introspection, il porte les marques de ses origines: un idéalisme inavoué, une hétérogénéité sans espoir, une imprécision de mauvais aloi.[4]"

Ou encore:

"La conscience pourrait-elle naître dans de pareilles machines? [...] il n'y a pas actuellement de réponse possible. Conscients que nous sommes de notre pensée propre, nous admettons que la pensée existe également chez nos semblables, en leur accordant une sorte de "bénéfice de l'analogie". Mais nous répugnons généralement à accorder ce même bénéfice à une machine. Il n'y a cependant aucune expérience au monde qui puisse prouver que l'homme "pense" réellement. Tout ce qu'on peut observer, c'est qu'il se comporte "comme s'il pensait". Et le cybernéticien est bien forcé d'admettre que, si une machine offre toutes les manifestations d'une conscience naissante, cette conscience existe effectivement ou - ce qui du point de vue pratique est équivalent, - que tout se passe "comme si" cette conscience existait." [5]

Schank et Weizenbaum ont un point de vue comparable. Le premier affirme ce qui suit:

"Nous faisons face à un dilemne. Il serait préférable de juger de la capacité de comprendre d'un système par autre chose que ses sorties (*output*). Malheureusement, il ne faut attendre rien d'autre que des sorties. En dépit du fait que toute tentative de juger de l'intelligence d'une entité soit incertaine, il se peut bien que l'évaluation de leurs sorties représente la méthode la plus

---

[4] Josse Lemaire, "Variations sur la pensée cybernétique", dans Le dossier de la cybernétique, Marabout Université, Paris, 1968, p. 67.

[5] "Qu'est-ce que la cybernétique?", Georges R. Boulanger, dans Le dossier de la cybernétique, *op. cit.*, p. 26.

> efficace pour évaluer le degré de compréhension des humains aussi bien que des ordinateurs. Après tout, tout système, humain ou mécanique, est jugé d'après ses résultats[6]."

Et selon l'auteur d'ELIZA:

> "Einstein nous enseigne que l'idée de mouvement est sans signification en elle-même, qu'on ne peut raisonnablement parler du mouvement d'un corps que par rapport à un certain cadre de référence et non d'un quelconque mouvement ABSOLU [...] Il en est de même pour l'intelligence. L'intelligence par elle-même est un concept sans signification. Elle exige, pour devenir significative, un cadre de référence ainsi que la spécification d'un domaine de pensée et d'action[7]."

La plupart des chercheurs en I.A. n'aiment pas spéculer sur ces fonctions insaisissables de la pensée que sont la conscience, la subjectivité ou en termes plus modernes l'intentionnalité. La question d'une présence psychique derrière des manifestations observables est au pire considérée sans fondements, au mieux comme étant insoluble dans la mesure où on préfère aborder l'esprit humain comme une "boîte noire" dont seules les entrées/sorties peuvent faire l'objet d'une étude rigoureuse; ainsi, pour les défenseurs de ce point de vue, l'ordinateur représente au contraire la "boîte blanche" (un système transparent) qui, à

---

[6] SCHANK, Roger C., The Cognitive Computer:..., *op. cit.*, p. 55.

[7] WEIZENBAUM, Joseph, Puissance de l'Ordinateur et Raison..., *op. cit.*, p. 135.

partir d'entrées semblables à celles que reçoit le cerveau, produit des sorties comparables. La tentation est alors très forte - certains y succombent volontiers - de croire que de la connaissance des processus de la boîte blanche on puisse déduire ceux de la noire.

La preuve du contraire n'est par définition pas facile à faire. Les ordinateurs sont produits et programmés par l'intelligence humaine: dans la solution de nombreux problèmes relevant ou non de la science cognitive, mais plus spécialement dans celle-ci, il est tout à fait fréquent que le programmeur approfondisse sa recherche en étudiant la démarche intellectuelle utilisée par l'humain vis-à-vis ce genre de problèmes. Mais il est tout aussi vrai qu'à cause des particularités techniques de l'ordinateur ou de celles des langages de programmation, on puisse obtenir, pour telle opération, des résultats comparables sinon supérieurs à ceux de l'esprit humain par des procédures tout à fait différentes: c'est le cas, entre autres, des programmes de jeux d'échecs, de reconnaissance de formes, etc.

De la même façon que la "boîte blanche" qui joue aux échecs ne le fait de toute évidence pas selon les mêmes processus que la "boîte noire" humaine, on ne pourrait soutenir ce genre d'analogie devant un système plus complexe

qui, à partir d'entrées semblables, mimerait les sorties de l'esprit humain.

C'est pourtant la position soutenue par Turing et ses successeurs lorsqu'ils affirment que nous n'avons pas le choix de reconnaître la "pensée" (ou conscience) à une machine qui en donne tous les signes extérieurs, de la même façon que nous le faisons pour les êtres humains; et à moins de se cantonner sur un point de vue solipsiste, il semble difficile de réfuter cet argument.

A ce sujet, la philosophie a peut-être besoin d'un humble auxiliaire - le bon sens - qu'elle néglige souvent parce qu'il est à la source de nombreuses erreurs dont seule la raison est venue à bout à travers l'histoire. Or, le behaviorisme de l'argumentation de Turing aussi bien que le point de vue opposé qui aboutit au solipsisme heurtent tous les deux notre bon sens. Arrêtons-nous pour nous demander pourquoi.

Il est bien connu qu'une démarche purement rationnelle peut conduire à des conclusions parfaitement absurdes, de même qu'une approche trop empirique ne conduit nulle part. Tout en étant une forme de pensée très concrète, le bon sens (ou sens commun) n'en constitue pas moins une première synthèse d'expériences particulières: il comporte alors une logique qui, bien que n'ayant pas la pureté et le degré de

généralisation correspondant aux exigences de la science, maintient le contact avec le réel dans toute sa variété changeante à l'échelle de l'expérience humaine. Bien que la physique nous apprenne que la matière en apparence opaque et solide est en réalité constituée surtout de vide, le bon sens nous retiendra de vouloir passer à travers le mur.

Ne se pourrait-il pas que ce bon sens, cette faculté dite "heuristique" par l'I.A. elle-même, en considérant un plus grand nombre de variables que certaines positions en apparence rationnelles, mais doctrinaires et unilatérales sur le fond, prenne mieux en ligne de compte toute la réalité dans sa complexité? Par exemple, il est vrai que nous reconnaissons la conscience à un autre être humain par analogie avec nous-même, et donc en tenant compte en partie de ses comportements; mais l'expérience courante nous montre que des états subjectifs peuvent être simulés (un acteur peut simuler la colère sans l'éprouver réellement): on peut donc dissocier un état subjectif d'un comportement. Par ailleurs, à de multiples occasions il a été démontré que la conscience peut être active chez un humain sans que celui-ci ne puisse le manifester extérieurement (dans certains types de paralysie ou de comas, par exemple). Notre bon sens ne se fie donc pas seulement aux comportements pour juger de la présence d'une vie psychique chez l'autre, mais tient compte d'autres facteurs dont le plus important est sans aucun doute

la ressemblance de morphologie et de structure que l'autre présente avec nous-même.

Ainsi, il y a fort peu de chances que l'on s'interroge sur les états mentaux d'une machine automatique quelconque car derrière ses comportements nous ne percevons pas de vie organique. Quant aux êtres vivants, par exemple en éthologie animale, on ne s'est sans doute pas fondé sur le seul fait que le chimpanzé se montrait du doigt devant un miroir pour conclure qu'il témoignait d'une conscience (même rudimentaire) de soi, car ce geste aurait pu être le résultat d'un simple conditionnement: ce qui a rendu cette conclusion plausible, c'est la corrélation de ce comportement du chimpanzé avec la similitude de son organisation cérébrale jusqu'à un certain point comparable à celle de l'être humain.

Ceci dit, les mystères relatifs à la nature profonde de la conscience demeurent. Toutefois, il semble bien raisonnable de conclure de ce qui précède que, dans la mesure où un système de structure à peu près comparable avec un autre (en l'occurence le mien), produit, pour les mêmes entrées, des sorties du même genre, les processus qui s'y déroulent, incluant les états subjectifs, sont vraisemblablement de même nature que ceux que je ressens.

Les systèmes mécaniques ou électroniques ne se prêtent malheureusement pas au genre d'analogie établie par Turing,

même d'un strict point de vue scientifique. Acceptons, par exemple, l'hypothèse (fort imprécise, mais tout de même largement partagée) que les phénomènes subjectifs - volonté, conscience, intentionnalité - constituent des **propriétés émergentes** d'un système nerveux central composés de plusieurs dizaines de milliards de neurones dont l'agencement et le fonctionnement nous sont encore peu connus. Avançons un autre postulat aussi largement admis en science , soit qu'il y a une relation causale entre la **complexité** d'un système nerveux et son **potentiel psychique:** les ordinateurs actuels - et même éventuels - ainsi que les techniques de programmation nous apparaissent d'une simplicité inouïe comparés, par exemple, à la structure connue d'un simple neurone. De plus, une propriété émergente n'est pas le seul résultat d'une quelconque complexité, mais d'un agencement particulier d'éléments ayant des propriétés bien spécifiques (mal connues dans le cas de la cellule vivante, encore moins à l'échelle du système nerveux central): enfin, la structure des ordinateurs actuels repose sur des principes et des éléments fort différents de ceux des êtres biologiques. Rien de ces observations ne laisse présager qu'une complexification de composantes électroniques, si grande soit-elle, conduise à des propriétés émergentes du même ordre que celles manifestées par le système nerveux.

En conséquence, en dépit de notre grande ignorance des phénomènes psychiques qu'éprouvent l'humain et certaines espèces animales, rien n'indique que l'I.A., dans les développements prévisibles de la technologie (programmation d'ordinateurs électroniques), soit engagée sur la voie de construction d'une entité capable d'états mentaux semblables à ceux que nous éprouvons.

Simulation ou reproduction?

Un des usages les plus spectaculaires de l'ordinateur est sans doute la **simulation**, i.e. la représentation du comportement de systèmes physiques modélisés: qu'il s'agisse de la résistance d'un édifice aux vents, de la réponse du marché boursier à une baisse des taux d'intérêt ou de la réaction d'une protéine à tel enzyme, tout système formalisable peut être traduit en données traitables par ordinateur.

Les comportements humains, dans la mesure où ils sont formalisables, ne font pas exception: on a pu ainsi simuler des désordres psychologiques (le programme PARRY simule la paranoïa), des processus du raisonnement (dans les systèmes-experts), certains mécanismes mis en jeu lors de l'analyse d'images, la réaction d'un politicien de droite à certaines

situations (le programme CYRUS) et même faire manifester par l'ordinateur des réactions apparemment émotives.

A des degrés divers, ces programmes produisent des réactions semblables à celles des humains. Dans certains cas, leurs auteurs ont cherché délibérément, par introspection, observations, etc., à mimer la procédure suivie par l'esprit humain; dans d'autres, on s'est contenté de faire reproduire, par un système donné, des résultats comparables à ceux produits par l'humain, sans vraiment se soucier d'en reproduire les processus.

Pour éviter toute confusion, nous reprendrons la distinction établie par Dreyfus (voir chapitre précédent) et réserverons le nom de **reproduction** aux premiers cas, c'est-à-dire ceux où l'expérience n'a pas pour seul objectif d'arriver aux mêmes résultats que l'esprit humain, mais d'y arriver à partir de processus similaires, le terme **simulation** convenant mieux aux cas du deuxième type.

Notons toutefois que dans l'un et l'autre cas, l'ordinateur opère sur des **modèles** qui ne sont, après tout, que des réductions, c'est-à-dire les produits d'une abstraction; ce sont ces objets qui néammoins permettent de faire des analogies entre des systèmes appartenant à deux mondes séparés - l'un de nature biologique, l'autre électronique.

Aux fins de la problématique qui nous intéresse, nous laisserons de côté les modèles **minimaux** que sont les simulations pour nous concentrer sur ces modèles plus significatifs qui prétendent **reproduire** des opérations intelligentes. Ces modèles, comme tous les autres, résultent nécessairement d'une distinction, parmi les données sensibles de l'expérience, entre l'essentiel et l'accessoire, et procèdent donc d'un schéma plus ou moins simplificateur. Ainsi, certains modèles peuvent être qualifiés d'**analogiques** dans la mesure où ils reproduisent des propriétés homologues à celles du phénomène étudié (ainsi une soufflerie produira des effets semblables à ceux du vent, l'étude des mécanismes de la pompe fera comprendre certaines propriétés du muscle cardiaque, le Perceptron (voir chapitre II) reproduisait certains mécanismes de la perception, etc.).

Or, l'informatique a recours à des modèles différents, dits **numériques** lorsque le schéma représenté par le modèle est formé d'expressions mathématiques (fréquent dans les sciences de la nature) ou **algorithmiques**[a] lorsque le modèle a la forme d'un organigramme définissant un ensemble de relations entre différents éléments d'un système et un "initinéraire" de procédures.

---

[a] Au sens large, et non par opposition à "heuristique".

L'intelligence artificielle expérimente surtout sur ce dernier genre de modèles; or, y retrouve-t-on des propriétés suffisantes pour parler d'une **identité substantielle** entre ces modèles et la pensée humaine? Evidemment non, dans la mesure où ces modèles opèrent sur un substrat (électronique) de nature fort différente de celui de l'esprit humain. Tout au plus pouvons-nous reconnaître une similitude de structure - un isomorphisme - entre les processus **logiques** suivis par ces deux ordres de systèmes. Dans certains cas, la ressemblance peut être étonnante, car l'ordinateur se prête avec une extrême souplesse à tous les itinéraires algorithmiques qu'on veut bien lui faire suivre.

Toutefois, les modèles ainsi obtenus ne se superposent qu'aux procédures d'**un seul niveau** de l'esprit, n'ont rien à voir avec les processus neurophysiologiques qui les sous-tendent, et en conséquence la similitude ne vaut qu'au niveau **fonctionnel.** Evidemment, rien n'interdit en principe qu'on puisse éventuellement modéliser à l'échelle moléculaire les réactions neurophysiologiques et y superposer des modèles pour l'ensemble des opérations de l'esprit humain: un tel super-modèle n'en serait toutefois qu'un simulacre, même si chacun des sous-modèles qui le composent reproduisait fidèlement les processus logiques de l'esprit humain. En effet, cette **reproduction** du fonctionnement logique ne serait pas le produit de réactions physiologiques, ces dernières

n'existant qu'à l'état de **simulation** dans le super-modèle. De la même façon qu'il est impensable qu'un modèle de neurone éprouve de la douleur (même s'il en simule les effets), ou qu'un modèle de la molécule d'ADN se reproduise effectivement (même s'il simule toutes les étapes de cette opération), rien ne nous autorise à penser qu'un modèle informatique de la pensée humaine - même dans sa totalité - puisse jamais faire autre chose qu'en simuler le fonctionnement - et ce, sans éprouver les états mentaux qui accompagnent la pensée chez l'être humain.

Margaret Boden, qui ne peut être accusée de scepticisme outrancier à l'égard de l'I.A., et selon qui les modèles de l'I.A. ont un intérêt considérable pour toute science cognitive, fait tout de même une sérieuse mise en garde contre des analogies trop serrées:

> "La question importante sur laquelle achoppe (et achoppera toujours) l'analogie avec l'ordinateur relève du caractère artificiel de celui-ci. Parce que les ordinateurs ne peuvent avoir de buts **de la même façon** que l'on affirme que les humains (et les animaux) poursuivent des buts, les concepts courants de la psychologie qui parlent de but, de connaissance, d'intelligence, de langage, de communication.... ne peuvent s'appliquer à aucun ordinateur imaginable [...] En dernière analyse, les ordinateurs ne sont pas des

produits humains du même type que les bébés, et sont dépourvus de tout but intrinsèque.[9]"

Est-ce à dire que le rêve millénaire de l'être humain de créer **artificiellement** une entité dotée des mêmes pouvoirs psychiques que les siens est à jamais inaccessible? Pas du tout, dans la mesure où le système nerveux humain est une entité matérielle, obéissant à des lois naturelles, et donc en principe connaissable et reproductible. Comme rien n'interdit **a priori** aux sciences de la vie l'élaboration d'une forme synthétique de vie, la perspective d'un humain (ou analogue) de synthèse, bien que fort lointaine, demeure dans l'horizon du possible.

Quand aux espoirs entretenus sur un quelconque potentiel psychique de l'ordinateur, nous partageons le scepticisme ou l'incrédulité manifestés par ces chercheurs de différents milieux tels Dreyfus, Searle, Kent et Boden: rien ne permet, dans les limites des connaissances actuelles, de supposer que l'esprit, en tant que propriété d'un type de système biologique, puisse être **reproduit** dans sa totalité sur un système non vivant. En dépit de grandes différences qualitatives et quantitatives, il y a, croyons-nous, continuité entre la sensibilité à l'environnement manifestée

---

[9] Margaret A. Boden, Minds and Mechanisms: philosophical psychology and computative models. Harvester Press, 1981, pp. 86-87.

par les créatures monocellulaires dont nous sommes issus et les états subjectifs que nous percevons au sein même de notre système nerveux central. La croyance de nombreux experts de l'I.A. que la conscience n'est qu'une simple représentation interne de soi, qu'elle peut résulter comme par magie de l'interaction de divers programmes et sous-programmes, bref qu'une reproduction purement formelle des opérations de l'esprit humain se mette à **penser** au plein sens du terme relève d'une simplification outrancière que le développement scientifique révélera peut-être plus clairement un jour.

Soulignons que les états subjectifs chez l'humain ne sont pas le produit de **données pures**, mais qu'ils sont le produit d'une genèse complexe fondée sur des **expériences** entre la structure biologique et neurophysiologique de l'humain et son environnement; dans cette optique, le **mentalisme** - cette position plus ou moins explicite de l'I.A. à l'effet que les processus de l'esprit sont entièrement réductibles à des programmes - ne nous semble rien d'autre qu'une manifestation nouvelle de l'idéalisme philosophique, version I.A.

Ce qui ne signifie pas que les recherches de l'I.A. sont dépourvues de sens, au contraire: plusieurs des modèles construits par l'I.A. ont fait mieux comprendre de nombreux

processus de la pensée humaine[10], tout en permettant, comme l'ont fait le boulier de l'Antiquité, puis la calculatrice de Pascal, et enfin les premiers ordinateurs, d'élaborer des processus infiniment plus rapides et efficaces que ceux de l'esprit humain: ce dernier repose sur des mécanismes électro-chimiques particulièrement lents, sans compter les nombreux avatars que lui imposent les inévitables considérations affectives et émotives, les idéologies, les préjugés... prix qu'il nous faut payer pour avoir accès à la subjectivité.

Il est donc prévisible que les échecs de l'I.A. "forte" pour qui l'ordinateur est un moyen de reproduction de l'intelligence humaine vont conduire au renforcement des recherches de l'I.A. "faible" sur les processus de la pensée. L'ordinateur constitue ainsi un outil privilégié pour mieux comprendre des fonctions de l'esprit jusqu'alors surtout accessibles à l'introspection: cette dernière voie ne permet généralement qu'une perception très grossière et approximative des sentiers parcourus par l'esprit lors d'une opération intellectuelle quelconque. Mais la simulation de ce processus sur ordinateur permet l'élaboration de modèles supérieurs en complexité et en testabilité susceptibles de fournir des indices inédits sur le fonctionnement de l'esprit

---

[10] Voir à ce sujet les travaux de Margaret Boden, particulièrement dans Minds and Mechanisms, *op. cit.*

humain. Par exemple, dans les années soixante, les chercheurs en I.A. avaient tendance à assimiler intelligence et langage: c'est ainsi que les premiers programmes de traduction automatique étaient surtout fondés sur l'analyse syntaxique. Leur échec a confirmé la prédominance du concept sur le mot, mettant en évidence des opérations d'une grande complexité derrière toute communication linguistique. Toutes les recherches actuelles en intelligence artificielle convergent dans la même direction, soit la **représentation de connaissances:** c'est en tentant de reconstruire sur un appareillage objectif des reconstructions mentales du monde qu'on peut étudier et mieux comprendre, même si ce n'est que de façon analogique, les opérations de l'esprit.

L'ordinateur jouera un rôle de plus en plus important comme prolongement **instrumental** de l'esprit. Son existence et son utilisation vont modifier l'environnement humain et la perception que l'homme a de lui-même. Mais le rêve (ou la crainte) d'une machine non biologique dotée d'intériorité, de projets et d'intentions qui lui sont propres n'est que peu défendable à la lumière des réalisations actuelles et prévisibles de l'I.A.

APPENDICE

Texte original de nos traductions

CHAPITRE I:

#12, p. 18: "The literature concerned with computers and robots quite properly dismisses that question as so loose and ambiguous that there is no way of deciding what the correct answer is. In its use by psychologists, neurologists, computer technologists, and philosophers, the word "think" has so many meanings in its application to animals, men, and machines, that if anyone asks, "Can animals think?" the answer must be "Yes" - in some senses of the word, and probably also "No" - in other senses of the word. No one has yet produced an acceptable definition of human thinking in all its variety that will serve as a standard for measuring the success of efforts to produce a robot that will simulate the whole range of human thinking." (Adler)

#14, p.21: "Today, AI researchers have started to ask questions about the real nature of human intelligence. We are trying to define precisely what it means for humans to understand, to learn, to think, and to change over time. This seemingly simple shift in approach actually makes a profound difference for the definition of AI and for the course of AI research. Rather that concentrating on a specific task, such as getting a computer to play chess, we are addressing the essential theoretical and philosophical questions of human intelligence." (Schank)

CHAPITRE II:

#2, p. 27: "By "artificial intelligence", I therefore mean the use of computer programs and programming techniques to cast light on the principles of intelligence in general and human thought in particular." (Boden)

#3, pp. 27-28: "The scientists who work in the field called artificial intelligence (AI) try to develop theories about the specific knowledge needed to solve problems and about the

processes that use more general knowledge to put specific knowledge to use [...] With philosophy, we share problems about mind, thought, reason, and feeling." (Minsky)

#17, p. 41: "...these programs respond to, rather that understand, language [...] Many of them can therefore deal with only one area of discourse - unlike men and women who can talk about baseball and algebra, psychology and genealogy [...] Instead of making intelligent interpretative use of a background system of knowledge ranging over widely varied semantic domains, these programs employ relatively isolated and inflexible rules to determine their verbal response to linguistic input. In short, such "speaking machines" do not behave like someone conversing in her native language. Rather, they resemble a person resorting to trickery and semantic sleight-of-hand in order to hide her lack of understanding of a foreign tongue." (Boden)

#25, p. 49: "Expert systems, while potentially useful, are not a theoretical advance in our goal of creating an intelligent machine. Real intelligence demands the ability to learn, to reason from experience, to **shoot from the hip**, to use general knowledge, to make inferences using gut-level intuition. Expert systems can do none of these. They don't improve as a result of experience. They just move on to the next if/then rule." (Schank)

CHAPITRE III

#2, p. 57: "In attempting to construct such machines we should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates." (Turing)

#7, p. 61: "Gödel's theorem must apply to cybenernetical machines, because it is of the essence of being a machine, that it should be a concrete instantiation of a formal system. It follows that given any machine which is consistent and capable of doing simple arithmetic, there is a formula which it is incapable of producing as being true - i.e. the formula is unprovable-in-the-system - but which we can see to be true. It follows that no machine can be a complete or adequate model of the mind, that minds are essentially different from machines." (Lucas)

#8, pp. 62-63: "[...] the concept of a conscious being is, implicitely, realized to be different from that of an unconscious objet. In saying that a conscious being knows

something, we are saying not only that he knows it, but that he knows that he knows it, and that he knows that he knows that he knows it, and so on, as long as we care to pose the question: there is, we recognize, an infinity here [...] Although conscious beings have the power of going on, we do not wish to exhibit this simply as a succession of tasks they are able to perform, nor do we see the mind as an infinite sequence of selves and super-selves and super-super-selves. Rather, we insist that a conscious being is a unity, and though we talk about parts of the mind, we do so only as a metaphor, and will not allow it to be taken literally.

The paradoxes of consciousness arise because a conscious being can be aware of itself, as well as of other things, and yet cannot really be construed as being divisible into parts. It means that a conscious being can deal with Gödelian questions in a way in which a machine cannot, because a conscious being can both consider itself and its performance and yet not be other than that which did the performance. A machine can be made in a manner of speaking to "consider" its own performance, but it cannot take this "into account" without thereby becoming a different machine, namely the old machine with a "new part" added." (Lucas)

#9, p. 63: "We are trying to produce a model of the mind which is mechanical - which is essentially "dead" - but the mind, being in fact "alive", can always go one better than any formal, ossified, dead, system can. Thanks to Gödel's theorem, the mind always has the last word." (Lucas)

#20, pp. 70-71: "The sense in which an automatic door "understands instructions" from its photoelectric cell is not at all the sense in which I understand English. If the sense in which Schank's programmed computers understand stories is supposed to be the metaphorical sense in which the door understands, and not the sense in which I understand English, the issue would not be worth discussing. But Newell and Simon (1963) write that the kind of cognition they claim for computers is exactly the same as for human beings. I like the straightforwardness of this claim, and it is the sort of claim I will be considering. I will argue that in the literal sense the programmed computer understands what the car and the adding machine understand, namely, exactly nothing." (Searle)

#21, p. 74: "The problem in this discussion is not about how I know that other people have cognitive states, but rather what it is that I am attributing to them when I attribute cognitive states to them. The thrust of the argument is that it couldn't be just computational processes and their output because the computational processes and their output can

exist without the cognitive state. It is no answer to this argument to feign anesthesia. In "cognitive sciences" one presupposes the reality and knowability of the mental in the same way that in physical sciences one has to presuppose the reality and knowability of physical objects." (Searle)

#22, p. 76: "Could a machine think?" My own view is that only a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as brains [...] Of course the brain is a digital computer. Since everything is a digital computer, brains are too. The point is that the brain's causal capacity to produce intentionality cannot consist in its instantiating a computer program, since for any program you like it is possible for something to instantiate that program and still not have any mental states. Whatever it is that the brain does to produce intentionality, it cannot consist in instantiating a program since no program, by itself, is sufficient for intentionality." (Searle)

#23, p. 76 (note): "No program we have written can be said to truly understand yet. Because of that, no program we have written "explains the human ability to understand"." (Schank)

#24, p. 77: "Searle's accusation of dualism in AI falls wide of the mark because the mechanist does not insist on a particular mechanism in the organism, but only that "mental" processes be represented in a physical system when the system is functioning. A program lying on a tape spool in a corner is no more conscious than a brain preserved in a glass jar, and insisting that the program if read ~~into~~ the appropriate computer would function with intentionality asserts only that the adequate machine consists of an organization imposed on a physical substrate. The organisation is no more mentalistic that the substrate itself." (Bridgeman)

#27, p. 78: "I find it hard to believe that what philosophers have been after for centuries were chemical explanations for the phenomena that pervade our lives.

Yet that is the position that Searle forces himself into. Because, apart from chemical explanation, what is left? [...] Does the brain understand? Certainly we humans understand, but does that lump of matter we call our brain understand? All that is going on there is so many chemical reactions and electrical impulses, just so many Chinese symbols." (Schank)

#29, p. 79: "If more and more of the cells in your brain were to be replaced by integrated circuit chips, programmed in such a way as to keep the input-output function of each unit

identical to that of the unit being replaced, you would in all likelihood just keep right on speaking exactly as you are doing now except that you would eventually stop **meaning** anything by it. What we outside observers might take to be words would become for you just certain noises that circuits caused you to make." (Pylyshyn)

#30, p. 80: "But what is the right kind of stuff? Is it cell assemblies, individual neurons, protoplasm, protein molecules, atoms of carbon and hydrogen, elementary particles? Let Searle name the level, and it can be simulated perfectly well using "the wrong kind of stuff." (Pylyshyn)

#32, p. 82: "In my own case I have an extra difficulty with behaviorism and functionalism because I cannot imagine anybody actually believing these views. I know that people say they do, but what am I to make of it when Rachlin says that there are no "mental states underlying...behavior" [...]? Are there no pains underlying Rachlin's pain behavior?" (Pylyshyn)

CHAPITRE IV

#38, pp. 110-111: "I sympathise with "humanists" who fear that technical developments threaten our social structure, our traditional image of ourselves and our cultural values [...] The steady encroachment of the computer must be **faced.** It is cowardice to respond by filling "humanities" departments with "phenomenologists" who assure us that the computer is barred by its finite number of states from encroaching further into the areas of activity they regard as "uniquely human" [...]

One third of the book is gossip, and has nothing to do with AI. As for the rest - well, if you take any Dreyfus quotation and you go back and see where it came from, look at it in context, it's always wrong. And it's not that occasionally he missed the point because he didn't see the context. There just aren't any exceptions, and so that must either prove something about Dreyfus, or that it's very hard to find real examples of bad things that were said by leading AI people." (Papert)

#39, p. 111: "What artificial intelligence needs is a good Dreyfus [...] We do have problems, and they could be illuminated by a first-class philosopher. But Dreyfus bludgeons us over the head with stuff he's misunderstood and is obsolete anyway - and every time you confront him with one more intelligent program, he says, "I never said a computer couldn't do **that.**" And what does he offer us instead?

Phenomenology! That ball of fluff! That cotton candy!" (Feigenbaum)

#42, pp. 113-114: "This can only reveal a basic misunderstanding as to the function of scientific understanding. As Einstein is said to have remarked, it is not the function of science to produce the taste of the soup! *The scientis's taks is not to duplicate phenomena but to make them accessible to the intellect* [...] The scientist must substitute for the "real thing", a system built on principles which he can understand. The "ultimate reality" is approachable in its manifest entirety by neither science nor revelation, neither by poetry nor mystic illumination [...] The scientist's task is a never-ending one of unfolding a description which relates both to the phenomena (i.e. the evidence) and to his capacity to intellectually grasp the description (i.e. to his rational capacities)." (Pylyshyn)

#43, p. 114: "By the time he is an adult a person's intelligence depends on his possessing a body only in the obvious sense that his body contains the mechanisms in which intelligence is realized and provides the means for perception, locomotion, etc." (Pylyshyn)

#45, pp. 115-116: "So long as the set of percepts has some structure, so that one percept resembles another in certain respects and yet another in other respects, we know of no way to capture this quality without describing the individual percepts in terms of relations among more primitive elements or qualities." (Pylyshyn)

#46, p. 116: "...the type of phenomenological description which fills the pages of Dreyfus' book can be logically explicated [...] in such a way that all the functional aspects are retained and the only thing lost is the poetic style." (Pylyshyn)

CHAPITRE V

#6, pp. 126-127: "We thus are faced with a dilemna. We would rather use something other than output to tell us if a system really **understands**. Output, however, is all we reasonably can expect to get. Despite the fact that setting out to assess the intelligence of an entity using **any** method at all is a highly dubious pursuit, it may be that evaluation of output is the most effective approach we have for estimating the degree of understanding of both humans and computers. In the end, any system, human or mechanical, is judged by its output." (Schank)

#9, pp. 137-138: "I have argued elsewhere that one important point where the computer analogy breaks down (and will always break down) is the artificiality of computers. Because computers can have no purposes of their own **in the same sense** in which human beings (and animals) have purposes of their own, psychological terms - such as purpose, knowledge, intelligence, speak, communicate... - cannot be litterally applied to any conceivable computer [...] in the last analysis, computers are man-made as babies are not, and have no **intrinsic** purposes at all." (Boden)

BIBLIOGRAPHIE

ADLER, MORTIMER J., "The Challenge to the Computer.", Proceedings of the American Catholic Association XLII, 1968, pp. 20-29. (Extrait du volume The Difference of Man and the Difference It Makes, N.Y., Holt, Rinehart and Winston, Inc., 1967, chap. 14 "From Descartes to Turing".)

BODEN, Margaret A., Artificial Intelligence and Natural Man, New York, Basic Books, 1977.
- Minds and Mechanisms: philosophical psychology and computative models, Harvester Press, 1981.

BOLTER, J. David., Turing's Man: Western Culture in the Computer Age, The University of North Carolina Press, Chapell Hill, 1985.

BONNET, ALAIN, L'intelligence artificielle: promesses et réalités, InterEditions, Paris, 1984.

BOUVRESSE, JACQUES, "Les machines sont-elles intelligentes?", La Recherche, no 170, octobre 1985, pp. 1126-1127.

BRIDGEMAN, BRUCE, "Brains + Programs = Minds", The Behavioral and Brain Sciences, Vol.3, Cambridge University Press, 1980, pp. 427-428.

BRIOT MAURICE ET ARNAUD ROBERT DE SAINT VINCENT, "La vision des robots", La Recherche, no 170, octobre 1985, pp. 1264-1273.

CORDELL, ARTHUR J., Le grand dérangement: à l'aube de la société d'information, Conseil des Sciences du Canada, Ottawa, 1985.

CORDIER, MARIE-ODILE, "Les systèmes experts", La Recherche, no 151, janvier 1984, pp. 60-70.

DENNETT, DANIEL C., Brainstorms: Philosophical Essays on Mind and Psychology, Bradford Books, 1978.
- "The Milk of Human Intentionality", The Behavioral and Brain Sciences, Vol.3, Cambridge University Press, 1980, pp, 428-431.

DREYFUS, Hubert L., Intelligence artificielle: mythes et limites, Paris, Flammarion, 1984. (Traduction de la deuxième édition américaine de What Computers Can't Do: the limits of artificial intelligence, New York, Harper & Row Publishers, 1979.)

FALLER, BENOIT, "L'ordinateur et les jeux de l'esprit", La Recherche, no 170, octobre 1985, pp. 1164-1174.

FEIGENBAUM, EDWARD et PAMELA McCORDUCK, La Cinquième Génération: le pari de l'intelligence artificielle à l'aube du 21ème siècle, InterEditions, Paris, 1984, (Traduction de The Fifth Generation, Addison-Wesley Publishing Co., Reading, Mass., 1983.)

GANASCIA, JEAN GABRIEL, "La conception des systèmes experts", La Recherche, no 170, octobre 1985, pp. 1142-1151.

GALLAIRE, HERVE, "La représentation des connaissances", La Recherche, no 170, octobre 1985, pp. 1240-1248.

GEORGE, F.H., "Minds, Machines, and Gödel: another reply to Mr Lucas", Philosophy, no 37, (1962), pp. 62-63.

GODEL, KURT, On Formally Undecidable Propositions, New York: Basic Books, 1962. Original publié sous le titre "Über Formal Unentscheidbare Sätze der Principia Mathematica und Verwandter Systeme, I." Monatshefte für Mathematik und Physik, 38 (1931), 173-198.

HAUGELAND, JOHN, "Programs, Causal Powers, and Intentionality", The Behavioral and Brain Sciences, Vol. 3, Cambridge University Press, 1980, pp. 432-3.

HOFSTADTER, Douglas R., Gödel, Escher, Bach: an eternal golden braid, Basic Books, 1979.

HOFSTADTER, Douglas R. and Daniel C. Dennett, The Mind's I: Fantasies and Reflections on Self and Soul, Bantam Books, N.Y., 1982.

KAYSER, DANIEL, "Des machines qui comprennent notre langue", La Recherche, no. 170, octobre 1985, pp. 1198-1209.

KENT, ERNEST W., The Brains of Men and Machines, McGraw-Hill Publications Co., N.Y., 1981.

KODRATOFF, YVES, "Quand l'ordinateur apprend", La Recherche, no 170, octobre 1985, pp. 1252-1262.

LEMAIRE JOSSE, "Variations sur la pensée cybernétique", Le dossier de la cybernétique, Marabout Université, Paris, 1968, p.67.

LUCAS, J.R., "Minds, Machines, and Gödel", Philosophy, 36, (1961), pp. 112-127. (Reproduit dans la collection d'articles Minds and Machines, par Alan Ross Anderson, Engelwood Cliffs, N.J., Prentice-Hall, 1964.)

MCCORDUCK, PAMELA, Machines Who Think, W.H. Freeman & Co., San Francisco, 1979.

MINSKY, MARVIN L., "Computer Science and the Representation of Knowledge", The Computer Age: A Twenty-Year View, publié par Michael L. Dertouzos et Joel Moses, MIT Press, Cambridge, 1979, pp.392-421.

MOTO-OKA, TOHRU, "Les ordinateurs de cinquième génération", La Recherche, no 154, avril 1984, pp. 516-525.

NEWELL ALLEN, J.C. SHAW ET H.A. SIMON, "Empirical Explorations with the Logic Theory Machine: A Case Study in Heuristics", in Computers and Thought, Edward A. Feigenbaum et Julian Feldman, eds. (New York, McGraw-Hill, 1963)

PAPERT, SEYMOUR, Jaillissement de l'esprit: ordinateurs et apprentissage, Flammarion, Paris, 1981. (Traduction de Mindstorms, Basic Books, N.Y., 1980).

PARENT, RICHARD, Point de vue québécois sur l'intelligence artificielle, Ministère des Communications du Québec, 1984.

PITRAT, JACQUES, "La naissance de l'intelligence artificielle", La Recherche, no 170, octobre 1985, pp. 1130-1141.

PYLYSHYN, ZENON W., "Minds, machines and phenomenology: Some reflections on Dreyfus' "What Computers Can't Do". Cognition, 3, (1974)

RINGLE, MARTIN, "Mysticism as a Philosophy of Artificial Intelligence", The Behavioral and Brain Sciences, Vol. 3, Cambridge University Press, 1980, pp. 444-445.

RITCHIE, David, Le Cerveau Binaire, Editions Robert Laffont, Paris, 1984. (Traduction de The Binary Brain, L.,B. & Co., Boston).

ROSE, FRANK, "The Black Knight of AI", Science, mars 1985, pp. 46-51.

SCHANK, Roger C., The Cognitive Computer: on Language, Learning and Artificial Intelligence, Addison-Wesley Publishing Co., 1985.
- "Understanding Searle", The Behavioral and Brain Sciences, Vol. 3, Cambridge University Press, 1980, pp. 446-447.

SEARLE, JOHN R., "Minds, Brains and Programs" (suivi de réponses de 27 auteurs), The Behavioral and Brain Sciences, Vol. 3, Cambridge University Press, 1980. (Reproduit dans The Mind's I de Dennett et Hofstadter)

STEVENS, LAWRENCE, Artificial Intelligence: the Search for the Perfect Machine, Hayden Book, Hasbrouck Heights, N.J., 1985.

WALDROP, M. MITCHELL, "Machinations of Thought", Science, Mars 1985, pp. 38-45.

VON NEUMANN, J., The Computer and the Brain, Yale University Press, N.H., Conn., 1958.

WEIZENBAUM, Joseph, Puissance de l'Ordinateur et Raison de l'Homme, Paris, Editions d'Informatique, 1981. (Traduction de Computer Power and Human Reason, W.H. Freeman and Co., 1976).

WIENER, NORBERT, The Human Use of Human Beings, Houghton Mifflin, Boston, 1950.
- God and Golem, Inc, The MIT Press, Cambridge, Mass., 1964.

WINOGRAD, TERRY, "Towards a Procedural Understanding of Semantics", Revue Internationale de Philosophie, #117-118, 1976, pp. 260-303.

INDEX

RESUME DE LA THESE

Ce travail traite d'une question à caractère philosophique chaudement débattue depuis l'apparition des premiers ordinateurs: une machine peut-elle penser?

Les techniques de programmation ont permis, dans les années 50, l'émergence d'une nouvelle discipline, **l'intelligence artificielle**. Ce développement a donné lieu à de grands espoirs: les pionniers de l'I.A. espéraient réaliser, avant 1968, un ordinateur capable de rivaliser avec les plus grands champions d'échecs et de démontrer un important théorème de mathématiques. Ces premiers projets, tout comme ceux encore plus ambitieux de nos jours, reposent sur l'hypothèse suivante: tout aspect de l'apprentissage, ou de toute faculté associée à l'intelligence, peut être décrit avec suffisamment de précision pour être reproduit sur ordinateur.

Trente ans plus tard, malgré des réalisations remarquables de l'I.A., l'ordinateur "pensant" se fait encore attendre. S'agit-il d'un simple problème d'évolution technique que les progrès éventuels des connaissances vont résoudre? Ou bien la voie adoptée pour reproduire artificiellement l'intelligence humaine n'est pas la bonne?

Les projets de l'I.A. ont déclenché un débat vigoureux sur la nature de l'esprit, de l'intelligence, de la pensée: informaticiens, psychologues et philosophes ont repris à la lumière des réalisations et des échecs de l'I.A. des questions au coeur de la réflexion philosophique.

Cet exposé a pour objectif de faire une synthèse des principaux arguments sur le sujet. Au chapitre premier, nous ferons voir l'émergence des premières interrogations soulevées par l'apparition des ordinateurs ainsi que la pertinence, pour la philosophie, d'une réflexion particulière sur l'intelligence artificielle. Le chapitre II brossera un tableau des origines de cette nouvelle discipline, des espoirs et des échecs auxquels elle a donné lieu, ainsi que de ses perspectives à moyen et long terme. Les troisième et quatrième chapitres traiteront de façon toute particulière des divers aspects du débat philosophique sur la question de l'intelligence des machines: au chapitre III, nous aborderons plus spécialement certains problèmes soulevés par les limites des systèmes logiques (selon le théorème de Gödel), et nous présenterons une synthèse de la polémique lancée par John R. Searle, philosophe américain, sur l'**intentionnalité** inhérente à la nature biologique de l'esprit humain: le quatrième chapitre sera consacré aux thèses du philosophe américain Hubert L. Dreyfus dont la perspective résolument phénoménologique a causé de sérieux remous dans les milieux de l'intelligence artificielle aux Etats-Unis. En guise de conclusion, le dernier chapitre proposera une appréciation personnelle des présupposés à caractère philosophique qui sous-tendent les projets de réalisation d'une **intelligence de synthèse**.