Des collaborations possibles entre philosophie et intelligence artificielle

Robin Lamarche-Perrin

Mémoire de recherche

présenté par

Robin Lamarche-Perrin

Master 2 Philosophie
Histoire de la philosophie et philosophies du langage
Option *Discours, Savoirs, Représentations*

Université Pierre-Mendès-France
UFR SH Grenoble

Des collaborations possibles entre philosophie et Intelligence Artificielle

---

Soutenu le 11 septembre 2012
devant les membres du jury

Directeur	Denis Perrin
Examinateur	Denis Vernant
Responsable du Master	Marie-Laurence Desclos
Responsable de la spécialité	Éric Dufour

Remerciements

Je tiens à remercier Denis Perrin pour ses cours et ses conseils durant les cinq dernières années. Je remercie également Anna Zielinska pour son travail d'édition de longue haleine. Merci à Jean-Michel Roy d'avoir partagé ses avis sur la version préliminaire de ce travail et, par avance, merci aux futurs lecteurs qui voudront en faire autant.

Un an avant l'échéance, je pense bien évidemment à tous ceux qui participent, de près ou de loin, à mon travail de doctorat : Yves et Jean-Marc, les chefs d'orchestre de ce marathon quotidien ; Marta et Claude, avec qui je prends toujours autant de plaisir à collaborer ; Lucas, qui est un coéquipier hors pair et un chercheur infatigable ; Cyrille et Jérémy, qui savent bien que ce sont les meilleurs qui partent en premier ; Julie, pour son éternelle bonne humeur.

Enfin, je tiens à remercier mon irréductible relectrice, Alice, ainsi que Mathieu et Régis pour leur soutien quotidien. Merci à Séverine et Jean de m'avoir hébergé pendant cet été brûlant. Merci également à Hachikō d'avoir fini par comprendre que, quand j'étais dehors, ce n'était pas nécessairement pour lui lancer un bout de bois.

Ce mémoire est bien évidemment dédié à Alan Turing, qui aurait eu cent ans cette année.

Table des matières

Partie 1. Introduction

Ce mémoire s'intéresse aux relations possibles entre une discipline millénaire et une discipline qui, en comparaison, en est encore à ses débuts[1] : la philosophie et l'Intelligence Artificielle. En ce sens, nous sommes moins intéressés par les résultats particuliers de ces deux disciplines, par leurs théories et par leurs objets, que par les formes générales de collaboration qui peuvent exister entre les deux communautés de recherche. Cette introduction sert avant tout à expliciter la stratégie adoptée dans ce mémoire pour mettre en place de telles collaborations (*cf.* figure 1 ci-dessous).

Le chapitre 1.1 présente deux manières de penser les rapports entre philosophie et Intelligence Artificielle. La première milite pour une « philosophie *de* l'Intelligence Artificielle », élaborant ainsi des rapports unilatéraux d'ordre *épistémologique*. La seconde, celle qui nous intéresse dans ce mémoire, propose au contraire une collaboration réciproque entre la philosophie *et* l'Intelligence Artificielle, à partir de rapports *scientifiques* bilatéraux. Le chapitre 1.2 explicite les problématiques générales concernant la notion d'« intelligence artificielle ». La problématique particulière de ce mémoire concerne les méthodes de résolution de telles problématiques générales. Elle a donc une visée *épistémologique*. Plus précisément, il s'agit de savoir comment la philosophie et l'Intelligence Artificielle peuvent collaborer pour répondre à des problématiques *scientifiques* concernant la notion d'« intelligence artificielle ». Le chapitre 1.3 introduit une distinction importante entre ce qu'on appelle l'IA faible et l'IA forte (*cf.* figure 1, étape 1). La première regroupe des problèmes propres à l'Intelligence Artificielle et la seconde des problèmes propres à la philosophie de l'esprit (étape 2). Le chapitre 1.4 explicite la stratégie du mémoire. Elle consiste à mettre en relation l'IA faible et l'IA forte (étape 3) pour définir différents modes de collaborations entre l'Intelligence Artificielle et la philosophie (étape 4). On se sert donc de relations (logiques ou épistémologiques) entre deux problèmes scientifiques pour opérer le rapprochement des disciplines concernées.

---

[1] Margaret A. Boden propose l'année 1943 comme date de naissance symbolique pour l'Intelligence Artificielle. Boden, M.A. 1995. « AI's Half-Century. » *AI Magazine*, vol. 16, n°4. Cette année correspond à la publication des travaux de Warren S. McCulloch et Walter H. Pitts liant « le calcul, la logique et les systèmes nerveux ». McCulloch, W.S., Pitts, W.H. 1943. « A Logical Calculus of the Ideas Immanent in Nervous Activity. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 22-39.

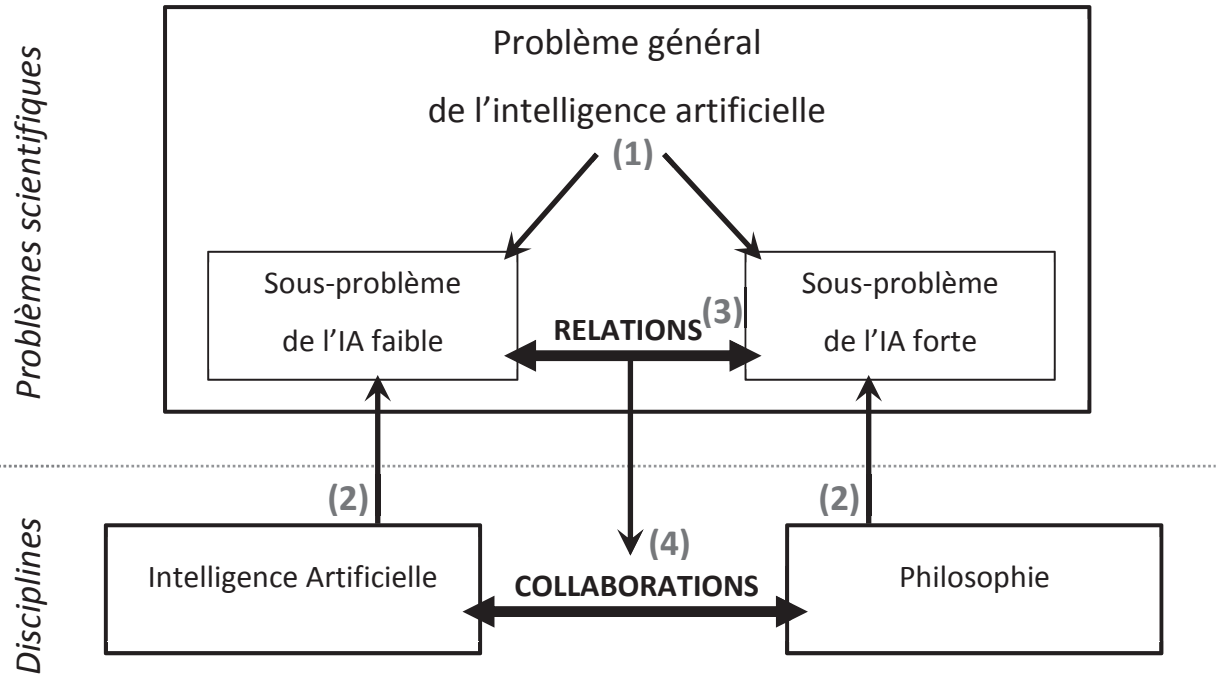

Figure 1 : stratégie globale du mémoire

Chapitre 1.1. Deux manières de penser la collaboration

Il convient de présenter, dans un premier temps, les démarches communément défendues concernant le rapprochement de la philosophie et de l'Intelligence Artificielle. À ce sujet, l'avant-propos de Daniel Andler à la version française de *What Computers Can't Do*[2] – ouvrage qui sera par ailleurs très utile à nos recherches – fournit de bonnes bases. Nous distinguons deux catégories d'approches : celles qui font de la philosophie une discipline « méta », au-delà de l'Intelligence Artificielle, et qui porte sur elle des jugements *épistémologiques*, et celles qui envisagent des échanges couplés entre les deux disciplines, une collaboration « à un même niveau » où la philosophie et l'Intelligence Artificielle entretiennent des rapports *scientifiques* bilatéraux.

La philosophie de l'Intelligence Artificielle

La façon la plus commune de concevoir les rapports entre Intelligence Artificielle et philosophie s'inscrit dans le cadre de la *philosophie des sciences*. Dans ce contexte, si on accorde à l'Intelligence Artificielle le statut de science[3], il doit exister une « philosophie de l'Intelligence Artificielle », tout comme il existe une « philosophie de la physique » et une « philosophie de la biologie ». La philosophie est alors un discours *sur* l'Intelligence Artificielle : sur son activité, ses méthodes, ses outils, sur la nature de sa pensée scientifique et sur la manière d'évaluer ses résultats. À ce titre, la « philosophie de l'Intelligence Artificielle » est une *épistémologie*. Elle fait l'étude de la science elle-même, et non de ses objets. Elle travaille à un niveau « méta », au-delà des théories et des expériences particulières. Les rapports sont donc unilatéraux : l'Intelligence Artificielle ne peut pas porter de jugements *sur* la philosophie, dans la mesure où celle-ci travaille en-dehors de son champ d'action scientifique, et la philosophie n'est pas inquiétée par les résultats particuliers de l'Intelligence Artificielle.

Pour John McCarthy, une telle « philosophie de l'Intelligence Artificielle » ne peut avoir beaucoup d'effets sur les pratiques de la discipline, « pas plus que la philosophie des sciences n'en a généralement sur la pratique des sciences. »[4] Sans commenter le peu d'optimisme qu'a McCarthy

---

[2] Dreyfus, H.L. 1979. *Intelligence Artificielle : mythes et limites*. [*What Computers Can't Do: The Limits of Artificial Intelligence*, 2nd ed.] Vassallo-Villaneau, R.-M. (trad.), Andler, D. (pref.), Perriault, J. (pref.). Paris : Flammarion, 1984.

[3] Ce qui est déjà un point polémique si l'on en croit Andler.

[4] « [The philosophy of artificial intelligence] is unlikely to have any more effect on the practice of AI research than philosophy of science generally has on the practice of science. » McCarthy, J. 1995. « What has AI in Common with Philosophy? » *International Joint Conference on Artificial Intelligence (IJCAI'95)*, vol. 2, p. 2041-2042.

concernant l'intérêt de la philosophie des sciences pour la pratique des sciences, nous retenons surtout que, pour un certain nombre de chercheurs, il existe d'autres démarches collaboratives pour opérer un rapprochement entre philosophie et Intelligence Artificielle.

La philosophie et l'Intelligence Artificielle

Certaines de ces démarches sont mises en évidence par Andler[5]. Elles consistent à voir que la philosophie et l'Intelligence Artificielle peuvent échanger à propos d'objets communs. Pour McCarthy par exemple, la philosophie est plus utile à l'Intelligence Artificielle sur le plan conceptuel que sur le plan méthodologique. Elle dispose notamment de théories concernant des qualités humaines difficiles à programmer. Les développeurs pourraient ainsi bénéficier de ces théories philosophiques pour implémenter au sein de leurs machines « des croyances », « des intentions », « un libre-arbitre », *etc.*[6] Ce que cherche McCarthy, ce sont donc des modèles sur lesquels fonder un travail pratique de conception. Avec cette première démarche, la philosophie est envisagée comme une source de *fondements conceptuels*. Ainsi, Stuart J. Russell et Peter Norvig consacrent un chapitre entier aux « fondements philosophiques » de l'Intelligence Artificielle dans leur ouvrage de référence[7] et, dans ce mémoire, nous abordons le rôle fondationnel de la philosophie à plusieurs reprises : concernant le socle philosophique du computationnalisme (section 2.4.2), celui de la « nouvelle IA » (chapitre 3.1) et pour importer en Intelligence Artificielle le concept d'« émergence épistémique » développé par la philosophie britannique (chapitre 3.2).

Ce rôle fondationnel est parfois pris en charge par la « philosophie de l'Intelligence Artificielle ». C'est par exemple la position de Boden qui, dans l'introduction de son recueil précisément intitulé *The Philosophy of Artificial Intelligence*[8], donne pour objectif à la philosophie de préciser les conditions de possibilité de l'Intelligence Artificielle[9]. Il s'agit d'expliquer comment

---

[5] Dreyfus, H.L. 1979. *Op. cit.*, p. XI-XIV.

[6] McCarthy, J. 1995. *Op. cit.*, p. 2041. On peut reprocher à McCarthy son optimisme débordant concernant l'applicabilité des concepts philosophiques aux programmes de l'Intelligence Artificielle. C'est pourquoi nous retenons avant tout sa démarche générale, qui a le mérite de concevoir des rapports intéressants entre les deux disciplines, plutôt que sa connaissance réelle des concepts manipulés par la philosophie de l'esprit.

[7] Russell, S.J., Norvig, P. 2003. « Chapter 26. Philosophical Foundations. » *Artificial Intelligence: A Modern Approach,* 3rd ed. Upper Saddle River, NJ : Prentice Hall, p. 1020-1043.

[8] Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press.

[9] « The many philosophical problems associated with AI arise from the question whether (and if so, how) this ambitious enterprise could be achieved or whether it is radically misconceived. » *Ibid.*, p. 1.

l'Intelligence Artificielle peut mener à bien son projet – quel qu'il soit[10] – et de lui fournir les concepts fondamentaux dont elle aura besoin. La particularité des démarches bilatérales ne repose donc pas sur ce rôle fondationnel, mais sur un renversement de la relation unilatérale défendue par la « philosophie de l'Intelligence Artificielle ». C'est alors au tour de l'Intelligence Artificielle de mettre ses résultats au service de la philosophie. Pour de telles démarches, les travaux pratiques de l'Intelligence Artificielle constituent notamment une source de clarifications conceptuelles. Mieux, ils offrent des méthodes pour tester les théories et pour valider empiriquement les hypothèses philosophiques[11]. Le renversement est alors complet quand Andler baptise l'Intelligence Artificielle la « science de la philosophie »[12], par opposition à la « philosophie des sciences ». Dans ce contexte, l'Intelligence Artificielle peut mettre en place des expériences de falsification de théories, ou au contraire valider expérimentalement des hypothèses, et ainsi s'immiscer dans des controverses proprement philosophiques. Elle peut, enfin, évaluer les méthodes et les travaux de la philosophie.

Cette approche, selon laquelle l'Intelligence Artificielle peut être utile aux débats philosophiques, est également abordée à plusieurs reprises dans ce mémoire. C'est le cas de la démarche d'Hector J. Levesque (section 2.2.3) et de celle de Timothy van Gelder (chapitre 4.2). De manière générale, dans la partie 4, nous donnons des exemples de collaboration allant dans ce sens : de l'Intelligence Artificielle vers la philosophie.

Positionnement du mémoire

Andler résume admirablement la collaboration bilatérale qui peut exister entre l'Intelligence Artificielle et la philosophie de l'esprit. « La philosophie peut aider l'IA à progresser sur le plan des concepts, donc à terme sur le plan pratique. »[13] « L'IA oblige à préciser [les] choix [de la philosophie], à dissiper des malentendus millénaires, à formuler et à tester de manière scientifique de nouvelles hypothèses. »[14] L'objectif de ce mémoire est d'expliciter ces attentes et de préciser comment de

---

[10] Boden a une définition particulière de la discipline, faisant de l'Intelligence Artificielle « *la science de l'intelligence en général* ». *Ibid.*, p. 1. Nous ne souscrivons pas cette acception trop générale, mais le rôle de la philosophie que Boden propose (concernant l'analyse de conditions de possibilité) peut être appliqué à une définition plus stricte et plus commune de l'Intelligence Artificielle (*cf.* chapitre 1.2).

[11] Pour Inman Harvey, les théories philosophiques peuvent ainsi être testées scientifiquement « dans le monde réel. » Il a d'ailleurs pour cela une formule amusante : « faire de la philosophie de l'esprit à l'aide d'un tournevis. » Harvey, I. 2000. « Robotics: Philosophy of Mind Using a Screwdriver. » *Evolutionary Robotics: From Intelligent Robots to Artificial Life*, vol. III, p. 207-230.

[12] Dreyfus, H.L. 1979. *Op. cit.*, p. XIV.

[13] *Ibid.*, p. XIII.

[14] *Ibid.*, p. XIII-XIV.

telles collaborations peuvent être établies. En ce sens, si nous affirmons que la philosophie ne doit pas être limitée à une *épistémologie*, ce mémoire a par contre une visée profondément épistémologique. En effet, il fait l'analyse de processus *scientifiques* et, plus précisément, de processus de *collaboration scientifique*. Les objets, les concepts et les problématiques des deux disciplines ne sont pas *directement* les objets de nos recherches. En quelque sorte, les objets principaux de ce mémoire sont les démarches collaboratives entre les deux disciplines. Il s'agit donc de la « philosophie des rapports scientifiques entre la philosophie et l'Intelligence Artificielle ».

La base de notre argumentation est néanmoins constituée d'exemples particuliers de collaborations faisant intervenir des concepts, des théories et des controverses en philosophie et en Intelligence Artificielle. En effet, comme l'exige Andler, « il faut, avant tout, juger sur pièces »[15], c'est-à-dire qu'il faut aborder les relations possibles à partir d'exemples concrets. Dans le cadre de ces exemples, nous nous intéressons donc ponctuellement à des problématiques propres aux deux disciplines. Mais il faut garder en tête que ces discussions visent à être généralisées et qu'elles ont pour objets essentiels les *rapports scientifiques* entre philosophie et Intelligence Artificielle.

[15] *Ibid.*, p. XIV.

Chapitre 1.2. Le problème général de l'intelligence artificielle

Qu'est-ce que l'Intelligence Artificielle ?

Margaret A. Boden propose deux définitions de l'Intelligence Artificielle. La première est une définition relativement classique : l'Intelligence Artificielle a pour objectif de construire des machines (ordinateurs, robots, programmes) capables de faire ce que les esprits font[16] :

> L'intelligence artificielle est parfois définie comme étudiant les façons de construire et/ou de programmer des ordinateurs pour leur permettre de faire le genre de choses que les esprits savent faire.[17]

Cette définition généralise la toute première mention du terme « Intelligence Artificielle », formulée lors d'une école d'été au *Dartmouth College* en 1956 par quatre initiateurs de la discipline. L'hypothèse sur laquelle ils proposent alors de réfléchir est la suivante :

> Chaque aspect de l'apprentissage, ou de tout autre caractéristique de l'intelligence, peut en principe être décrit si précisément qu'une machine peut être construite pour le simuler.[18]

À l'origine, l'Intelligence Artificielle est donc une discipline qui vise à implémenter des facultés de l'esprit, notamment liées à la notion d'intelligence, au sein de machines, notamment numériques. La problématique générale de l'Intelligence Artificielle peut ainsi être formulée : « comment construire de telles machines ? » Selon la démarche de Boden, la philosophie, en s'intéressant aux conditions de possibilité d'un tel projet, pose des questions préliminaires : « les machines *peuvent-elles* être intelligentes ? » « *Que faut-il* pour qu'une machine soit intelligente ? » Toutes ces questions constituent ce que nous appelons par la suite « le problème général de

---

[16] La seconde définition, celle que Boden utilise (*cf.* note 10 ci-dessus), ne permet pas de bien distinguer les objectifs philosophiques des objectifs techniques de l'Intelligence Artificielle. Nous utilisons donc la définition classique.

[17] « Artificial intelligence is sometimes defined as the study of how to build and/or program computers to enable them to do the sort of things that minds can do. » [Notre traduction] Boden, M.A. 1990. *Op. cit.*, p. 1.

[18] « Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. » [Notre traduction] McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E. 1955. « A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. » *AI Magazine*, 2006, vol. 27, n°4, p. 12-14.

l'intelligence artificielle[19] ». Elles expriment très généralement les objectifs philosophiques, scientifiques et techniques liés à la notion de « machine intelligente ». Dès l'abord, l'Intelligence Artificielle (en tant que discipline) et la philosophie n'abordent pas cette problématique de la même façon. Cette divergence concernant le problème général de l'intelligence artificielle sera explicitée dans le chapitre suivant.

Problématique de recherche

Ce mémoire n'a pas pour but de répondre *directement* au « problème général de l'intelligence artificielle ». Il s'intéresse en effet aux méthodes possibles pour apporter une réponse à un tel problème. À ce titre, il a une visée *épistémologique*, orientée vers la méthode scientifique, et non vers les théories et les objets particuliers (*cf.* chapitre 1.1). Dans un premier temps, la problématique de ce mémoire peut ainsi être formulée : « comment résoudre le problème général de l'intelligence artificielle ? » Plus précisément, nous nous intéressons aux apports d'une collaboration entre philosophie et Intelligence Artificielle. « Est-ce que le travail couplé des deux disciplines peut aider à la résolution du problème général de l'intelligence artificielle ? Et, si c'est le cas, de quelle manière devraient-elles collaborer ? » Ainsi, les parties 2 et 3 exposent et analysent différentes stratégies de collaboration. Celles-ci sont évaluées en fonction de leurs capacités à répondre aux problèmes scientifiques : « les machines peuvent-elles être intelligentes ? » et « comment construire de telles machines ? »

Dans un second temps, grâce au renversement annoncé par Andler, ces stratégies de collaboration sont évaluées relativement à la résolution de problèmes purement philosophiques, concernant la nature de l'esprit et de l'intelligence (partie 4). De manière générale donc, ce mémoire cherche à montrer comment la philosophie peut aider à résoudre des problèmes qui sont chers à l'Intelligence Artificielle (et réciproquement), mais il montre également que ces problèmes, même s'ils semblent à première vue spécifiques à chacune des deux disciplines, peuvent également présenter un grand intérêt pour l'une et l'autre : d'une part, les problèmes philosophiques sont également des problèmes d'Intelligence Artificielle et, d'autre part, les problèmes pratiques de l'Intelligence Artificielle constituent à leurs tours des problèmes philosophiques.

---

[19] L'« intelligence artificielle » (sans majuscule) désigne le concept même d'« intelligence artificielle », c'est-à-dire une intelligence artificiellement conçue. Elle est à distinguer de l'« Intelligence Artificielle » (avec des majuscules) qui désigne une discipline, dont l'objectif est de réaliser de telles « intelligences artificielles ». Le « problème de l'intelligence artificielle » est donc un problème scientifique général, dont l'Intelligence Artificielle n'a pas le monopole *a priori*.

Le chapitre suivant précise une distinction importante en ce qui concerne le problème général de l'intelligence artificielle. Elle permet de bien différencier, dans un premier temps, le travail de la philosophie et celui de l'Intelligence Artificielle concernant la « problématique générale de l'intelligence artificielle ».

Chapitre 1.3. Les deux sous-problèmes de l'intelligence artificielle

Boden précise deux acceptions possibles de la définition classique de l'Intelligence Artificielle. L'une présuppose que les machines, de la même façon que les esprits, « peuvent réellement diagnostiquer, conseiller, inférer et comprendre. »[20] L'autre s'intéresse seulement aux « performances observables »[21] de ces machines, sans faire de supposition quant à leurs réelles capacités à comprendre et à penser. John R. Searle introduit le même genre de distinction afin de clarifier les préoccupations de l'Intelligence Artificielle. Il distingue deux projets, nommément l'« IA faible » et l'« IA forte » (« *weak AI* » et « *strong AI* » en anglais), qui abordent la problématique générale de l'intelligence artificielle de deux manières différentes. Dans ce chapitre, nous explicitons les objectifs particuliers de ces deux projets et nous précisons les termes qui seront utilisés dans ce mémoire.

La genèse des termes par Searle

Les termes « IA faible» et « IA forte » apparaissent pour la première fois en 1980, dans un article où Searle amorce sa longue critique de l'Intelligence Artificielle[22]. Ils sont définis dans le premier paragraphe de l'article, de la manière suivante :

> Pour l'IA faible, la valeur principale de l'ordinateur dans l'étude de l'esprit est de nous offrir un outil très puissant. Par exemple, il nous permet de formuler et de tester des hypothèses d'une façon plus rigoureuse et précise.[23]
>
> Pour l'IA forte, au contraire, l'ordinateur n'est pas simplement un outil pour l'étude de l'esprit ; bien plus, programmé de façon appropriée, il *est* réellement un esprit, au sens où

[20] « they might really diagnose, advice, infer, and understand » [Notre traduction] Boden, M.A. 1990. *Op. cit.*, p. 1.

[21] « observable performance » *Ibid.*

[22] Searle, J.R. 1980. « Minds, Brains, and Programs. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 67-88.

[23] *Ibid.*, p. 67. « According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. » Duyckaerts, E. (trad.). *In Quaderni* : *Genèse de l'intelligence artificielle*, Printemps 1987, n°1, p. 65.

> l'on peut dire littéralement d'ordinateurs dotés de programmes corrects qu'ils comprennent et ont d'autres états cognitifs.[24]

La distinction entre les deux projets réside ainsi dans une différence quant à la nature des ordinateurs. Dans le cas de l'IA faible, ils sont des outils puissants qui assistent le scientifique dans ses recherches sur la nature et le fonctionnement de l'esprit. Dans le cas de l'IA forte, les ordinateurs sont *à proprement parler* des esprits. Plus tard, Searle accole à l'hypothèse de l'IA forte la théorie computo-représentationnelle, selon laquelle l'esprit est lui-même un programme d'ordinateur[25] :

> À l'extrême de cette théorie [IA forte], l'idée que le cerveau n'est qu'un ordinateur numérique, et l'esprit un programme d'ordinateur. [L]'esprit est au cerveau ce que le programme est au 'hardware' informatique.[26]

Cette distinction IA faible/IA forte permet donc de décomposer le problème général de l'intelligence artificielle en deux catégories de problématiques scientifiques distinctes (*cf.* figure 1, étape 1). Dans la suite, nous explicitons les acceptions que nous utiliserons dans ce mémoire et nous montrons qu'elles permettent d'identifier les préoccupations particulières de l'Intelligence Artificielle, d'une part, et celles de la philosophie, d'autre part.

La notion de « simulation »

Dans ce mémoire, nous travaillons à partir d'acceptions plus générales des termes « IA faible » et « IA forte ». Nous pensons que la notion fondamentale qui permet de distinguer les deux IA est la notion de « simulation », en tant qu'action de « faire paraître comme réel, effectif (ce qui ne l'est pas). »[27] Ainsi, l'IA faible propose d'assister les recherches sur la nature de l'esprit en simulant ses facultés, c'est-à-dire en produisant des comportements intelligents similaires, en faisant

---

[24] *Ibid.*, p. 67. « According to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states. » Duyckaerts, E. (trad.). *Op. cit.*

[25] *Cf.* la section 2.3.1 concernant la théorie computo-représentationnelle de l'esprit et la section 2.3.2 concernant la critique de cette théorie par Searle.

[26] « According to the most extreme version of this view [strong AI], the brain is just a digital computer and the mind is just a computer program. [T]he mind is to the brain, as the program is to the computer hardware. » [Notre traduction] Searle, J.R. 1984. *Du cerveau au savoir : conférences Reith 1984 de la BBC*. [*Minds, Brains and Science: The 1984 Reith Lectures.*] Chaleyssin, C. (trad.). Paris : Hermann, 1985, p. 28. Cette association entre l'IA forte et la théorie computo-représentationnelle est réaffirmée dans Searle, J.R. 1997. *The Mystery of Consciousness*. New York, NY : The New York Review of Books, p. 9.

[27] Première définition de « simuler » dans le *Nouveau Petit Robert*, édition 2002.

comme si les machines étaient réellement intelligentes. Pour l'IA forte, les machines peuvent faire mieux qu'une simple simulation de l'intelligence : elles peuvent être *réellement* intelligentes. Les définitions proposées par Stuart J. Russell et Peter Norvig dans leur ouvrage de référence *Artificial Intelligence: A Modern Approach* sont à ce titre adaptées à la distinction que nous proposons :

> L'assertion selon laquelle les machines pourraient agir *comme si* elles étaient intelligentes est appelée hypothèse de l'IA faible par les philosophes, et l'assertion selon laquelle les machines qui agissent ainsi pensent *effectivement* (et ne font pas que *simuler* le fait de penser) est appelée hypothèse de l'IA forte.[28]

L'IA faible et l'IA forte se définissent donc à partir de la distinction entre « intelligence simulée » et « intelligence véritable ». Cette distinction s'appuie sur une différence de nature des machines étudiées et permet ainsi d'expliciter les problématiques propres à la notion d'« intelligence artificielle ». Le problème d'une telle définition est qu'elle dépend du critère utilisé pour identifier ce qui est de l'ordre de la véritable intelligence et ce qui ne l'est pas. Ainsi, comme nous le verrons dans la partie 2, de nombreux critères ont été avancés dans les débats historiques pour évaluer la « véritable intelligence » des machines : est-ce que leurs fonctionnements cognitifs internes sont similaires à ceux des hommes ? Est-ce qu'elles sont douées de consciences, de vies phénoménales ? Est-ce qu'elles ont des états mentaux, une intentionnalité ? L'IA forte dépend donc du contexte philosophique et de ce que l'on cherche à opposer à la « fausse intelligence », c'est-à-dire à l'« intelligence simulée ». Pareillement, l'IA faible peut endosser différentes acceptions suivant ce que l'on met derrière le terme de « simulation » : est-ce que les machines peuvent reproduire les comportements humains en général ? Des comportements dialogiques ? Des comportements moteurs ? Est-ce qu'elles peuvent également reproduire des fonctions non cognitives (reproduction, duplication, régénération) ? Est-ce qu'elles peuvent reproduire le résultat *externe* de fonctions (associations entre des entrées et des sorties) ou également leurs structures *internes* (*e.g.*, les règles de manipulations des données) ?

Ainsi, il faut veiller à bien distinguer les critères retenus lorsqu'on parle d'IA faible et d'IA forte. Nous les préciserons autant que faire se peut. Mais, par souci de généricité, nous parlerons également d'« intelligence simulée » et d'« intelligence véritable » pour désigner les objectifs généraux des deux IA. Ce qui nous intéresse, ce sont les relations possibles entre ces deux formes d'intelligences, indépendamment de leurs acceptions particulières.

---

[28] « The assertion that machines could act *as if* they were intelligent is called the weak AI hypothesis by philosophers, and the assertion that machines that do so are *actually* thinking (not just *simulating* thinking) is called the strong AI hypothesis. » [Notre traduction] Russell, S.J., Norvig, P. 2003. *Op. cit.*, p. 1020.

Spécialistes de l'IA et philosophes de l'esprit

La distinction IA faible/IA forte formulée par Searle permet d'identifier le travail de deux communautés de chercheurs (*cf.* figure 1, étape 2). D'une part, les spécialistes de l'IA[29] s'efforcent de concevoir des méthodes et des outils informatiques pour aider à l'étude de phénomènes complexes, et notamment à l'étude de la cognition (IA faible). D'autre part, la philosophie de l'esprit s'interroge sur la nature même des machines, sur leur vie phénoménale, leurs états mentaux, *etc.* (IA forte) Cette divergence s'applique également à la distinction IA faible/IA forte reposant sur la notion de « simulation ». Les spécialistes de l'IA s'intéressent alors à la *simulation* de comportements intelligents et les philosophes à la *réalité* d'une telle intelligence. La problématique générale de l'intelligence artificielle est ainsi divisée en deux sous-problèmes qui, à première vue, appartiennent à deux champs disciplinaires distincts : les spécialistes de l'IA répondent au problème de l'IA faible « les machines peuvent-elles *simuler* l'intelligence ? » et les philosophes au problème de l'IA forte « les machines peuvent-elles être *réellement* intelligentes ?»[30]

La distinction IA faible/IA forte repose également sur une différence de nature des travaux réalisés par leurs communautés respectives. D'une part, les philosophes établissent et évaluent des « théories de l'esprit », c'est-à-dire des modèles spéculatifs concernant l'ontologie de l'esprit et de ses facultés. La philosophie est à ce titre une *science explicative*. D'autre part, les spécialistes de l'IA s'intéressent à des « machines concrètes », c'est-à-dire à des objets particuliers, à des individus[31]. Elle est avant tout une *science applicative*. Les parties 3 et 4 montrent comment cette nature duale entre *explication* et *application* peut être unifiée en ce qui concerne le problème général de l'intelligence artificielle et comment, de manière générale, une science explicative comme la philosophie peut aider une science applicative comme l'Intelligence Artificielle (partie 3), et

---

[29] La notion de « spécialistes » pour désigner la communauté de chercheur travaillant sur l'Intelligence Artificielle est proposée par Andler dans l'avant-propos de Dreyfus, H.L. 1979. *Op. cit.*, p. X-XI.

[30] La section 2.1.1 décrit plus en détail les axes de recherche de ces deux communautés de chercheurs et leurs manières respectives d'aborder la notion d'intelligence. Elle montre également que ces deux communautés sont très souvent exclusives et que, du même coup, le problème de l'IA faible et celui de l'IA forte sont traités, dans la majorité des cas, de manière indépendante.

[31] Ces « machines concrètes » peuvent désigner aussi bien des robots (entités physiques) que des programmes (entités virtuelles), dans la mesure où les programmes informatiques peuvent recevoir une implémentation physique. Dans ce contexte, le rôle de l'Intelligence Artificielle n'est pas strictement limité à la réalisation physique de programmes ou de robots, mais cela reste néanmoins son objectif final. On s'éloigne ainsi de la « science informatique » en général qui, notamment avec l'algorithmique, a également des objectifs purement théoriques (*i.e.* des objets qui ont une valeur scientifique en dehors de toute réalisation physique).

réciproquement (partie 4). Les collaborations présentées dans ces parties peuvent ainsi être interprétées comme des cas particuliers de collaboration entre une science appliquée et la philosophie.

Chapitre 1.4. Stratégie de recherche

Comme annoncé en chapitre 1.2, dans ce mémoire, nous sommes moins intéressés par les solutions apportées au problème général de l'intelligence artificielle que par les *méthodes de résolution* elles-mêmes, et plus précisément celles qui font intervenir une collaboration entre Intelligence Artificielle et philosophie. Comme nous l'avons vu dans le chapitre précédent, ces deux disciplines semblent travailler sur deux sous-problèmes divergents : l'IA faible et l'IA forte. La stratégie de ce mémoire consiste à mettre en évidence des relations logiques ou épistémologiques entre ces deux sous-problèmes (*cf.* figure 1, étape 3) pour inférer des modes de collaboration au niveau des disciplines : est-ce que les travaux portant sur un des deux problèmes peuvent aider à la résolution du second problème ? Les rapports entre IA faible et IA forte servent alors à *exemplifier* les liens qui peuvent exister entre Intelligence Artificielle et philosophie (*cf.* figure 1, étape 4).

Relations possibles entre IA faible et IA forte

À ce titre, les deux sous-problèmes de l'intelligence artificielle peuvent être liés de plusieurs manières. Premièrement, on peut trouver des rapports *logiques*, par exemple dans le cas du béhaviourisme logique (section 2.2.1), indiquant qu'un résultat concernant l'un des deux problèmes induit logiquement un résultat quant au second problème. Il s'agit alors d'une relation *analytique* au sens où elle dépend de la seule formulation des deux problèmes, indépendamment de toute découverte épistémique concernant la nature des machines. Le second type de rapports est *épistémologique*. Il correspond à des relations *synthétiques* entre IA faible et IA forte, c'est-à-dire des relations qui ne dépendent pas seulement de la forme logique des deux problèmes, mais qui résultent d'une découverte sur le plan épistémique (issu de résultats théoriques ou empiriques).

Tout d'abord, nous utilisons plusieurs formules pour exprimer les positions quant aux possibilités de l'IA faible et de l'IA forte, indépendamment l'une de l'autre :

- « IA faible » signifie que le problème de l'IA faible admet une solution, *i.e.* qu'« il est possible de construire une machine qui *simule* l'intelligence. »
- « non IA faible » signifie au contraire que l'IA faible est impossible, *i.e.* qu'« aucune machine ne peut *simuler* l'intelligence. »
- « IA forte » signifie réciproquement que le problème de l'IA forte admet une solution, *i.e.* qu'« il est possible de construire une machine qui est *réellement* intelligente. »
- « non IA faible » signifie que l'IA forte est impossible, *i.e.* qu'« aucune machine ne peut être *réellement* intelligente. »

Ces formules peuvent s'appliquer à des ensembles particuliers de machines au lieu d'être interprétées de manière générale. On notera par exemple « IA faible (X) » pour dire qu'au moins une machine de l'ensemble X simule l'intelligence et « IA faible (x) » pour dire que c'est le cas de la machine x plus particulièrement.

D'autres formules servent à exprimer les relations (analytiques ou synthétiques) entre les deux problèmes :

- « IA faible → IA forte » signifie qu'une machine simulant l'intelligence est toujours « véritablement intelligente ». Il peut s'agir d'une relation analytique (comme dans le cas du béhaviourisme logique, section 2.2.1) ou d'une relation synthétique (reste du chapitre 2.2). Cela ne prouve pas la possibilité de l'un ou l'autre des deux problèmes. Pour autant, cette relation peut être défendue sans qu'on soit certain qu'une machine puisse effectivement simuler l'intelligence (« IA faible »). En outre, si tel est le cas, on conclura immédiatement à partir de cette relation qu'une telle machine est « véritablement intelligente » (« IA forte »).

- « IA forte → IA faible » signifie qu'une machine « véritablement intelligente » parvient toujours à simuler l'intelligence.

- « IA faible ↔ IA forte » signifie que les machines simulant l'intelligence et les machines « véritablement intelligentes » forment un même ensemble d'individus (l'un ne va pas sans l'autre). Cette relation est la conjonction des deux relations précédentes.

- On parlera également d'« indépendance des deux problèmes » lorsqu'aucune de ces relations ne tient. Une machine pourrait éventuellement simuler l'intelligence sans être « véritablement intelligente » et réciproquement.

Ces formules peuvent également être niées. « non (IA faible → IA forte) » signifie par exemple que la relation « IA faible → IA forte » n'est pas valide. Cette affirmation reposera notamment sur la mise en évidence d'un contre-exemple du type « IA faible (x) et non IA forte (x) » selon lequel la machine x simule l'intelligence sans être « véritablement intelligente ». D'autres formules enfin, plus nuancées que celles présentées ici, peuvent apparaître dans ce mémoire. C'est par exemple le cas de l'affirmation « non IA forte → IA faible peu probable » défendue par Dreyfus (*cf.* section 2.4.2) et signifiant qu'une machine qui n'est pas « véritablement intelligente » a beaucoup de mal à simuler l'intelligence. Selon cette formule, il est donc très difficile de construire une telle machine, bien que cela ne soit pas impossible. Il s'agit d'une version moins stricte de la relation « non IA forte → non IA faible ».

<table>
<tr><td>« Deux communautés distinctes »</td><td rowspan="3">Indépendance des deux problèmes</td><td>Section 2.1.1</td></tr>
<tr><td>Béhaviourisme méthodologique (Turing)</td><td>Section 2.1.2</td></tr>
<tr><td>Béhaviourisme logique (Ryle)</td><td rowspan="2">Section 2.2.1</td></tr>
<tr><td>Béhaviourisme logique (Hempel)</td><td rowspan="3">« IA faible → IA forte »</td></tr>
<tr><td>Impossibilité par principe de l’IA faible</td><td>Section 2.2.2</td></tr>
<tr><td>Nécessité en pratique de l’IA forte (Levesque)</td><td>Section 2.2.3</td></tr>
<tr><td>Critique de la relation « IA faible → IA forte »</td><td>« non (IA faible → IA forte) »</td><td>Section 2.2.4</td></tr>
<tr><td>Computationnalisme (Newell & Simon)</td><td>« IA faible ↔ IA forte »</td><td>Section 2.3.1</td></tr>
<tr><td rowspan="2">Critique de l’IA (Searle)</td><td>« non (IA faible → IA forte) »</td><td rowspan="2">Section 2.3.2</td></tr>
<tr><td>« non IA forte »</td></tr>
<tr><td rowspan="3">Critique de l’IA (Dreyfus)</td><td>« non IA faible (GPS) »</td><td>Section 2.4.1</td></tr>
<tr><td>« non IA forte → IA faible peu probable »</td><td>Section 2.4.2</td></tr>
<tr><td>« IA forte → IA faible »</td><td>Section 2.4.3</td></tr>
</table>

Figure 2 : résumé des relations présentées dans la partie 2

Plan du mémoire

La partie 2 présente les positions de défenseurs de l’Intelligence Artificielle (Turing, Newell et Simon) et celles de critiques célèbres (Searle et Dreyfus) en ce qui concerne la possibilité des deux problèmes et leurs relations possibles (*cf.* figure 2 ci-dessus). Les travaux présentés retracent sommairement les débuts de l’Intelligence Artificielle et les grands paradigmes qu’elle a empruntés. Ils sont résumés dans un tableau synthétique page 66 (figure 3). Chacune des positions présentées, lorsqu’elle présuppose des relations entre IA faible et IA forte, permet de définir une stratégie collaborative entre Intelligence Artificielle et philosophie. Nous soutenons notamment que la relation « IA forte → IA faible », défendue par Dreyfus, est propice à un véritable échange entre philosophie et Intelligence Artificielle. Dans les parties 3 et 4, cette relation est exploitée dans des travaux faisant intervenir les deux disciplines. Ces exemples de collaboration montrent que la démarche empruntée par Dreyfus est féconde et qu’elle permet à la philosophie et à l’Intelligence Artificielle de se soutenir l’une et l’autre dans leurs recherches. D’autres relations, comme celle défendue par Levesque, sont également exploitées dans cette optique.

La partie 3 met la philosophie (et pas seulement la philosophie de l’esprit) au service de l’Intelligence Artificielle. Nous montrons comment, à partir de la relation « IA forte → IA faible », des théories et des concepts philosophiques peuvent être utilisés pour définir de nouveaux modèles pratiques en Intelligence Artificielle et ainsi résoudre des difficultés techniques. Le travail de collaboration présenté dans le chapitre 3.2 montre par exemple comment le concept d’« émergence

épistémique », développé par la philosophie britannique au tournant du XIX^{e} siècle, peut être exploité en Intelligence Artificielle pour la simulation de systèmes complexes et la résolution de problèmes distribués. Plus généralement, il s'agit d'un exemple de collaboration où une discipline *explicative* est utilisée dans un cadre *applicatif*.

La partie 4 opère le renversement formulé par Andler en mettant l'Intelligence Artificielle, ainsi que d'autres sciences de l'ingénieur, au service de la philosophie de l'esprit. À partir de la contraposée « non IA faible → non IA forte » de la relation utilisée par Dreyfus, nous montrons comment des machines concrètes de l'Intelligence Artificielle permettent de tester des théories philosophiques sur la nature de la cognition. Notons que d'autres relations sont exploitées dans cette partie, notamment la relation « IA faible → IA forte » dans le chapitre 4.3. Ces travaux constituent enfin des exemples de collaborations où une science applicative est utilisée dans le cadre de polémiques proprement philosophiques.

Contexte de recherche

Ce travail de recherche a été réalisé dans le cadre d'un Master 2 Recherche en philosophie à l'Université Pierre-Mendès-France (UPMF, Grenoble) sous la direction de Denis Perrin, enseignant-chercheur du groupe Philosophie, Langages & Cognition (PLC, Grenoble). Il continue un mémoire de recherche[32] réalisé en Master 1, également à l'UPMF, sous la direction de Max Kistler et dont les résultats ont été en partie repris dans le chapitre 4.3. Durant l'année de Master 2, un travail de publication[33] a été réalisé pour le *3^{e} colloque des doctorants et des jeunes chercheurs*, organisé par le groupe PLC en juin 2011, dont l'objectif était de « repenser les rapports entre science(s) et philosophie. » À cette occasion, les travaux exposés dans ce mémoire ont été présentés comme des cas particuliers de collaboration entre une discipline scientifique (l'Intelligence Artificielle) et la philosophie (notamment la philosophie de l'esprit). L'article rédigé à la suite de cette présentation, accepté pour publication aux actes du colloque, est fourni en annexe (pages 119-133) en tant que résumé étendu des questions traitées dans ce mémoire. Néanmoins, la partie 4 concernant le renversement des rapports entre philosophie et Intelligence Artificielle n'est présentée que très succinctement dans l'article, sous la forme de perspectives de recherches.

---

[32] Lamarche-Perrin, R. 2010. *Le Test de Turing pour évaluer les théories de l'esprit*. Mémoire de Master, Kistler, M. (dir.). Grenoble : Université Pierre-Mendès-France, sept. 2010.

[33] Lamarche-Perrin, R. 2011b. « Des collaborations possibles entre Intelligence Artificielle et philosophie de l'esprit. » *Colloque des doctorants et jeunes chercheurs du groupe de recherche PLC : Repenser les rapports entre science(s) et philosophie*. Grenoble : Philosophie, Langages et Cognition, juin 2011. En cours d'édition.

Par ailleurs, ces travaux de philosophie ont été menés en parallèle à des recherches en Intelligence Artificielle dans le cadre d'une thèse à l'Université de Grenoble (UdG), sous la direction d'Yves Demazeau et de Jean-Marc Vincent du Laboratoire d'Informatique de Grenoble (LIG). Certains résultats ont été repris dans le chapitre 3.2 de ce mémoire pour donner un exemple de collaboration étroite entre philosophie et Intelligence Artificielle. Elle a également fait l'objet d'une présentation scientifique[34] dans le cadre de la plate-forme annuelle de l'Association Française d'Intelligence Artificielle (AFIA) en mai 2011.

[34] Lamarche-Perrin, R. 2011a. « Conceptualisation de l'émergence : dynamiques microscopiques et analyse macroscopique des SMA. » *Atelier Futur des Agents et des Multi-Agents (FUTURAMA'11)*. Chambéry : Plateforme AFIA 2011, mai 2011.

Partie 2. Des relations possibles entre IA faible et IA forte

Cette partie retrace les débuts de l'Intelligence Artificielle et les critiques qui ont été formulées à son égard dans les années 80. Les différentes approches qui y sont présentées reposent implicitement sur des relations entre les deux sous-problèmes de l'intelligence artificielle, nommément l'IA faible et l'IA forte (*cf.* chapitre 1.3). Ces relations nous permettent de formaliser des stratégies collaboratives génériques, qui seront exploitées dans la suite du mémoire. Le chapitre 2.1 s'intéresse à l'indépendance des deux problèmes, avec notamment le béhaviourisme méthodologique d'Alan M. Turing. Le chapitre 2.2 présente des positions en faveur de la relation « IA faible → IA forte », avec notamment le béhaviourisme logique. Le chapitre 2.3 s'intéresse à l'hypothèse computationnaliste, telle que formulée par Allen Newell et Herbert A. Simon, et à la relation d'équivalence « IA forte ↔ IA faible ». La critique de John R. Searle y est également abordée. Le chapitre 2.4, enfin, se penche sur l'ouvrage critique de Dreyfus et les modes de collaboration qu'il permet de définir. Celui-ci débouche notamment sur la relation « IA forte → IA faible » (section 2.4.3) qui sera largement utilisée dans la suite de ce mémoire.

Chacune de ces positions, et les relations associées, autorisent ou interdisent des modes de collaborations entre philosophie et Intelligence Artificielle. Ceux-ci sont explicités dans chaque section et synthétisés dans le tableau page 67 (figure 3). Nous pensons que les stratégies collaboratives induites par le travail de Dreyfus sont les plus favorables à un véritable échange entre les deux disciplines. Elles seront exploitées dans la partie 3 pour mettre la philosophie au service de l'Intelligence Artificielle (à partir de la relation « IA forte → IA faible ») et dans la partie 4 pour réaliser le renversement imaginé par Andler (*cf.* chapitre 1.1) et pour mettre l'Intelligence Artificielle au service de la philosophie (à partir de la relation « non IA faible → non IA forte »).

Chapitre 2.1. L'indépendance des deux problèmes

Section 2.1.1. Deux communautés distinctes

Dans le chapitre 1.3, nous avons évoqué le fait que les chercheurs travaillant sur le problème de l'IA faible et ceux travaillant sur le problème de l'IA forte constituent deux communautés de recherche distinctes : les spécialistes de l'IA, d'une part, et les philosophes de l'esprit, d'autre part. Cette section explicite les préoccupations de ces deux domaines et montre que, dans la majorité des cas, les uns ne s'intéressent pas aux travaux des autres.

Les spécialistes de l'IA et le problème de l'IA forte

D'une part, les spécialistes de l'IA, selon le constat de Russell & Norvig, « prennent l'hypothèse de l'IA faible pour acquise et ne se préoccupent pas de l'hypothèse de l'IA forte. »[35] Pour beaucoup, l'Intelligence Artificielle est avant tout un ensemble de théories, d'outils et de méthodes destinées à assister la résolution de problèmes scientifiques, de problèmes techniques ou même de problèmes de la vie quotidienne. Les programmes sont donc évalués en fonction de la qualité des solutions qu'ils apportent, et non en fonction des similitudes qu'ils entretiennent avec l'intelligence humaine ou avec toute autre « intelligence véritable » (IA forte). Généralement, un problème est caractérisé par une *mesure de performance* que l'on cherche à optimiser. Dans le cas du « joueur d'échecs », par exemple, il s'agit de gagner un maximum de parties contre un joueur de référence (un champion d'échecs ou un programme étalon). La manière dont le programme détermine ses stratégies importe peu à partir du moment où le résultat qu'il engendre est satisfaisant[36].

Notons que de nombreux problèmes font intervenir simultanément plusieurs mesures de performance. De plus, elles peuvent être définies relativement aux ressources computationnelles nécessaires pour le bon fonctionnement du programme. Dans le cas du joueur d'échecs, on cherche par exemple à minimiser le temps de calcul afin d'interagir avec un utilisateur humain. Les

---

[35] « Most of AI researchers take the weak AI hypothesis for granted, and don't care about the strong AI hypothesis. » [Notre traduction] Russell, S.J., Norvig, P. 2003. *Artificial Intelligence: A Modern Approach,* 3rd ed. Upper Saddle River, NJ : Prentice Hall, p. 1020.

[36] Dans la plupart des cas, les programmes d'échecs évaluent les coups possibles en fonction des parties engendrées. Les méthodes d'évaluation peuvent différer des stratégies que l'on apprend dans les écoles d'échecs, mais la différence principale réside dans le nombre de coups évalués. La puissance de calcul d'un ordinateur lui permet en effet de parcourir de manière plus systématique l'arbre des parties possibles et ainsi de compenser son manque de « stratégies intuitives ». *Cf.* également la critique des programmes computationnalistes par Dreyfus (section 2.4.1) et plus particulièrement la page 56 concernant le jeu d'échecs.

contraintes temporelles éliminent ainsi certaines façons de procéder (*e.g.*, parcours complet de l'arbre des parties possibles) et introduisent une contrainte de similitude avec l'intelligence humaine : un programme doit pouvoir fonctionner *à une échelle humaine* pour une bonne interaction. Ainsi, nous pouvons dire que l'optimisation du coût computationnel d'un programme (en terme de temps de calcul ou d'espace mémoire) induit un rapport de commensurabilité entre intelligence artificielle et intelligence humaine. Il s'agit cependant d'une notion faible de similitude dans la mesure où elle répond à des contraintes pratiques. Quels sont les moyens technologiques dont nous disposons ? Quelles sont les contraintes d'interactivité nécessaires à l'utilisation humaine de tels programmes ? Il ne s'agit pas, comme dans le cas du problème de l'IA forte, d'une notion de similitude profonde, répondant à des contraintes de « véritable intelligence » issues par exemple de la psychologie du comportement, des sciences cognitives ou de la philosophie de l'esprit. Les contraintes d'interactivité, au contraire, relèvent bien du domaine de l'IA faible. Elles assurent la simulation *efficace* des comportements intelligents.

Les mesures de performance définissent ainsi des critères d'intelligence *ad hoc*, en fonction du problème à résoudre et des exigences technologiques sous-jacentes. Aucune similitude avec les modes de résolutions humains n'est *a priori* requise. Les ingénieurs en aéronautique, lorsqu'ils conçoivent un avion, n'ont pas pour objectif d'imiter le vol des oiseaux[37]. L'objectif est avant tout de construire un moyen de transport rapide, sûr et économique. La question de savoir si un avion « simule le vol » ou s'il « vole véritablement » a peu d'importance pour un ingénieur. Russell & Norvig affirment ainsi que, pour les spécialistes de l'IA, la question relève surtout d'aspects linguistiques : est-ce que pour « voler véritablement » il faut *battre des ailes* ? Ou est-ce qu'il suffit de *se déplacer au-dessus du sol* ? La question de la « véritable intelligence », au même titre que celle du « vol véritable », ne présente en pratique que peu d'intérêts pour l'activité quotidienne des spécialistes de l'IA. « Tant que leurs programmes fonctionnent, ils ne se préoccupent pas de savoir s'il faut appeler cela simulation de l'intelligence ou intelligence réelle. »[38] En pratique, il faudrait préférer la notion d'« implémentation » à celle de « simulation »[39], puisqu'elle traduit mieux l'activité

[37] Ce type de remarques, ici empruntée à Russell, S.J., Norvig, P. 2003. *Op. cit.*, p. 3 et 1021, raille la tendance anthropomorphique de certaines acceptions de l'intelligence, notamment en philosophie (*cf.* par exemple le test de Turing dans la section suivante). Les chercheurs en IA utilisent en réalité une multitude de définitions *ad hoc* qui n'ont rien à voir avec de telles définitions empruntées à la nature (vol des oiseaux, intelligence humaine).

[38] « as long as their program works, they don't care whether you call it a simulation of intelligence or real intelligence. » [Notre traduction] *Ibid.*, p. 1020.

[39] *Ibid.*, p. 1027.

des spécialistes consistant à réaliser une fonction donnée à l'aide d'une machine physique. La distinction IA faible/IA forte, reposant justement sur cette notion de *simulation* des comportements (*e.g.* « battre des ailes »), est de ce fait ignorée par la plupart des chercheurs en Intelligence Artificielle qui se contentent d'*implémenter* des fonctions (*e.g. «* se déplacer au-dessus du sol »).

Pour une grande majorité de spécialistes donc, l'IA faible est *implicitement* possible (« IA faible »). L'optimisation des mesures de performance en témoigne. Cependant, la distinction IA faible/IA forte n'est pas *explicitement* prise en compte et le problème de l'IA forte, quant à lui, n'est pas même évoqué.

Les philosophes et le problème de l'IA faible

D'autre part, il apparaît que la majorité des philosophes intéressés par le problème de l'IA forte, ne s'intéressent pas pour autant aux travaux de l'IA faible. L'exemple le plus marquant est celui de John R. Searle qui, dans un article introduisant son argument séminal contre l'Intelligence Artificielle[40], déclare « n'avoir aucune objection contre les prétentions de l'IA faible. »[41] Il annonce ainsi que son travail concerne *uniquement* la question de la « véritable intelligence », ici liée à la notion d'intentionnalité. On ne trouve effectivement aucune référence au problème de l'IA faible dans la suite de l'article. Avec cette brève déclaration, Searle semble indiquer que les outils développés dans le contexte de l'IA faible ont très peu de chose à voir avec la philosophie. Dans un second ouvrage, issue des *1984 Reith Lectures* données pour la BBC[42], dans lequel Searle traite également de la question de l'IA forte, le terme « IA faible » n'est pas même mentionné une seule fois, alors que le terme « IA forte » apparaît huit fois en vingt-sept pages[43].

Ce désintéressement pour les questions propres à l'IA faible est également présent dans la très polémique expérience de pensée du « remplacement de cerveau ». Introduite par Clark Glymour

---

[40] Searle, J.R. 1980. « Minds, Brains, and Programs. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 67-88. *Cf.* également la section 2.3.2 concernant l'argument de la « chambre chinoise ».

[41] « I have no objection to the claims of weak AI, at least as far as this article is concerned. » [Notre traduction] *Ibid.*, p. 67. Notons cependant que dans cet article, Searle définie l'IA faible (*weak AI*) comme un ensemble d'outils informatiques pour l'étude de l'esprit. Il ne s'agit pas à proprement parler de l'acception que nous avons retenue, mais nous généralisons ici la position de Searle quant au projet plus général des spécialistes de l'IA.

[42] Searle, J.R. 1984. *Du cerveau au savoir : conférences Reith 1984 de la BBC*. [*Minds, Brains and Science: The 1984 Reith Lectures.*] Chaleyssin, C. (trad.). Paris : Hermann, 1985.

[43] *Ibid.*, p. 37-63.

dans les années 70, elle a été longuement discutée comme un point essentiel du débat concernant la conscience des machines[44]. Elle fait donc intervenir une question emblématique de l'IA forte : « les machines peuvent-elles avoir une conscience et, en particulier, une vie phénoménale ? » L'expérience propose de remplacer un-à-un les neurones d'un cerveau humain par autant de composants électroniques capables de reproduire précisément le comportement des neurones amputés (à supposer qu'une telle technologie soit disponible). On part du principe que tout le monde s'accorde à dire que le cerveau humain dispose, au début de l'expérience, d'une vie consciente. Comme le remplacement se fait neurone par neurone, il est difficile d'imaginer, si on se met à la place de l'individu conscient, que sa vie phénoménale disparaisse brutalement à un instant donné du processus. Dans ce cas, il faut admettre, à la fin de l'expérience, que rien n'a changé de ce côté. L'expérience constitue donc un argument en faveur de l'IA forte dans la mesure où le cerveau finalisé, puisqu'il est entièrement fait de composants électroniques, est une machine numérique douée de vie consciente (« IA forte (cerveau remplacé) »). Nous ne souhaitons pas ici discuter de la pertinence d'un tel argument, ni en présenter les nombreuses critiques[45], mais souligner que ce débat important de la philosophie de l'esprit, concernant le problème de la conscience appliqué aux machines, ne participe *d'aucune manière* aux polémiques de l'IA faible. Il s'agit en vérité d'une question hautement philosophique, faisant intervenir un appareillage conceptuel propre à la philosophie de l'esprit[46], dont l'enjeu est avant tout la défense ou la critique du fonctionnalisme[47] (une théorie avant tout philosophique), et dont les retombées ne sauraient perturber le travail des spécialistes de l'IA. Dans l'expérience de pensée, on suppose toujours que le cerveau final se *comporte* de la même façon que le cerveau initial (« IA faible (cerveau remplacé) »), mais cela n'est pas la question centrale et, surtout, cet exemple de machine simulant l'intelligence est sans intérêt pour les spécialistes de l'IA. Après tout, la technologie imaginée par Glymour est aberrante pour un spécialiste, et surtout son objectif *n'est pas* la conception de telles machines conscientes (*cf.* sous-section précédente). Réciproquement, les recherches en IA faible ne peuvent avoir la moindre

[44] *Cf.* par exemple Moravec, H. 1988. *Mind Children*. Cambridge, MA : Harvard University Press ; Searle, J.R. 1992. *The Rediscovery of the Mind*. Cambridge, MA : MIT Press, p. 65-82 ; Russell, S. J., Norvig, P. 2003. *Op. cit.*, p. 1029-1031 ; Chalmers, D. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, p. 253-263.

[45] Voir notamment la critique de Searle et la réponse de Chalmers. Searle, J. 1992. *Op. cit.*, p. 65-70 ; Chalmers, D. 1996. *Op. cit.*, p. 253-263.

[46] David Chalmers fait référence à cette expérience de pensée sous le terme « les qualia qui s'estompent » (*fading qualia*), rejoignant ainsi la question fondamentale de la philosophie de l'esprit concernant la cause et la nature des expériences sensibles. *Ibid.*

[47] Russell, S. J., Norvig, P. 2003. *Op. cit.*, p. 1029.

influence sur celles des philosophes dans la mesure où elles divergent quant à leurs objectifs fondamentaux : produire un comportement intelligent et produire une vie consciente.

Dès l'abord, il y aurait un fossé scientifique, au moins au niveau des pratiques disciplinaires, entre les deux sous-problèmes de l'intelligence artificielle. D'une part, les spécialistes de l'IA, occupés par des difficultés technologiques, s'intéressent au problème de l'IA faible et, d'autre part, les philosophes de l'esprit s'occupent de celui de l'IA forte. Aucune des deux communautés ne considère que les travaux réalisés par l'autre communauté peuvent avoir des conséquences sur leurs propres recherches. Dans la suite de ce mémoire, nous présentons cependant la position de chercheurs qui s'intéressent simultanément aux deux sous-problèmes : pour en affirmer l'indépendance (*cf.* section suivante) ou, au contraire, pour montrer qu'il existe des liens logiques ou épistémologiques entre IA faible et IA forte.

Section 2.1.2. Le béhaviourisme méthodologique de Turing

Un dispositif expérimental pour l'IA faible

Dans son article de 1950, Alan M. Turing propose un dispositif concret pour déterminer si « les machines sont capables de penser »[48]. Il s'agit du désormais fameux « test de Turing » : un observateur humain communique avec deux autres individus à l'aide d'un téléscripteur, d'un terminal, ou de tout autre appareil permettant de recevoir et d'émettre des messages écrits. Il est impossible à l'observateur de percevoir directement ses interlocuteurs, d'en connaître l'aspect ou d'en entendre la voix. Il peut seulement converser avec eux, à l'écrit, via le dispositif. Parmi les deux individus, une machine se livre au « jeu de l'imitation »[49] : elle est programmée de manière à se faire passer pour un homme. La tâche de l'observateur consiste à déterminer, par le seul jeu de la conversation, lequel des deux individus est un homme et lequel est une machine. On dit alors que la machine « passe le test avec succès » lorsqu'elle parvient à tromper l'observateur quant à sa nature[50]. Il faut alors conclure, selon Turing, que la machine *pense réellement*. En effet, si l'observateur ne détecte aucune différence essentielle lors de sa conversation avec l'homme (de chair et de sang) et avec la machine, s'il est incapable de distinguer de cette manière la « nature

[48] « 'Can machines think?' » Turing, A.M. 1950. « Computing Machinery and Intelligence. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 40.

[49] « the 'imitation game' » *Ibid.*, p. 40.

[50] En réalité, il faudrait considérer que le test est réussi lorsque la machine trompe son interlocuteur au moins une fois sur deux. En effet, quand il ne peut distinguer l'homme de la machine, l'observateur peut répondre au hasard. On dit également que le test de Turing est réussi « à X% » lorsque la machine trompe son interlocuteur X% du temps, 50% étant considéré comme un succès complet.

profonde » de ses interlocuteurs, c'est qu'une telle distinction *n'est pas* essentielle. Ainsi, si on affirme que l'homme *pense* lorsqu'il répond à l'observateur, il faut être prêt à dire la même chose de la machine.

Le test de Turing définit ainsi une méthode empirique pour détecter une faculté psychologique à partir de ses propriétés observables. En pratique donc, le test utilise l'acception comportementale des facultés de l'esprit. Par exemple, on dira qu'une machine *est* intelligente lorsqu'elle *se comporte* à la manière d'un homme intelligent[51]. Turing hérite ainsi du *béhaviourisme méthodologique*, apparu en psychologie au début du siècle, selon lequel les facultés psychologiques devraient être étudiées en fonction de leurs seules qualités observables, c'est-à-dire des comportements qu'elles engendrent[52]. Le test de Turing est avant tout un dispositif expérimental permettant de répondre au problème de l'IA faible : une machine *simule* l'intelligence lorsqu'elle passe avec succès le test de Turing. Pour une machine x donnée, le test permet de déterminer si on a « IA faible (x) » ou « non IA faible (x) ». Le béhaviourisme méthodologique incite ainsi à étudier les comportements et à ne pas faire grand cas des propriétés psychologiques internes. En évitant ainsi de se confronter à de telles propriétés, le test de Turing reste muet quant au problème de l'IA forte : puisque, en dehors des comportements engendrés, on ne peut faire de distinctions concernant les propriétés internes des individus, celles-ci ne peuvent pas servir à la définition des concepts psychologiques. La notion de « véritable intelligence », inaccessible à l'examen scientifique, est ainsi laissée pour compte[53].

L'indépendance épistémologique des deux problèmes

Cependant, il est intéressant de noter que Turing se positionne également vis-à-vis du problème de la conscience (appliqué aux machines). Il rappelle d'abord qu'il est strictement impossible de connaître la vie phénoménale d'une machine, à moins bien sûr d'« *être* la machine et

---

[51] Ici, le comportement intelligent est borné à l'échange langagier entre deux individus. On pourrait exiger une évaluation comportementale plus large (perception, motricité, aptitudes sociales, *etc.*) pour définir l'intelligence.

[52] C'est notamment la position de John B. Watson qui définit la psychologie comme étant « la science des comportements », et non « la science des phénomènes mentaux ». Watson, J.B. 1913. « Psychology as the Behaviorist Views it. » *Psychological Review*, vol. 20, p. 158-177.

[53] Aujourd'hui, de nombreuses compétitions proposent de confronter des programmes au test de Turing. Le *Loebner Prize*, inauguré en 1991, promet 100.000 dollars au premier ordinateur capable de passer le test avec succès (http://www.loebner.net/Prizef/loebner-prize.html). Les compétitions annuelles organisées à cet effet n'ont pas encore révélé de telles machines.

de se sentir penser soi-même. »[54] Devant cet obstacle épistémique, connu en philosophie sous le nom de « problème des autres esprits »[55], Turing recommande d'avoir « la convention polie (*the polite convention*) que tout le monde pense. »[56] C'est effectivement la solution que nous adoptons quotidiennement pour échapper au solipsisme : par convention, lorsqu'on interagit avec un autre individu, on suppose que celui-ci est conscient, et notamment qu'il *pense*.

Cependant, cette recommandation n'implique pas pour Turing un engagement profond en ce qui concerne le problème de la conscience, mais seulement une facilité méthodologique. Turing décide de travailler seulement à partir de ce qui est directement observable, en accord avec le projet du béhaviorisme méthodologique. Mais pour ce qui est de la conscience, de la vie phénoménale et des questions relatives à la notion de « véritable intelligence », Turing botte en touche : il s'agit d'un autre problème, indépendant de la question qui l'intéresse ici. En particulier, le problème de la conscience n'a pas besoin d'être résolu pour mettre en place le test de Turing[57]. En d'autres termes, la résolution du problème de l'IA forte n'est pas *nécessaire* à la résolution du problème de l'IA faible : « non (IA faible → IA forte) ». Elle n'est pas non plus *suffisante*, puisque le test de Turing définit des conditions pratiques pour identifier des comportements intelligents, ce que ne fait pas l'étude de la conscience : « non (IA forte → IA faible) ». Turing ne prétend pas que son test résolve le problème de la conscience des machines. Réciproquement, il affirme que la résolution des mystères de la conscience ne peut pas remplacer l'utilisation d'un test empirique. En outre, il se pourrait qu'une machine sans conscience parviennent à passer le test de Turing (« IA faible (x) et non IA forte (x) ») et, réciproquement, qu'une machine consciente ne répondent pas aux conditions de vérification du test, et que ces comportements soit ainsi déclarés non-intelligents (« IA forte (x) et non IA faible (x) »).

Ainsi, le béhaviorisme de Turing défend l'*indépendance épistémologique* des deux problèmes. Ceci explique notamment le constat de la section précédente selon lequel les philosophes et les spécialistes de l'IA forment des communautés scientifiques distinctes. Ils

---

[54] « to *be* the machine and to feel oneself thinking » [Notre traduction] Turing, A.M. 1950. *Op. cit.*, p. 52.

[55] *Cf.* par exemple Hyslop, A. 2009. « Other Minds. » *In Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/other-minds/, mis en ligne le 6 oct. 2005, révisé le 23 nov. 2009, consulté le 19 juillet 2011.

[56] « the polite convention that everyone thinks » [Notre traduction] Turing, A.M. 1950. *Op. cit.*, p. 52.

[57] « I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper. » *Ibid.*, p. 53.

travaillent en effet sur des problèmes indépendants. Leur collaboration n'est donc ni nécessaire, ni utile, à l'avancée des deux domaines.

Bilan sur le béhaviourisme méthodologique

- Le test de Turing est un dispositif expérimental permettant de répondre au problème de l'IA faible : est-ce qu'une machine simule ou non des comportements intelligents ? Turing ne se prononce pas sur une éventuelle réponse positive, mais fait part de son optimiste quant au fait qu'une machine passera un jour le test avec succès.

- Turing ne se prononce pas quant au problème de l'IA forte lié à la conscience des machines. Il s'en remet à une « convention polie » qui constitue une position relativement faible, mais qui a le mérite d'être compatible avec sa méthode comportementale.

- Il affirme l'indépendance épistémologique des deux problèmes. Il est ainsi possible de répondre à la question de l'IA faible, sans avoir résolu le problème de l'IA forte (et réciproquement).

Chapitre 2.2. « IA faible → IA forte »

Section 2.2.1. La réduction du béhaviourisme logique

La réduction logique de l'IA forte

Une forme plus radicale de béhaviourisme remet en cause la notion même de « véritable intelligence ». Il s'agit du *béhaviourisme logique*, dont Carl G. Hempel, figure importante de l'empirisme logique, est l'un des théoriciens les plus connus. Son *Analyse logique de la psychologie* défend l'idée que « *toutes les conditions de vérification* [*d'un*] *énoncé psychologique* se présentent sous la forme de formules de contrôles physiques. »[58] Dire que « Paul à mal aux dents », par exemple, c'est dire qu'il hurle de douleur, que son activité neurologique répond à certains critères bien définis et qu'un professionnel découvrirait certainement, lors d'un examen dentaire, une molaire cariée. Par conséquent, en psychologie, la signification des concepts mentaux est réductible aux relations entre des stimuli et des comportements observables. À la différence de son homologue méthodologique, la version logique du béhaviorisme opère ainsi une élimination des énoncés mentalistes.

Le test de Turing pourrait bien constituer un tel outil de « contrôle physique ». Dès lors, il acquiert une place prépondérante dans la définition des facultés psychologiques. La version *logique* du test, à opposer au test *méthodologique* véritablement défendu par Turing, est un dispositif expérimental permettant de détecter la « véritable intelligence ». Il remplace ainsi les énoncés de l'IA forte (*e.g.*, « cette machine est consciente » ou plus généralement « cette est *véritablement* intelligente ») par des énoncés de l'IA faible (*e.g.*, « cette machine passe le test de Turing avec succès »). La « véritable intelligence » est alors une condition nécessaire aux comportements intelligents, de manière strictement analytique[59]. Si Turing avait opté pour un béhaviourisme logique, il aurait donc défendu l'implication logique « IA faible → IA forte », et non l'indépendance des deux problèmes (*cf.* section précédente). Cela ne signifie pas que la résolution du problème de l'IA forte est nécessaire à la résolution du problème de l'IA faible (*relation épistémologique* à laquelle s'oppose Turing), mais que la signification de la « véritable intelligence » réside dans ses « contrôles physiques » (*relation logique* défendue par le béhaviorisme logique). Ainsi, pour certains philosophes

---

[58] Hempel, C. G. 1935. « L'analyse logique de la psychologie. » [« The Logical Analysis of Psychology. »] *In* Fisette, D. (éd.), Poirier, P. (éd.). 2002. *Philosophie de l'esprit : psychologie du sens commun et sciences de l'esprit*. Paris : Vrin, p. 197-215.

[59] Au sens où la définition de la « véritable intelligence » est incluse dans celle des « comportements intelligents ». Autrement dit, « IA faible → IA forte » est vrai en vertu de sa seule signification.

comme Stevan Harnad, le test de Turing, *quoi qu'il arrive*, a le dernier mot en ce qui concerne la détection de la conscience[60]. Dès lors, la notion même de conscience n'a pas de signification en dehors de ses conditions de vérification comportementales. De même, pour les empiristes logiques, les significations des termes introspectifs de la psychologie classique sont équivalentes à leurs acceptions comportementales. Par conséquent, les seuls termes dont nous pouvons disposer en science sont issus d'outils empiriques, à la manière du test de Turing.

Le béhaviourisme logique élimine ainsi le rôle de la philosophie de l'esprit en ce qui concerne le problème de l'IA forte. En effet, celui-ci peut être résolu par les spécialistes de l'IA en s'appuyant sur la relation « IA faible → IA forte »[61]. Les philosophes n'ont pas leur mot à dire concernant le problème général de l'intelligence artificielle et les spécialistes prennent leur place concernant leurs sujets de prédilection tels que la conscience et les états mentaux. En allant plus loin que le béhaviourisme méthodologique, le béhaviourisme logique élimine entièrement la philosophie du champ scientifique, laissant la responsabilité épistémologique aux seuls spécialistes de l'IA, aussi bien pour l'IA faible que pour l'IA forte. Dans ce contexte, aucune collaboration n'est évidemment possible.

L'indépendance logique des deux problèmes

Bien qu'il soit difficile d'attribuer une étiquette à la philosophie de Gilbert Ryle, sa position est parfois interprétée comme une forme « douce » de béhaviourisme logique[62]. Selon l'opposition de Ryle au cartésianisme (critique d'origine wittgensteinienne), la conception mentaliste du langage commet une « erreur de catégorie » lorsqu'elle affirme l'*existence* d'états mentaux au même titre

---

[60] « The answer to our revised question – "What kinds of machines can be conscious (and how)?" has now come into methodological focus. The answer is: The kinds that can pass the Turing Test, and by whatever means are necessary and sufficient to pass the Turing Test. » Harnad, S. 2003. « Can a Machine Be Conscious? How? » *Journal of Consciousness Studies*, vol. 10, n°4, p. 67-75.

[61] Après réflexion, la relation complète défendue par le béhaviourisme logique est « IA faible ↔ IA forte ». En effet, pour Hempel, les énoncés mentalistes sont *équivalents* à des énoncés comportementaux. Ainsi, le béhaviourisme logique aurait pu être présenté dans le chapitre 2.3, justement intitulé « IA forte ↔ IA faible », puisqu'il défend une équivalence logique des deux problèmes. Cependant, pour respecter la chronologie des débuts de l'IA, cette position est présentée à la suite du béhaviourisme de Turing. En outre, c'est la partie suffisante « IA faible → IA forte » qui est utilisée pour réduire l'IA forte à l'IA faible (et non l'inverse), en toute cohérence avec le béhaviourisme méthodologique.

[62] *Cf.* la préface de Julie Tanney intitulée « Une cartographie des concepts mentaux » dans Ryle, G. 1949. *La Notion d'esprit*. [*The Concept of Mind*.] Stern-Gillet, S. (trad.), Tanney, J. (pref.). Paris : Payot & Rivages, 2005, p. 7-73.

que l'*existence* d'objets matériels. Dans le cas qui nous intéresse, l'erreur catégorielle consisterait à affirmer la dichotomie entre IA faible et IA forte, entre « intelligence simulée » et « véritable intelligence ». En effet, dire qu'une machine *est* intelligence (et lui attribuer par exemple des états mentaux) n'est en rien comparable au fait de dire qu'elle *simule* l'intelligence. Les propriétés comportementales et les propriétés psychologiques n'appartiennent tout simplement pas à la même catégorie logique. Dès lors, la conjonction des deux problèmes est impossible et les réductions d'un problème à l'autre sont « des réponses à des questions mal posées »[63]. Il apparaît ainsi que Ryle aurait pu défendre l'*indépendance logique* des deux problèmes. Cependant, la mise en garde de Ryle contre le cartésianisme et contre l'application abusive des prédicats mentaux, et sa volonté de se concentrer sur ce qui est observable, tendent à affirmer sa préférence (au moins sur le plan méthodologique) pour la réduction béhaviouriste. Il ne s'agit plus de réduire les termes de l'IA forte à ceux de l'IA faible, mais de considérer que l'IA forte n'a pas de sens lorsqu'on s'interroge sur les possibilités de l'IA faible.

Bilan sur le béhaviourisme logique

- Le béhaviourisme logique de Hempel, appliqué au test de Turing, réduit le problème de l'IA forte au problème de l'IA faible (« IA faible → IA forte »)[64]. La « véritable intelligence » peut être exprimée en termes de « comportements intelligents » et détectée empiriquement par le test de Turing.

- Le béhaviourisme logique de Ryle postule pour une indépendance logique des deux problèmes. Les problématiques de l'IA forte n'ont pas de sens du point de vu de l'IA faible, et réciproquement. Il s'agit d'une position plus radicale que celles présentées dans le chapitre précédent, puisque l'indépendance est formulée au niveau logique, et non au niveau épistémologique (section 2.1.2) ou au niveau disciplinaire (section 2.1.1).

- Dans les deux cas, le béhaviourisme logique rend impossible toute forme de collaboration. Soit parce que la philosophie n'est pas apte à résoudre le problème de l'IA forte, soit parce que celui-ci n'a pas de sens pour l'Intelligence Artificielle.

---

[63] *Ibid.*, p. 90.

[64] Nous avons vu que la relation véritablement défendue par le béhaviourisme logique est en fait « IA faible ↔ IA forte ». Celle-ci comprend la relation « IA faible → IA forte » utilisée pour la réduction. *Cf.* note 61.

Section 2.2.2. L'impossibilité *par principe* de l'IA faible

La relation « IA faible → IA forte », utilisée comme réduction *logique* par le béhaviourisme (relation analytique), peut être défendue dans le cadre d'une stratégie *épistémologique* (relation synthétique[65], *cf.* section 1.4). En effet, il est possible de défendre que la « véritable intelligence » est nécessaire à la simulation de l'intelligence. Par exemple, on peut défendre que la conscience est une qualité nécessaire à la production de comportements intelligents. Une telle relation est dangereuse pour l'IA faible lorsqu'on considère sa contraposée (« non IA forte → non IA faible »). En effet, si on souscrit à cette relation et que l'on dispose d'une preuve de l'absence de « véritable intelligence » chez les machines (« non IA forte »), alors l'IA faible est définitivement mise en échec (« non IA faible »). Par exemple, Searle soutient que les puces de siliciums qui constituent les machines numériques n'ont pas le pouvoir causal nécessaire à l'émergence de l'intentionnalité, et donc d'une « véritable intelligence »[66]. Dès lors, si on considère que l'intentionnalité est nécessaire à la production de comportements intelligents, on en déduit que le projet de l'IA faible est impossible. Il s'agit d'une impossibilité *par principe*, au sens où un principe externe (issu de l'IA forte) interdit à l'IA faible d'aboutir en pratique[67].

S'il advient que l'IA faible produise un contre-exemple *en pratique*, c'est-à-dire une machine qui simule effectivement l'intelligence, deux conclusions sont possibles. La première consiste à remettre en cause l'implication épistémologique « non IA forte → non IA faible ». Autrement dit, on découvre que la « véritable intelligence » n'est en fait pas nécessaire à la simulation des comportements intelligents. Ceci conduit à l'indépendance des deux problèmes, comme pour le béhaviourisme méthodologique de Turing (*cf.* section 2.1.2). La seconde conclusion possible consiste à remettre en cause l'affirmation « non IA forte ». De manière synthétique, on peut découvrir que les machines ont en fait accès à la « véritable intelligence ». Le *principe* qui contraignait l'IA faible est ainsi relâché. Les philosophes apprennent alors d'un tel contre-exemple. En effet, s'ils ont de bonnes raisons de croire que la relation épistémologique « IA faible → IA forte » est correcte, alors les résultats de l'IA faible peuvent amener à redéfinir les caractérisations de la « véritable intelligence ». Par exemple, si on affirme que l'intentionnalité est nécessaire aux comportements intelligents et si

---

[65] Au sens où « IA faible → IA forte » n'est pas « vraie en vertu de sa seule forme », mais relativement à une découverte épistémique. Pour parler en termes kantiens, il y a « accroissement de la connaissance ».

[66] Searle, J.R. 1980. *Op. cit*. *Cf.* également la section 2.3.2 à ce sujet.

[67] Searle ne soutient pas la relation épistémologique « non IA forte → non IA faible », en témoigne son expérience dite de la « chambre chinoise » (*cf.* section 2.3.2). Ici, seule sa thèse concernant l'absence de conscience chez les machines est utilisée (« non IA forte »). Elle offre un bon exemple de principe d'impossibilité.

un contre-exemple est mis en évidence par l'IA faible, il faut remettre en cause l'affirmation de Searle concernant le pouvoir causal des puces de silicium et préciser un concept d'intentionnalité qui s'applique également aux machines numériques.

La version synthétique de la relation « IA faible → IA forte » permet donc une collaboration entre Intelligence Artificielle et philosophie de l'esprit. L'« impossibilité *par principe* de l'IA faible » consiste à donner des causes philosophiques à l'impossibilité de produire des comportements intelligents (« non IA forte → non IA faible »). Cette démarche est également utilisée par Dreyfus pour expliquer les difficultés pratiques du computationnalisme à partir d'une critique de son socle philosophique (*cf.* section 2.4.2). Cependant, cette même relation peut être exploitée en inversant le sens de la collaboration. En effet, les résultats de l'Intelligence Artificielle peuvent renseigner la philosophie de l'esprit sur les formes d'intelligence possibles. Les machines de l'IA faible deviennent alors des objets d'étude pour les philosophes. Cette stratégie collaborative est explicitée ci-dessous à partir du travail d'Hector J. Levesque.

Section 2.2.3. La nécessité *en pratique* de l'IA forte

Une seconde façon d'envisager la relation synthétique « IA faible → IA forte » repose sur des considérations pratiques. Il ne s'agit pas d'utiliser l'IA forte comme un principe de l'impossibilité de l'IA faible, mais de considérer que l'IA faible nécessite *en pratique* une « véritable intelligence ». Hector J. Levesque, spécialiste en *knowledge representation and reasoning* depuis le début des années 80, défend ainsi que la réalisation d'une tâche complexe ne peut être menée à bien sans qu'une compréhension profonde n'ait elle-même lieu[68]. Par exemple, un programme sachant additionner 20 nombres de 10 chiffres chacun ne pourrait le faire sans « apprendre intelligemment » la notion d'addition. La solution « stupide » consistant à stocker en mémoire toutes les réponses possibles (soit 10^{200} réponses), pour que le programme puisse répondre « par cœur » au problème, est irréalisable. Une solution « raisonnable » consiste à implémenter un algorithme d'addition séquentielle avec retenues et donc à apprendre au programme à additionner intelligemment. On dira alors que cet algorithme « raisonnable » *comprend véritablement* la notion d'addition dans la mesure où il utilise des propriétés essentielles de l'opération mathématique. Pour Levesque, ce qui est vrai pour l'addition l'est aussi pour l'apprentissage d'une langue ou pour toute autre tâche complexe.

Ce qu'invoque Levesque, ce sont donc des raisons pratiques : il est impossible de construire une machine qui passe le test de Turing en établissant une simple liste des questions-réponses

[68] Levesque, H.J. 2009. « Is It Enough to Get the Behavior Right? » *International Joint Conference on Artificial Intelligence (IJCAI'09)*, p. 1439-1444.

possibles. L'explosion combinatoire du nombre de cas possibles interdit *en pratique* une telle approche. Bien qu'un tel programme existe *en théorie*, il ne peut être réalisé[69]. Levesque soutient alors qu'un programme, s'il passe effectivement le test de Turing, s'adapte nécessairement à la complexité du test à l'aide de boucles récursives. Celles-ci permettent de traiter aisément tous les cas possibles. Pour Levesque, ces structures récursives sont justement la marque d'une véritable assimilation de la tâche à réaliser, c'est-à-dire d'un raisonnement « véritablement intelligent » (« IA faible → IA forte »). Au minimum, un programme qui réussit à simuler l'intelligence n'est pas « complétement stupide » et mérite qu'on se penche sur son fonctionnement (« IA faible → IA forte probable »).

Dans ce cas, la relation « IA faible → IA forte » est une constatation empirique : en pratique, les comportements intelligents ne peuvent être simulés qu'en passant par l'implémentation d'une « véritable intelligence ». Deux conclusions sont à retenir. Pour les spécialistes de l'IA, il s'agit de ne pas choisir les modes d'implémentation simplistes tels que l'enregistrement complet de tous les résultats possibles, sous peine de ne jamais parvenir en pratique à l'émergence de comportements intelligents. Comme dans la section précédente, l'IA faible produit des objets d'étude intéressants pour les philosophes et les cogniticiens. Si un programme arrive *en pratique* à résoudre une tâche complexe, sans doute que la manière dont il s'exécute constitue un processus cognitif intéressant. Peut-être même qu'il peut servir à décrire et à expliquer la nature de la cognition. Par ailleurs, la réciproque n'est pas nécessairement vraie : le fait qu'on n'ait pas encore réalisé de programmes passant le test de Turing ne signifie pas que ceux que nous avons réalisés jusqu'à présent sont inintéressants.

Cette stratégie paraît féconde pour mettre l'Intelligence Artificielle au service de la philosophie de l'esprit. En outre, elle réalise le renversement proposé par Andler en introduction consistant à faire de l'Intelligence Artificielle une « science de la philosophie ». Cette démarche est développée dans le chapitre 4.2 à partir du travail de Timothy van Gelder sur les régulateurs de

[69] Levesque constate que le nombre de bits d'information nécessaires à l'implémentation d'un tel programme est largement supérieur au nombre d'atomes dans l'univers. Dans son article, la difficulté pratique devient donc une *stricte impossibilité*. Nous retenons essentiellement la première idée, en remarquant que d'autres programmes un peu moins « stupides » et un peu moins « coûteux », mais argumentant dans le même sens, pourraient *en principe* être implémentés malgré de très lourdes difficultés *en pratique*.

vitesses pour les machines à vapeur[70]. Ces objets techniques, répondant à un problème pratique délicat, servent de modèles à la philosophie de la cognition pour préciser ses théories (« IA faible → IA forte probable »).

Section 2.2.4. Critique de la relation « IA faible → IA forte »

La critique des approches présentées dans cette section repose sur l'existence de contre-exemples « IA faible (x) et non IA forte (x) ». Le contre-exemple le plus simple étant celui que nous avons nommé « zombie[71] chanceux » dans notre mémoire de Master 1[72]. Il s'agit d'un programme qui produit des comportements de manière complétement aléatoire. Dans le cas du test de Turing, par exemple, le « zombie chanceux » engendre des suites de symboles en choisissant aléatoirement parmi les 26 caractères de l'alphabet latin, plus quelques signes de ponctuations. Ces « phrases » sont ensuite transmises à l'interlocuteur humain. Il existe une chance, certes infime, que de telles suites de symboles aient un sens, voire une certaine pertinence, et qu'elles trompent ainsi le jugement de l'observateur. À l'issue du test, celui-ci conclura que le zombie a bien un comportement intelligent. Or, il paraît invraisemblable d'affirmer que ce programme possède une « véritable intelligence ». En vérité, il s'agit du programme le plus « stupide » qui puisse exister. On dispose donc d'un cas de comportement intelligent engendré par un programme non-intelligent (« IA faible (zc) et non IA forte (zc) »). La relation « IA faible → IA forte » est alors réfutée.

- Sur le plan analytique, c'est le fait d'intégrer des considérations concernant le fonctionnement interne des machines qui rend caduque la réduction du béhaviourisme logique. Ici, on connaît la manière dont le « zombie chanceux » produit ses comportements. Or, il est difficile d'affirmer que son fonctionnement aléatoire est *analytiquement* équivalent à ses comportements intelligents. En effet, plusieurs fonctionnements différents peuvent produire les mêmes comportements et les formules de contrôles physiques, telles que le test

---

[70] van Gelder, T. 1998. « Dynamique et cognition. » [« Dynamics and Cognition. »] Lapointe, S. (trad.). *In* Fisette, D. (éd.), Poirier, P. (éd.). 2003. *Philosophie de l'esprit : problèmes et perspectives*. Paris : Vrin, p. 329-369.

[71] Le terme « zombie » est utilisé en référence à Kirk, R., Squires, J.E. 1974. « Zombies v. Materialists. » *The Aristotelian Society*, vol. 48, p. 135-163. Il désigne un individu qui se comporte en tout point comme un homme, mais à qui il manque une propriété essentielle, telle que les *qualia*. Dans notre cas, un zombie se comporte *comme s'*il était intelligent (« IA faible »), mais il n'est pas *véritablement* intelligent (« non IA forte »). À ce titre, la chambre chinoise de Searle est, elle aussi, un zombie (*cf.* section 2.3.2 et section 4.3.2).

[72] Lamarche-Perrin, R. 2010. *Le Test de Turing pour évaluer les théories de l'esprit*. Mémoire de Master, Kistler, M. (dir.). Grenoble : Université Pierre-Mendès-France, sept. 2010.

de Turing, ne suffisent pas à les identifier. De manière générale, le béhaviourisme est mis à mal lorsqu'on laisse de côté les qualités fondamentalement inobservables (telle que la conscience) pour s'intéresser à des qualités internes accessibles à l'examen empirique. Les sciences cognitives ont ainsi mis fin au béhaviourisme en s'intéressant aux cerveaux et à leurs processus cognitifs.

- Sur le plan synthétique, le « zombie chanceux » réfute toute impossibilité *par principe* de l'IA faible (section 2.2.2). En effet, il existe au moins un programme qui, dans de bonnes circonstances, agit comme s'il était intelligent. Nous préférons alors des relations moins radicales telles que « non IA forte → IA faible peu probable ». Ainsi, même si le « zombie chanceux » peut se comporter intelligemment, cela constitue un cas exceptionnel et, en général, nous avons bien « non IA forte → non IA faible ».

- En ce qui concerne la nécessité *en pratique* de l'IA forte, cependant, l'argument du « zombie chanceux » ne suffit pas. En effet, sur une assez longue durée, un tel programme ne parviendra jamais à se comporter convenablement. En pratique donc, il ne faut pas considérer que le fonctionnement aléatoire est une forme de « véritable intelligence ». Il est possible, ici aussi, de soutenir une relation moins radicale de la forme « IA faible → IA forte ».

Chapitre 2.3. « IA faible ↔ IA forte »

Section 2.3.1. Le computationnalisme de Newell & Simon

Dans les années 1960, les sciences cognitives se sont organisées autour d'une critique véhémente du béhaviourisme. De nombreux chercheurs ont alors soutenu que les sciences modernes avaient leur mot à dire quant au fonctionnement interne du cerveau et que la seule analyse des comportements observables ne suffisait pas à expliquer la complexité de l'esprit humain. L'avènement des sciences cognitives a ainsi permis de dépasser la position béhaviouriste faisant de la psychologie une science des comportements (*cf.* section 2.1.2 et section 2.2.1). Dans le cas de l'Intelligence Artificielle, il devient évident que la production des comportements intelligents (IA faible) et les facultés cognitives internes aux machines (IA forte) doivent, elles aussi, être mises en relation. Cela induit de nouvelles façons de considérer les rapports possibles entre IA faible et IA forte.

La double hypothèse des systèmes symboliques physiques

Parmi les hypothèses fondatrices des sciences cognitives, le *computationnalisme* hérite directement des avancées de l'Intelligence Artificielle. Dans cette section, nous en analysons une acception technique proposée en 1976 par Allen Newell et Herbert A. Simon, deux spécialistes de l'IA. Le cœur de la mise en relation entre comportements intelligents et facultés cognitives réside alors dans leur *hypothèse des systèmes symboliques physiques* qui s'énonce ainsi : « un système symbolique physique a les moyens nécessaires et suffisants de l'action intelligente générale. »[73] Pour le dire en quelques mots, un « système symbolique » est un ensemble de symboles et un ensemble de règles qui permettent de les manipuler. Un « système symbolique physique » ou « SSP » est la réalisation physique d'un tel système symbolique, c'est son implémentation matérielle via un automate mécanique, électronique ou même biologique. Les ordinateurs, à ce titre, sont des exemples canoniques de SSP.

L'hypothèse des SSP est une hypothèse double :

1. Par « suffisant », on entend que les ordinateurs, en tant que SSP, ont les moyens suffisants à l'action intelligente. Il s'agit de l'hypothèse de l'IA faible, ni plus ni moins : selon Newell &

[73] « The Physical Symbol System Hypothesis. A physical symbol system has the necessary and sufficient means for general intelligent action. » [Notre traduction] Newell, A., Simon, H.A. 1976. « Computer Science as Empirical Inquiry: Symbols and Search. » *Communications of the ACM*, vol. 19, n°3, p. 116.

Simon, les machines peuvent en principe agir intelligemment (si leur taille et leur organisation le leur permet). On a donc « IA faible (SSP) ».

2. Par « nécessaire », on entend que le cerveau humain, puisqu'il est responsable de la production de comportements intelligents, doit lui aussi être *une sorte de SSP*. Le cerveau est donc (en un sens à préciser) similaire à l'ordinateur. Il s'agit de l'hypothèse computato-représentationnelle, donc Jerry A. Fodor est sans doute un des plus grands défenseurs du côté de la philosophie[74].

Une analogie en faveur de l'IA forte

La condition nécessaire de l'hypothèse des SSP défend l'analogie suivante : l'esprit est au cerveau ce que le programme est à l'ordinateur qui l'exécute. L'esprit est donc un système symbolique ; ses symboles et ses règles sont réalisés par un réseau neuronal ; et la cognition consiste en un calcul sur ces symboles, c'est-à-dire une *computation*.

De ce fait, le computationnalisme penche en faveur de l'IA forte. En effet, du fait de l'analogie, une machine peut prétendre aux capacités que l'on attribue aux autres SSP, tels que les cerveaux humains. Ainsi, à l'instar de leurs analogues biologiques, les ordinateurs sont capables d'engendrer une vie consciente, des états mentaux, *etc.* Bien que Newell, Simon et les autres partisans du computationnalisme en Intelligence Artificielle ne prennent pas directement position (*cf.* section 2.1.1), une lecture plus engagée de la thèse computationnaliste tend à résoudre le problème de l'IA forte en ce qui concerne la conscience et les états mentaux. Par exemple, pour Stevan Harnad, le computationnalisme affirme que « les états mentaux sont juste l'implémentation des (bons) programmes informatiques. »[75] On a alors « IA forte (SSP) ». Dans tous les cas, le computationnalisme ne formule pas d'impossibilité de principe contre l'IA forte.

Ainsi, selon le computationnalisme :

- L'IA faible est possible *en principe*. Nous verrons qu'*en pratique* les applications du computationnalisme ont rencontré d'importantes difficultés (section 2.4.1).

---

[74] Fodor, J.A. 1975. *The Language of Thought*. Cambridge, MA : Harvard University Press.

[75] « Mental states are just implementations of (the right) computer program(s). » [Notre traduction] Harnad, S. 2001. « Minds, Machine and Seale II: What's Wrong and Right About Searle's Chinese Room Argument? » *In* Bishop, M. (éd.), Preston, J. (éd.). 2001. *Essays on Searle's Chinese Room Argument*. Oxford University Press.

- Il n'y a pas d'impossibilité *de principe* contre l'IA forte. Il n'y a aucune raison pour laquelle les ordinateurs ne pourraient pas, au même titre que les cerveaux, avoir des états mentaux, une vie consciente, *etc.*

L'équivalence des deux problèmes

Il est également possible de voir dans le computationnalisme, et plus particulièrement dans la double hypothèse de Newell & Simon, une preuve synthétique de l'équivalence des deux problèmes de l'Intelligence Artificielle. En effet, du fait de l'analogie computo-représentationnelle, la notion de SSP peut être interprétée comme la définition même de ce qu'est la « véritable intelligence ». *Être intelligent*, c'est donc *être un SSP bien organisé*. Si l'hypothèse de Newell & Simon se révèle juste, on a alors :

1. Condition suffisante : il suffit d'être un SSP bien organisé pour pouvoir se comporter intelligemment. On a donc « IA forte → IA faible ».

2. Condition nécessaire : les comportements intelligents sont nécessairement le produit d'un SSP bien organisé. On a donc « IA faible → IA forte ».

Ainsi, pour le computationnalisme les problèmes de l'IA forte et de l'IA faible sont épistémologiquement équivalents (« IA forte ↔ IA faible »). Construire et étudier les machines intelligentes revient à construire et étudier des SSP. Ce postulat épistémologique a largement participé à l'essor des sciences cognitives. De nombreuses disciplines se mettent alors à travailler sur des problématiques connexes, autour d'un nouvel objet commun : les systèmes symboliques et leurs implémentations physiques (naturelles ou artificielles). L'hypothèse computationnaliste permet ainsi une « collaboration maximale » entre philosophie de l'esprit et Intelligence Artificielle, mais également avec d'autres disciplines telles que la psychologie, la linguistique et la neurobiologie. Par ailleurs, ces relations scientifiques sont intra-théoriques, c'est-à-dire qu'elle participe à la création d'un champ de recherche commun, plutôt que d'instaurer les rapports asymétriques (*cf.* chapitre 1.1). Ici, l'une et l'autre des deux disciplines peuvent partager leurs connaissances propres et leurs méthodes particulières : l'IA utilise les résultats des sciences de l'esprit pour construire des programmes intelligents[76] et la philosophie étudie ces programmes pour comprendre la structure et le fonctionnement de l'esprit[77].

[76] Newell et Simon se basent par exemple sur une analyse psychologique des raisonnements humains pour construire leur *General Problem Solver* (*cf.* chapitre 2.4). Newell, A., Simon, H.A. 1963. « GPS, A Program

Notons que l'hypothèse des SSP n'est pas un théorème, mais une « généralisation empirique »[78]. Newell & Simon ne prétendent pas donner de preuve analytique de l'équivalence entre IA forte et IA faible. La démonstration de l'hypothèse repose donc sur un travail synthétique des sciences cognitives. La condition suffisante « IA forte → IA faible » (les SSP peuvent engendrer des comportements intelligents) est abordée par l'Intelligence Artificielle qui doit montrer que les machines sont effectivement capables de comportements intelligents. La condition nécessaire « IA faible → IA forte » (les comportements intelligents sont nécessairement produits par un SSP) est abordée par la psychologie cognitive qui doit notamment montrer que les cerveaux humains sont bien des SSP. L'hypothèse du computationnalisme est bien le cœur de la collaboration qui prend place au sein des sciences cognitives.

Section 2.3.2. La critique searlienne du computationnalisme

La « chambre chinoise »

Dans sa célèbre expérience de pensée de la « chambre chinoise », Searle avance un argument contre le computationnalisme[79]. Il imagine un dictionnaire, rédigé en anglais, contenant des règles formelles de manipulation de sinogrammes du type « *if someone asks you* "你怎么样?", you should *answer* "我很好，谢谢你". » Le dictionnaire ne précise pas le sens des expressions chinoises ainsi manipulées[80], si bien qu'une locutrice anglophone peut avoir une longue et profonde conversation en chinois sans jamais *comprendre* les termes de la discussion. Si celle-ci s'enferme dans une « chambre » avec le dictionnaire, et communique avec un locuteur chinois en faisant glisser des feuilles de papier sous la porte, le locuteur affirmera que la femme dans la « chambre chinoise » maîtrise sa langue à la perfection. Il en conclura immédiatement que celle-ci *comprend* le chinois au sens de Turing. Pourtant, la locutrice anglophone manipule des symboles qui sont pour elle vides de sens : difficile donc d'affirmer que celle-ci *comprend réellement* le chinois.

---

that Simulates Human Thought. » *In* Feigenbaum, E.A. (éd.), Feldman, J. (éd.). 1995. *Computers and Thought*. Cambridge, MA : MIT Press, p. 279-293.

[77] Simon prédit par exemple que « within ten years most theories in psychology will take the form of computer programs, or of qualitative statements about the characteristics of computer programs. » Newell, A., Simon, H.A. 1958. « Heuristic Problem Solving: The Next Advance in Operations Research. » *Operations Research*, vol. 6, n°1, p. 7-8.

[78] Newell, A., Simon, H.A. 1976. *Op. cit.*, p. 118.

[79] L'expérience est rapportée en détail dans Searle, J.R. 1980. *Op. cit.*

[80] Pour les lecteurs non sinophones, « 你怎么样? » signifie « *how are you?* » et « 我很好，谢谢你 » signifie « *I am fine, thank you* ».

La femme de la « chambre chinoise » passe avec succès une variante du test de Turing, consistant à identifier les locuteurs chinois, alors que celle-ci ne parle pas « véritablement » chinois. La « chambre chinoise » constitue donc, de la même manière que le « zombie chanceux » présenté dans la section 2.2.4, un cas de *faux positif* pour le test de Turing, c'est-à-dire un individu passant le test avec succès alors que celui-ci n'est pas « véritablement intelligent »[81]. On a donc « IA faible (cc) et non IA forte (cc) » où « cc » désigne le dispositif imaginé par Searle. La relation « IA faible → IA forte » est alors remise en cause par ce contre-exemple.

Un argument contre le béhaviourisme

L'argument de Searle peut donc être interprété comme une critique du béhaviourisme méthodologique (la définition comportementale n'est pas suffisante : dans le cas de la « chambre chinoise » il faut également examiner le fonctionnement interne pour identifier l'intelligence du dispositif, *cf.* section 2.1.2), une critique du béhaviourisme logique (le comportement intelligent ne suffit pas à définir analytiquement la notion de « véritable intelligence », *cf.* section 2.2.1) et une critique des autres défenseurs de la relation « IA faible → IA forte » (*cf.* reste du chapitre 2.2). Seulement, il s'agit d'un argument complexe, alors que l'exemple minimaliste du « zombie chanceux » suffit à mettre en cause la relation « IA faible → IA forte » (*cf.* section 2.2.4). En effet, la « chambre chinoise » a d'autres objectifs que la seule critique du béhaviourisme.

Un argument contre le computationnalisme

L'objectif principal de la « chambre chinoise » est une critique du computationnalisme. Le dictionnaire décrit par Searle est en effet un système symbolique, dont les symboles sont des sinogrammes, et qui est manipulé (c'est-à-dire exécuté physiquement) par la locutrice anglophone. La « chambre chinoise » est donc un SSP. En affirmant « IA faible(SSP) et non IA forte (SSP) », Searle remet en cause le fait que les SSP sont « véritablement intelligents ». Ainsi, les cerveaux, puisqu'ils sont « véritablement intelligents », ne peuvent pas être eux-mêmes des SSP. C'est la fin de l'analogie computo-représentationnelle. Même si l'on parvient à résoudre le problème pratique de la simulation de l'intelligence grâce aux SSP (« IA faible (SSP) »), il reste des questions philosophiques que les spécialistes de l'IA ne peuvent régler, et notamment la question de la « véritable intelligence » (« non IA forte (SSP) »). Notons que Searle ne critique pas la *condition suffisante* de l'hypothèse des SSP de Newell & Simon. La « chambre chinoise » est elle-même un SSP qui parvient à simuler l'intelligence. Seule la *condition nécessaire* est mise en cause, dans la mesure où les

[81] Bien sûr, le contre-exemple de Searle ne se limite pas à la maîtrise du chinois. Il peut être généralisé à toute faculté de l'esprit et notamment à la notion d'intelligence.

cerveaux, qui ne sont pas des SSP, parviennent également à produire des comportements intelligents.

La véritable cible de la « chambre chinoise » est donc l'assimilation de la « véritable intelligence » au fait d'être « un SSP bien organisé ». Ainsi, pour Searle, la cognition n'est pas seulement syntaxique. Elle n'est pas une simple *computation*. L'hypothèse du computationnalisme est donc en partie fausse (condition nécessaire rejetée), et la philosophie perd le bénéfice des travaux en Intelligence Artificielle dans la mesure où l'objet d'étude varie par ses qualités fondamentales. Les cerveaux humains, contrairement aux SSP, sont notamment doués d'intentionnalité[82], ils peuvent engendrer la conscience, une vie phénoménale[83], *etc.* Le projet de collaboration au sein des sciences cognitives est ainsi avorté.

Cependant, lorsqu'on examine en détail l'argumentation de Searle, on se rend compte que celle-ci est mal dirigée. En effet, pour mettre à mal la théorie computo-représentationnelle, Searle a besoin d'affirmer « non IA forte (SSP) ». La seconde affirmation « IA faible (SSP) » n'est pas utilisée. En effet, nous n'avons pas besoin de savoir que la « chambre chinoise » parvient à simuler l'intelligence pour démontrer qu'elle n'est pas « véritablement intelligente ». Cette seconde affirmation n'est nécessaire que si l'on cible le béhaviourisme et si l'on veut montrer « non (IA faible → IA forte) ». En d'autres termes, Searle développe un argument trop encombrant pour sa seule cible : l'IA forte[84].

Un argument contre l'IA forte

En définitive, Searle émet surtout une objection radicale contre l'IA forte. Comme brièvement présenté dans la section 2.2.2, Searle défend l'idée que seuls les cerveaux ont le pouvoir causal nécessaire à la conscience et à l'intentionnalité. Au contraire, les puces de silicium au sein des machines numériques n'ont pas cette capacité et ne peuvent *de facto* pas être à l'origine d'une véritable compréhension des symboles manipulés, pas plus qu'ils ne sont capables d'avoir des états intentionnels[85]. En d'autres termes, les machines ne peuvent dépasser le niveau syntaxique. De nombreuses critiques de cette objection ont argumenté que, s'il paraissait incroyable qu'un amas d'objets inanimés – tels que des transistors – puissent faire émerger une vie consciente, il était tout

[82] Searle, J.R. 1980. *Op. cit.* ; Searle, J.R. 1984. *Op. cit.*

[83] Searle, J.R. 1992. *Op. cit.*

[84] Nous reviendrons sur les erreurs de l'argumentation de Searle dans la section 4.3.2.

[85] « Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena. » Searle, J.R. 1980. *Op. cit.*, p. 86-87.

aussi incroyable de constater qu'un amas de neurones en soit capable[86]. Ainsi, l'objection searlienne résulte plus d'une *pétition de principe* induite par notre incompréhension profonde du mystère de la conscience, que d'un véritable argument scientifique.

Searle nie donc la possibilité de l'IA forte : les machines *par principe* ne peuvent pas engendrer cette « véritable intelligence », au niveau sémantique, que les cerveaux humains atteignent en produisant des états intentionnels. Si Searle est optimiste quant au travail des sciences du cerveau pour résoudre le problème difficile de la conscience humaine, et donc pour fonder de nouvelles « sciences de l'esprit », il accuse l'inutilité fondamentale de l'Intelligence Artificielle dans ce domaine[87]. Cependant, en argumentant pour cette position de principe, Searle ne s'intéresse qu'au problème de l'IA forte. Les possibilités et difficultés pratiques de l'IA faible sont laissées pour compte par son analyse (*cf.* section 2.1.1) et les spécialistes de l'IA ne peuvent donc pas profiter de ses arguments philosophiques. L'objection de principe « non IA forte (SSP) » informe seulement la philosophie de l'esprit. Elle n'a aucune conséquence pratique pour l'Intelligence Artificielle. Dans le chapitre suivant, nous montrons comment la critique de Dreyfus, au contraire, ne se limite pas au domaine de l'IA forte.

Bilan de la critique searlienne

La critique de Searle souffre de nombreux défauts en ce qui concerne la problématique de ce mémoire :

- L'argument de la « chambre chinoise », pris dans son intégralité, s'oppose avant tout au béhaviourisme et à l'affirmation « IA faible → IA forte ». En effet, la « chambre chinoise » est un dispositif qui simule l'intelligence sans comprendre un traître mot de ce qu'il dit (« IA faible (cc) et non IA forte (cc) »). Si on en reste là, l'argumentation n'est pas nouvelle et les critiques du béhaviourisme, notamment par les sciences cognitives, ont largement précédées celle de Searle.

---

[86] « Comment un amas compliqué d'événements de traitement d'information dans un tas de puces de silicone *pourrait*-il être équivalent à des expériences conscientes ? Mais il est tout aussi difficile d'imaginer la façon dont un cerveau humain organique pourrait servir de support à la conscience. […] Et pourtant nous imaginons aisément que des êtres humains soient conscients, même si nous ne pouvons toujours pas imaginer *comment* c'est possible. » Dennett, D.C. 1991. *Consciousness Explained*. New York : Hachette Book, p. 537. Voir également la position de Chalmers dans Chalmers, D. 1996. *Op. cit.*, p. 314.

[87] « J'ai bon espoir que les sciences du cerveau accroîtront la compréhension que nous avons de nous-même. Je suis pessimiste sur l'avenir de l'intelligence artificielle. » Searle, J.R. 1984. *Op. cit.*, p. 3.

- La nouveauté de l'argumentation réside donc dans l'opposition frontale au computationnalisme, et plus généralement au projet de l'IA forte. L'impossibilité des machines numériques à engendrer une conscience et des états intentionnels met ainsi un terme à la théorie computo-représentationnelle et à la « collaboration maximale » au sein des sciences cognitives. Cependant, cette argument relève plus d'une pétition de principe que d'un argument empirique.

- Au-delà de cette limite scientifique, l'argument contre l'IA forte n'a aucune conséquence quant à la résolution du problème de l'IA faible. Searle interdit au spécialiste de l'IA de s'exprimer sur la question de la « véritable intelligence » et il n'incite par les philosophes à s'intéresser au problème pratique de la simulation de l'intelligence. Dans ce contexte, aucune collaboration n'est prévue entre Intelligence Artificielle et philosophie de l'esprit.

- Enfin, Searle est parfois mal renseigné sur les recherches en Intelligence Artificielle qui lui sont contemporaines. Il a notamment été soutenu que l'argument de la « chambre chinoise », à cause de sa méconnaissance des programmes réellement développés par les spécialistes de l'IA, ne pouvait être généralisé à *tout type* de computation et que, par conséquent, la critique de Searle n'atteignait qu'une partie limitée du computationnalisme. En particulier, le dictionnaire de la « chambre chinoise » n'a pas été réellement rédigé et il est possible que Searle sous-estime la complexité d'un tel programme[88].

[88] Pour Dennett, l'argument de la « chambre chinoise » repose sur « the (unwarranted) supposition that the giant program would work by somehow simply "matching up" the input Chinese characters with some output Chinese characters. » Cette vision très restrictive des programmes computationnalistes ne reflète pas la complexité des programmes développés à l'époque. Elle se limite éventuellement aux modèles de raisonnements séquentiels et linéaires, développés notamment dans le cas des « systèmes experts » (*cf.* section 2.4.1). Les autres exemples présentés dans l'article de Searle, plus sophistiqués, limitent cependant la clarté de son argument. Pour Dennett justement, « complexity does matter. » Dennett, D.C. 1991. *Op. cit.*, p. 431-455.

Chapitre 2.4. Critique dreyfusienne : un exemple de véritable collaboration

Ce chapitre reprend les trois parties de l'ouvrage de Hubert L. Dreyfus intitulé *What Computers Can't Do*[89]. Il rend explicite les stratégies collaboratives qui nous intéressent pour la suite.

Section 2.4.1. Partie 1 : l'échec constaté du computationnalisme

Contrairement au travail de Searle, la critique de Dreyfus est très bien renseignée sur les avancées et les résultats de l'Intelligence Artificielle. En témoigne la première partie de *What Computers Can't Do* dans laquelle Dreyfus fait le bilan de vingt années de recherche[90], de 1957 à 1977. Parmi les projets de l'Intelligence Artificielle dont Dreyfus fait l'analyse, le *General Problem Solver* (GPS) de Newell, Shaw & Simon constitue un exemple canonique d'application du computationnalisme[91].

Objectifs des programmes computationnalistes

Le GPS est construit pour résoudre tout type de problèmes rationnels qui requièrent « intelligence et adaptation ». En ce sens, il participe à l'effort de l'IA faible. Mais au-delà de généraliser et d'automatiser des méthodes de résolution de problèmes, le projet auquel appartient le GPS a pour objectif premier de « comprendre les processus d'information à l'origine des capacités intellectuelles, adaptatives et créatives de l'homme. »[92] Le GPS utilise notamment des méthodes de résolutions implémentées à partir de celles qu'utilisent quotidiennement les étudiants d'une université américaine[93]. Ainsi, les auteurs affirment non seulement que leur programme est capable de résoudre tout type de problèmes rationnels, mais qu'en plus il procède pour faire cela de manière similaire à l'homme. Newell, Shaw et Simon ont ainsi l'ambition de fonder une « théorie de la

---

[89] Dreyfus, H.L. 1979. *Intelligence Artificielle : mythes et limites*. [*What Computers Can't Do: The Limits of Artificial Intelligence*, 2nd ed.] Vassallo-Villaneau, R.-M. (trad.), Andler, D. (pref.), Perriault, J. (pref.). Paris : Flammarion, 1984.

[90] Dreyfus, H.L. 1979. « Première partie : vingt ans de recherche en intelligence artificielle (1957-1977). » *Op. cit.*, p. 35-188.

[91] Newell, A., Shaw, J.C., Simon, H.A. 1959. « Report on a general problem-solving program. » *International Conference on Information Processing*, p. 256-264 ; Newell, A., Simon, H.A. 1963. *Op. cit*.

[92] « to understand the information processes that underlie human intellectual, adaptive, and creative abilities. » [Notre traduction] Newell, A., Shaw, J.C., Simon, H.A. 1959. *Op. cit.*, p. II.

[93] Newell, A., Simon, H.A. 1963. *Op. cit.*, p. 280-282.

résolution humaine des problèmes »[94]. Ainsi, ils travaillent simultanément sur la production informatique de comportements intelligents et sur l'intelligence humaine.

On retrouve bien, à l'origine de ce projet ambitieux, la « collaboration maximale » des sciences cognitives, liant Intelligence Artificielle et sciences de l'esprit (ici notamment la psychologie cognitive). Cette collaboration est rendue possible par la double hypothèse des systèmes symboliques physiques (SSP), qui sera proprement énoncée quelques années plus tard par Newell & Simon, et par l'analogie computationnaliste entre ordinateurs et cerveaux (*cf.* section 2.3.1). Le GPS est donc développé en accord avec les crédos computationnalistes, et en particulier en accord avec l'équivalence des deux problèmes (« IA faible ↔ IA forte »). Dans ce projet, en proposant des modèles descriptifs des opérations cognitives humaines, l'Intelligence Artificielle se met au service de la psychologie.

Architecture des programmes computationnalistes

Si on regarde plus en détail l'histoire du GPS, on trouve en amont le *Logic Theorist*[95] également développé par Newell & Simon. Ce précurseur implémentait des méthodes de raisonnement capables de démontrer quelques théorèmes mathématiques issus des *Principia Mathematica* de Russell & Whitehead. En aval, le GPS est un proche ancêtre des « systèmes experts »[96], ces outils d'aide à la décision développés à partir de la fin des années 60 pour raisonner sur de grandes bases de connaissance, par exemple pour identifier des constituants chimiques (DENDRAL[97]) ou pour le diagnostic de maladie du sang et la prescription de médicaments (MYCIN[98]). Cette lignée de programmes, du *Logic Theorist* aux systèmes experts, est identifiable par son architecture logicielle commune. En effet, ces programmes partagent une conception et un fonctionnement largement similaires, reposant sur l'implémentation de systèmes symboliques et l'utilisation de principes d'inférences logiques[99]. En outre, on peut décrire ainsi leur fonctionnement : (1) les données du problème à résoudre sont formalisées dans une « base de faits » qui constitue le

---

[94] « a theory of human problem-solving » *Ibid.*, p. 279.

[95] Newell, A., Shaw, J.C., Simon, H.A. 1959. *Op. cit.*, p. 2 ; Russell, S.J., Norvig, P. 2003. *Op. cit.*, p. 17-18.

[96] Russell, S.J., Norvig, P. 2003. « Part III: Knowledge-based systems: The key to power? (1969-1979). » *Op. cit.*, p. 22-24.

[97] Buchanan, B.G., Sutherland, G.L., Feigenbaum, E.A. 1969. « Heuristic DENDRAL: a Program for Generating Explanatory Hypotheses in Organic Chemistry. » *In* Meltzer, B. (éd.), Michie, D. (éd.). 1969. *Machine Intelligence*, vol. 4. Edinburgh University Press, p. 209-254.

[98] Buchanan, B.G., Shortliffe, E.H. 1984. *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA : Addison-Wesley.

[99] Russell, S.J., Norvig, P. 2003. « Chapter 9: Inference in First-Order Logic. » *Op. cit.*, p. 322-365.

vocabulaire du système symbolique. (2) Des règles logiques opérant sur ce vocabulaire sont également formalisées. Elles constituent la « base de règles » et fixent la syntaxe du système. (3) Des algorithmes permettent d'appliquer les règles aux données du problème afin d'engendrer de nouvelles connaissances, jusqu'à la résolution du problème. Ces algorithmes sont des fonctions heuristiques destinées à faire évoluer le système symbolique, afin d'obtenir une solution. Par exemple, à partir de la position actuelle des pièces sur l'échiquier (« base de faits »), des règles de déplacement (« base de règles ») et de la solution recherchée (mettre « échec et mat » son adversaire), un programme comme le GPS peut utiliser ses heuristiques pour proposer un mouvement régulier et stratégique.

L'intérêt fondamental de ces architectures réside dans la généricité du modèle de raisonnement : les heuristiques sont ainsi définies indépendamment des « bases de faits » et des « bases de règles » ; elles utilisent notamment un « moteur d'inférence », simulant des raisonnements déductifs logiques et dérivant des conclusions à partir des connaissances. Ces algorithmes, en principe, sont donc applicables à tout système symbolique. Newell, Shaw et Simon supposent alors que tout problème rationnel peut être formalisé à l'aide d'un système symbolique (vocabulaire et syntaxe) et que les algorithmes heuristiques, en utilisant à dessein les règles d'inférence logique, peuvent y apporter une solution. Il s'agit bien d'une application directe du computationnalisme, faisant du cerveau un SSP et de la cognition un calcul symbolique. L'hypothèse supplémentaire défendue par le GPS consiste à affirmer que la méthode de calcul est principalement indépendante du problème traité. Il s'agit d'une hypothèse forte et non triviale, dont la validité sera mise en cause par les opposants au computationnalisme.

Difficultés des programmes computationnalistes

Dreyfus fait l'inventaire des difficultés rencontrées par le GPS. Nous en distinguerons trois catégories. La première est propre à la complexité des problèmes traités. Les opérations logiques possibles et les nouveaux faits, engendrés à partir du vocabulaire de base, tendent à se multiplier très rapidement. Les ressources nécessaires au traitement de la base augmentent alors de manière exponentielle. On parle d'« explosion combinatoire ». Pour pallier ce problème, il faut déployer des stratégies efficaces de sélection des règles et des faits pertinents, sous peine de se retrouver très rapidement à manipuler un système symbolique gigantesque. Dans le cas du joueur d'échec, la décision du prochain coup ne peut reposer sur une évaluation complète de l'arbre des parties possibles[100]. Il faut alors définir des stratégies moins systématiques, plus « intelligentes », pour

[100] Shannon a estimé à 10^{120} le nombre de parties possibles aux échecs. Il s'agit du nombre de parties « raisonnables », c'est-à-dire largement inférieur au nombre de parties légales, incluant une grande quantité de

sélectionner l'ensemble des coups pertinents[101]. Seulement, la sélection des faits pertinents pour la résolution du problème dépend énormément du problème en question. Le computationnalisme se heurte alors à deux types de difficultés bien plus graves.

Ces difficultés sont liées aux limites des « micromondes »[102]. Un « micromonde » est un univers simpliste, qui modélise éventuellement un problème réel complexe. Le nombre des variables d'état et leurs évolutions possibles sont grandement limités, notamment pour pallier les difficultés liées à l'« explosion combinatoire ». Les « micromondes » sont ainsi des univers clos, discrets et certains : on dispose d'un nombre borné d'objets, dont les caractéristiques sont fixées et entièrement connues. Par exemple, la position des pièces sur un échiquier, les différents coups possibles et la condition de victoire peuvent être entièrement formalisés et connus avec certitude, ce qui fait du jeu d'échecs un « micromonde » idéal. Mises à part les difficultés propres à un « micromonde » donné, l'utilisation de ces modèles simplistes induit deux difficultés de taille :

1. Tout d'abord, il est difficile d'imbriquer ou de généraliser les « micromondes » pour augmenter la portée d'un programme. Celui-ci détient alors des savoirs spécialisés (jouer aux échecs, prouver un théorème, diagnostiquer une maladie, *etc.*) dont les liens ne sont pas mis en évidence par la technique des « micromondes ». Pourtant, selon Dreyfus, il est impossible de décomposer notre monde quotidien en une telle myriade de problèmes simples et isolés. Le « monde réel » n'est pas un assemblage de « micromondes », il est impossible de le modéliser à partir de ces briques « même en les étirant et en les combinant entre [elles] »[103]. De ce fait, le GPS, mille fois spécialiste, n'atteint pas la généricité escomptée.

2. De plus, certains problèmes sont très difficiles à formaliser en termes de faits et de règles, notamment ceux dont l'espace des états possibles n'est pas autant contrôlé que celui du jeu d'échecs. L'Intelligence Artificielle se heurte à une grande difficulté concernant la

---

parties « absurdes ». Shannon, C. 1950. « Programming a Computer for Playing Chess. » *Philosophical Magazine*, vol. 41, n°314, p. 256-275.

[101] Par exemple, l'« équilibre des forces » et le « contrôle du centre » sont deux critères classiques permettant d'évaluer et de comparer les états possibles d'un échiquier. Dreyfus, H.L. 1979. *Op. cit.*, p. 149. Cependant, de nombreux autres coups sont éliminés par des critères bien plus simples, implicitement mis en place par les habitudes et le « bon sens » des joueurs initiés.

[102] Dreyfus, H.L. 1979. « Chapitre 3 : phase III (1967-1972) : la manipulation des micromondes. » *Op. cit.*, p. 115-144.

[103] *Ibid*., p. 126-127.

représentation les « connaissances usuelles »[104]. Par exemple, la motricité dans un environnement complexe, dynamique et incertain résiste à une description purement symbolique ; la compréhension du langage naturel n'est pas seulement régie par des opérations lexicales et syntaxiques, mais également par une sémantique et une pragmatique qui sont difficilement formalisables ; même dans le cas du joueur d'échecs, les méthodes d'évaluation des coups possibles qu'utilisent les professionnels ne sont pas nécessairement formalisables avec rigueur, ni même simplement exprimables. Ainsi pour Dreyfus, ces problèmes ne peuvent être formalisés par des systèmes symboliques complexes, et encore moins par des « micromondes ». Le GPS ne peut donc prétendre à la résolution de « tous les problèmes » et reste ainsi limité aux cas idéaux et simplistes.

Bilan : « non IA faible (GPS) »

Dreyfus montre enfin comment, face à ces difficultés, l'idée d'un programme générique, capable de résoudre « tout type de problème », s'est soldée par un échec et est finalement abandonnée par Newell & Simon. Les machines réalisées dans le cadre du computationnalisme, dont le GPS et les systèmes experts, ne parviennent pas, dans de nombreux cas, à résoudre les problèmes complexes qui leurs sont posés. Ils ne parviennent pas à simuler l'intelligence. La première partie de *What Computers Can't Do* est donc consacrée à montrer empiriquement que « non IA faible » *dans le cas du computationnalisme*. Cependant, il ne s'agit (1) ni d'une objection *de principe* (mais bien d'un constat empirique), (2) ni d'une objection *systématique* (le computationnalisme peut réussir à résoudre certaines catégories de problèmes tels que les « micromondes »), (3) ni d'une objection *globale* à l'encontre de l'IA faible (seul les projets computationnalistes sont visés par la critique de Dreyfus).

Section 2.4.2. Partie 2 : les erreurs de la tradition cartésienne

Les postulats implicites du computationnalisme

Dans la seconde partie de *What Computers Can't Do*, Dreyfus met en évidence les origines philosophiques du modèle computo-représentationnel[105]. Il identifie quatre postulats implicites du computationnalisme, dont trois sont d'ordre philosophique :

---

[104] Dreyfus, H.L. 1979. « Chapitre 4 : phase IV (1972-1977) : l'I.A. s'attaque à la question de savoir comment représenter les connaissances usuelles. » *Op. cit.*, p. 145-184.

[105] Dreyfus, H.L. 1979. « Deuxième partie : sur quoi se fonde l'inébranlable optimisme de l'I.A. ? » *Op. cit.*, p. 189-291.

- Le postulat psychologique : « l'esprit peut être envisagé comme un système opérant sur des éléments binaires d'information, selon des règles formelles »[106]. C'est ce que nous avons identifié comme étant la condition nécessaire de la double hypothèse des systèmes symboliques physiques, formulée par Newell & Simon : l'esprit est un système symbolique au sein duquel la cognition manipule des symboles à l'aide d'opérations logiques (*cf.* section 2.3.1).

- Le postulat épistémologique : « tout savoir peut être explicitement formulé »[107]. Cette hypothèse est nécessaire à la réussite de programmes tels que le GPS. En effet, pour qu'il soit véritablement un « *General Problem* » *Solver*, le GPS doit assimiler et manipuler « tout type de connaissance ». Puisqu'il fonctionne à partir de données symboliques, toute connaissance doit donc être formalisée sous la forme explicite d'un SSP.

- Le postulat ontologique : « tout ce qui existe est un ensemble de faits, dont chacun est logiquement indépendant de tous les autres »[108]. L'indépendance des faits est nécessaire à la méthode des « micromondes ». Un programme comme le GPS doit pouvoir résoudre un problème donné à partir d'un ensemble isolé et borné de faits. Si ceux-là ne sont pas indépendants, la résolution d'un problème peut exiger l'examen d'un ensemble considérable de connaissances interdépendantes qui ne peuvent pas être contenues dans un seul « micromonde ».

La critique des fondements philosophiques

Dreyfus opère une critique radicale de ces trois postulats. Il affirme que (1) la cognition humaine utilise des processus radicalement différents de ceux employés par les ordinateurs et autres SSP, notamment en évitant de manière intuitive les difficultés inhérentes à la formalisation des problèmes (rejet du postulat psychologique), (2) la volonté d'expliciter toute connaissance sous forme de règles conduit à une régression infinie pour avoir « recours à des règles indiquant comment appliquer les règles »[109] (rejet du postulat épistémologique), (3) la connaissance dépend fortement d'un contexte cognitif global, ce qui rend impossible l'indépendance des faits et l'isolement de « micromondes » (rejet du postulat ontologique). Les trois postulats endossent alors la responsabilité

---

[106] *Ibid.*, p. 192 et « Chapitre 5 : le postulat psychologique », p. 201-234.

[107] *Ibid.*, p. 192 et « Chapitre 6 : le postulat épistémologique », p. 235-260.

[108] *Ibid.*, p. 193 et « Chapitre 7 : le postulat ontologique », p. 261-287.

[109] *Ibid.*, p. 290.

de l'échec du GPS et des autres tentatives d'application du computationnalisme. Si l'IA faible est en difficulté, explique Dreyfus, c'est parce qu'elle s'appuie sur des principes philosophiques erronés.

Dreyfus montre également comment ces trois postulats héritent en fait d'une très longue tradition philosophique. Parmi les précurseurs du computationnalisme, on peut alors citer Descartes et l'idée que l'esprit est un « miroir de la nature »[110], Hobbes et son « *"reason"* […] *is nothing but "reckoning"* »[111], Leibniz et l'espoir de construire un langage algorithmique capable de venir à bout des problèmes philosophiques[112], Kant et le fait que tout comportement humain est régit par des règles transcendantales qu'il faut s'efforcer d'expliciter, le premier Wittgenstein et l'atomisme logique, *etc.* Cette lignée philosophique (on parlera également de « tradition cartésienne ») a donc fourni un socle solide au computationnalisme et aux sciences cognitives grandissantes. Les trois postulats en sont le témoignage direct. Le GPS, parmi d'autres projets, participe à la mise en pratique de cette tradition philosophique. Comme nous le verrons dans la section suivante, Dreyfus en est un fervent opposant. Les erreurs du computationnalisme viennent donc, selon lui, des erreurs de la philosophie cartésienne concernant la nature de l'esprit et celle de la connaissance.

Bilan : « non IA forte → IA faible peu probable »

Dreyfus explique l'échec du computationnalisme à partir des erreurs de son socle philosophique. Cette stratégie critique repose donc sur la relation « non IA forte → non IA faible » : si on utilise un modèle erroné de la cognition, et que par conséquent il manque aux machines que l'on construit un aspect de la « véritable intelligence », alors il est très difficile en pratique d'engendrer des comportements intelligents. Dans le cas du GPS, les postulats du computationnalisme induisent une conception fallacieuse de la cognition, ce qui explique les difficultés techniques rencontrées par Newell & Simon.

Notons qu'il ne s'agit pas d'une stricte impossibilité. Comme nous l'avons vu dans la section précédente, le computationnalisme, lorsqu'il est appliqué à des problèmes bornés et contrôlés (vérifiant donc localement les postulats épistémologique et ontologique), réussit parfois à simuler

---

[110] *Cf.* également la critique rortyenne, s'appuyant sur les « philosophies révolutionnaires » de Wittgenstein et Heidegger, pour amorcer la remise en cause des postulats ontologiques cartésiens concernant la nature de la connaissance. Pour Rorty, notamment, il n'y a pas de faits objectifs indépendants du sujet connaissant. Rorty, R. 1980. *L'homme spéculaire*. [*Philosophy and the Mirror of Nature*.] Marchaisse, T. (trad.). Paris : Seuil, 1990.

[111] « La "raison" n'est rien d'autre qu'un calcul. » Cette position, défendue dans la première partie du *Léviathan*, est un précurseur du postulat psychologique.

[112] *Cf.* le fameux « *calculemus* » de Leibniz.

l'intelligence. La relation réellement défendue par Dreyfus est donc « non IA forte → IA faible peu probable ». En particulier, la contraposée « IA faible → IA forte » est rejetée par Dreyfus : ce n'est pas parce qu'on arrive à produire des comportements intelligents que le modèle de la cognition implémenté est nécessairement correct. Une réussite locale n'implique en aucun cas une validation du modèle computationnaliste. Ainsi, le projet initial du GPS consistant à reproduire les opérations cognitives de l'homme est voué à l'échec (« non IA forte (GPS) »), indépendamment de sa réussite éventuelle sur certains « micromondes » particuliers (« IA faible (GPS) »).

La collaboration entre philosophie et Intelligence Artificielle est amorcée par la mise en relation de projets techniques et des théories philosophiques sous-jacentes. Ici, c'est la philosophie de l'esprit qui est au service de l'Intelligence Artificielle, en évaluant les postulats philosophiques à l'origine des programmes développés. Dreyfus montre qu'un philosophe bien informé a son mot à dire quant aux développements pratiques de l'Intelligence Artificielle. Dans le cas du computationnalisme, la critique des postulats philosophiques sur la nature de la cognition (IA forte) amène à prédire et à expliquer les échecs de l'IA faible. La section suivante présente une stratégie similaire permettant de dépasser ces échecs et de fonder une « nouvelle IA » à partir d'une réflexion philosophique. Dans le chapitre 4.1, nous verrons comment la philosophie peut bénéficier d'une stratégie collaborative de la forme « non IA faible → non IA forte », où c'est au tour de l'IA faible d'expliciter les erreurs de la philosophie.

Section 2.4.3. Partie 3 : de nouveaux fondements pour l'IA

Le poids du paradigme dominant

Dans d'autres travaux, Dreyfus montre comment le poids de la tradition philosophique a conduit l'Intelligence Artificielle et les sciences cognitives à se forger un paradigme dominant – on parle parfois de « cognitivisme »[113] – au sein duquel l'hypothèse computationnaliste et la théorie computo-représentationnelle jouent des rôles importants. Ce paradigme a parfois agi comme un dogme, étouffant les approches hétérodoxes qui faisaient pourtant l'économie de certains des trois postulats du computationnalisme.

[113] Varela, F.J. 1988. *Invitation aux sciences cognitives*. [*Cognitive Science. A Cartography of Current Ideas*.] Lavoie, P. (trad.). Paris : Seuil, 1996, p. 35-51.

En 1988, c'est l'histoire difficile du connexionnisme qui est rapportée par Dreyfus et son frère[114]. Cette approche, dont les prémisses apparaissent très tôt dans l'histoire de l'Intelligence Artificielle[115], se libère notamment du postulat épistémologique qui prétend que « tout savoir peut être explicitement formulé. » En effet, les programmes connexionnistes sont construits sur deux niveaux. Le premier, appelé « niveau subsymbolique », est un réseau de composants logiques élémentaires. La syntaxe opérant sur ces composants est extrêmement simple en comparaison de la multitude des règles syntaxiques opérant sur les systèmes symboliques du computationnalisme. La connaissance, en tant que telle, émerge à un second niveau. Elle est une structure, un motif au sein du réseau subsymbolique, élaborée au cours d'un « apprentissage ». Elle est donc causalement liée à l'objet qu'elle représente (contrairement aux représentations des systèmes symboliques qui sont toujours arbitraires). Ainsi, le connexionnisme défend un mode de représentation *implicite* de la connaissance. Les concepts ne sont pas *explicitement* codés comme des composants logiques complexes du niveau symbolique. Ils sont contenus en puissance dans la structure globale du réseau. Le savoir n'a plus besoin d'être « explicitement formalisé » puisque le processus de formalisation est réalisé par le réseau lui-même, au cours de l'apprentissage. Le postulat épistémologique est ainsi relâché[116]. Malgré son intérêt conceptuel[117], le connexionnisme n'a pas eu, selon Dreyfus, le succès qu'il méritait lors des commencements de l'Intelligence Artificielle. Il s'est même fait « écrasé » par le

---

[114] Dreyfus, H.L., Dreyfus, S.E. 1988. « Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 309-333.

[115] Pour Boden, la démarche connexionniste est amorcée dès 1943, par les travaux de McCulloch, W.S., Pitts, W.H. 1943. « A Logical Calculus of the Ideas Immanent in Nervous Activity. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 22-39. Le connexionnisme est donc aussi vieux que l'Intelligence Artificielle (*cf.* note 1), même s'il faudra attendre un peu moins de quarante ans pour que les recherches sur les réseaux de neurones soient à nouveau répandues. Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 3.

[116] Voir également l'analyse de Stevan Harnad concernant la capacité des systèmes connexionnistes à ancrer les symboles dans la réalité et à faire ainsi émerger un niveau sémantique non-arbitraire. Harnad, S. 1990. « The Symbol Grounding Problem. » *Physica D*, vol. 42, p. 335-346.

[117] Voir également le travail de Paul Smolensky concernant les bases conceptuelles du connexionnisme. Smolensky, P. 1988. « Le traitement approprié du connexionnisme. » [« On the Proper Treatment of Connectionism. »] *In* Fisette, D. (éd.), Poirier, P. (éd.). 2003. *Philosophie de l'esprit : problèmes et perspectives*. Paris : Vrin, p. 197-215. Varela explique également comment le connexionnisme se débarrasse du calcul symbolique et s'affranchit ainsi d'une vision erronée de la connaissance. Varela, F.J. 1988. *Op. cit.*, p. 53-87.

paradigme computationnaliste plus facile à concevoir, plus direct à implémenter et, surtout, mieux défendu sur le plan philosophique[118].

Aujourd'hui encore, le computationnalisme a une place privilégiée en Intelligence Artificielle. De nombreuses recherches procèdent à partir de représentations symboliques explicites : la formalisation de connaissances à l'aide de logiques propositionnelles ou de grammaires formelles, les méthodes de raisonnement spatial et de raisonnement temporel, les moteurs d'inférences, les outils du « web sémantique », *etc*. John Haugeland parle de *Gool Old Fashion Artificial Intelligence* (GOFAI)[119] pour désigner ces approches classiques majoritaires. Pourtant, comme nous le verrons dans le chapitre 3.1, la critique du computationnalisme a libéré l'Intelligence Artificielle de son dogme, même si les nouveaux paradigmes qui ont émergé à partir des années 80 restent minoritaires et doivent souvent se justifier vis-à-vis des approches classiques.

La phénoménologie contre le cartésianisme

La troisième partie de *What Computers Can't Do* est consacrée au dépassement de la tradition cartésienne[120]. Dreyfus y propose « une autre vision des choses » en empruntant aux critiques de cette tradition, et notamment à la phénoménologie dont il est un fin connaisseur. En précisant les erreurs philosophiques du cartésianisme, il prétend montrer à l'Intelligence Artificielle ce qui lui fait défaut et ce que ses fondements philosophiques ne peuvent lui apporter. Sommairement, la critique de Dreyfus à deux angles d'attaque :

1. Ce qu'il manque aux programmes computationnalistes, c'est une inscription corporelle. Dreyfus souligne ici la place importante du corps dans les capacités cognitives humaines. La psychologie de la forme (théorie gestaltiste) rappelle par exemple que l'on saisit les phénomènes comme des ensembles structurés, des formes entières. Le GPS au contraire procède en décomposant les problèmes en faits atomiques indépendants. Selon Dreyfus, cette intuition globale qui fait défaut au GPS est justement donnée par notre corps[121]. La philosophie cartésienne, selon laquelle le corps n'est qu'une interface entre le sujet connaissant et la matière, et qui procède de manière analytique pour comprendre le monde (*i.e.* par décomposition), a induit en erreur les modèles computationnalistes. Notamment, les

[118] Dreyfus, H.L., Dreyfus, S.E. 1988. *Op. cit.*

[119] Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*, Cambridge, MA : MIT Press.

[120] Dreyfus, H.L. 1979. « Troisième partie : face aux postulats traditionnels : une autre vision des choses. » *Op. cit.*, p. 293-364.

[121] Dreyfus, H.L. 1979. « Chapitre 9 : le rôle du corps dans l'exercice de l'intelligence. » *Op. cit.*, p. 301-327.

fonctions cognitives de « bas-niveau », telle que la perception et la motricité, sont bien plus des fonctions corporelles que des fonctions rationnelles. Elles ne peuvent donc pas être réalisées par des systèmes symboliques[122].

2. Il manque également au GPS une notion de contexte, de situation globale, qui lui permettrait d'identifier les faits importants de manière immédiate. Ici, c'est le pragmatisme de Wittgenstein qui est opposée à la philosophie classique : les faits de notre expérience dépendent de nos intentions et de nos objectifs[123]. En d'autres termes, il n'existe pas de faits absolus indépendants de notre rapport au monde. Contrairement au GPS, qui doit repérer l'information importante au milieu d'une multitude de données symboliques, nous « construisons » les faits à partir d'un contexte général non-formalisé.

Bilan : « IA forte → IA faible probable »

Nous ne sommes pas intéressés ici par la critique de la tradition cartésienne dans la mesure où il s'agit d'un débat profondément philosophique et dont les enjeux ne concernent pas directement l'Intelligence Artificielle. Dreyfus en effet est avant tout un philosophe, intéressé par la phénoménologie et spécialiste de Husserl, Heidegger et Merleau-Ponty. Son ouvrage *What Computer's Can't Do* et ses autres travaux sur l'Intelligence Artificielle peuvent à ce titre être considérés comme une longue parenthèse au milieu de travaux de « philosophie fondamentale » concernant la nature de l'expérience et le contenu de la conscience. Ce qui nous intéresse ici, à proprement parler, ce n'est pas la critique phénoménologique du cartésianisme, mais plutôt ses conséquences pour l'Intelligence Artificielle. Nous sommes intéressés par la « philosophie appliquée » de Dreyfus.

La phénoménologie permet en effet d'identifier ce qu'il manque aux programmes de l'Intelligence Artificielle : un *corps* et une *situation*. Sans cela, ils ne sont pas « véritablement intelligents » et, du même coup, ils auront beaucoup de difficultés à engendrer des comportements qui le soient (« non IA forte → IA faible peu probable » comme nous l'avons vu dans la section

[122] Francisco J. Varela fait une critique similaire concernant la négligence du corps dans la philosophie cartésienne. Son *modèle énactif* repose en grande partie sur la réhabilitation inaugurée par la phénoménologie. Varela, F.J., Thompson, E.T., Rosch, E. 1991. *L'Inscription corporelle de l'esprit : sciences cognitives et expériences humaine*. [*The Embodied Mind: Cognitive Science and Human Experience*.] Havelange, V. (trad.). Paris : Seuil, 1993.

[123] Dreyfus, H.L. 1979. « Chapitre 10 : la situation : conduite ordonnée sans recours à des règles. » *Op. cit.*, p. 329-350.

précédente). Si les spécialistes de l'IA parviennent à donner à leurs programmes ce *corps* et cette *situation* que Dreyfus préconise, qu'est-ce que cela impliquerait ?

Même si ce n'est pas l'intention première de Dreyfus, qui est globalement pessimiste quant à l'avenir de l'Intelligence Artificielle, les points critiques qu'il apporte peuvent servir à dépasser les échecs de l'IA faible. La conception phénoménologique de l'esprit, en s'opposant au socle philosophique du computationnalisme, fournit elle-même un socle nouveau. Elle permet d'établir de nouveaux postulats et de forger de nouveaux paradigmes en Intelligence Artificielle. Cette démarche se fonde sur la réciproque de la relation « non IA forte → IA faible peu probable » utilisée par Dreyfus pour expliquer les échecs du computationnalisme. Elle consiste à dire qu'on a également « IA forte → IA faible probable » : une machine « véritablement intelligente » aura beaucoup d'aisance à produire des comportements intelligents. En d'autres termes, l'implémentation d'une théorie correcte concernant la nature de l'esprit a de grandes chances d'aboutir en pratique. La philosophie de l'esprit, en comparant et en évaluant les modèles de la cognition, est donc une source positive pour l'Intelligence Artificielle. Son rôle n'est pas seulement de mettre en évidence les erreurs conceptuelles (« non IA forte → IA faible peu probable »), mais également de proposer de nouveaux concepts, plus justes et adéquats, pour simuler l'intelligence (« IA forte → IA faible probable »). Le chapitre 3.1 montre comment la critique du computationnalisme, à laquelle Dreyfus a participé, a en effet déclenché une crise paradigmatique dans les années 80 et a donné lieu à l'édification d'une « nouvelle IA ». Ces nouveaux paradigmes tentent notamment de pallier les limites du modèle classique en empruntant ses concepts à la phénoménologie et au pragmatisme.

Pour résumer, la critique de Dreyfus ne devrait pas être abordée de manière trop négative par les spécialistes de l'IA. Même si les limites exposées sont pour Dreyfus difficilement surmontables, elles présentent des axes de recherche pertinents pour dépasser le paradigme computationnaliste. De manière générale, la démarche de Dreyfus et de ses successeurs en IA faible donne un exemple de véritable collaboration : un travail philosophique exploité sur le plan technique par l'Intelligence Artificielle (*cf.* chapitre 3.1). Cette stratégie collaborative nous intéresse dans la suite de ce mémoire. Dans le chapitre 3.2, nous utilisons la relation « IA forte → IA faible » pour mettre la philosophie au service de l'Intelligence Artificielle. Dans le chapitre 4.1, c'est sa contraposée « non IA faible → non IA forte » que nous exploitons pour renverser les rôles et mettre l'Intelligence Artificielle au service de la philosophie.

Chapitre 2.5. Bilan des stratégies possibles

Le tableau en page suivante (figure 3) résume les positions présentées dans cette partie. Il indique pour chaque démarche les opinions éventuelles concernant les possibilités de l'IA faible et de l'IA forte, ainsi que la relation défendue entre les deux problèmes et la stratégie collaborative qui en découle éventuellement.

Dans les parties suivantes, nous développons certaines de ces démarches pour opérer le rapprochement de la philosophie et de l'Intelligence Artificielle. Nous retenons essentiellement les démarches de Dreyfus (« non IA forte → IA faible peu probable » et « IA forte → IA faible »), pour faire de la philosophie le socle conceptuel de l'Intelligence Artificielle (partie 3), et la démarche de Levesque (« IA faible → IA forte »), pour opérer le renversement annoncé en introduction et mettre l'Intelligence Artificielle au service de la philosophie de l'esprit.

	IA faible	IA forte	Relation IA faible/IA forte	Stratégie collaborative	Sections
Spécialistes de l'IA (en général)	« IA faible »	Ne s'intéressent pas à l'IA forte	Indépendance disciplinaire	Aucune	section 2.1.1
Philosophes de l'esprit (en général)	Ne s'intéressent pas à l'IA faible	« IA forte » pour certains « non IA forte » pour d'autres	Indépendance disciplinaire		
Béhaviourisme méthodologique (Turing)	L'IA faible est résolue par le test	Ne se prononce pas sur l'IA forte	Indépendance épistémologique		section 2.1.2
Béhaviourisme logique (Ryle)		L'IA forte n'a pas de sens	Indépendance logique		section 2.2.1
Béhaviourisme logique (Hempel)		L'IA forte est résolue par le test	« IA faible → IA forte » (analytique)		
Impossibilité *par principe* de l'IA faible	« non IA faible »	« non IA forte »	« IA faible → IA forte »	La philosophie au service de l'IA	section 2.2.2
Nécessité *en pratique* de l'IA forte (Levesque)	« IA faible »	« IA forte »		L'IA au service de la philosophie	section 2.2.3
Critique du béhaviourisme (« zombie chanceux »)	« IA faible (zc) »	« non IA forte (zc) »	« non (IA faible → IA forte) »	Remise en cause des deux stratégies ci-dessus	section 2.2.4
Computationnalisme (Newell & Simon)	« IA faible (SSP) »	« IA forte (SSP) »	« IA faible ↔ IA forte »	Collaboration maximale	section 2.3.1
Critique de Searle (« chambre chinoise »)	« IA faible (cc) »	« non IA forte (cc) »	« non (IA faible → IA forte) »	Fin de la collaboration maximale	section 2.3.2
Critique de Searle (pouvoirs causaux)	Ne s'intéresse pas à l'IA faible	« non IA forte » (pétition de principe)	Indépendance épistémologique	Aucune	section 2.3.2
Critique de Dreyfus (computationnalisme)	« non IA faible (SSP) »	« non IA forte (SSP) »	« non IA forte → IA faible peu probable »	La philosophie au service de l'IA	section 2.4.1 section 2.4.2
Critique de Dreyfus (« nouvelle IA »)	« IA faible (phénoménologie) »	« IA forte (phénoménologie) »	« IA forte → IA faible »		section 2.4.3

Figure 3 : résumé des démarches collaboratives exposées dans la partie 2

Partie 3. La Philosophie au service de l'Intelligence Artificielle

Parmi les collaborations possibles entre Intelligence Artificielle et philosophie, la démarche d'Hubert L. Dreyfus semble extrêmement féconde. Dans la troisième partie de *What Computers Can't Do*, après avoir constaté l'échec du computationnalisme (section 2.4.1) et avoir expliqué les erreurs de la tradition cartésienne (section 2.4.2), Dreyfus expose « une autre vision des choses » (section 2.4.3). Il s'agit de révéler ce qu'il manque à Intelligence Artificielle pour produire des comportements intelligents. En révélant les lacunes philosophiques du computationnalisme, Dreyfus propose implicitement de nouveaux modèles. Une démarche collaborative peut ainsi être élaborée à partir de la relation « IA forte → IA faible » : l'implémentation d'un bon modèle de la cognition, d'une « véritable intelligence », permet de produire en pratique des comportements intelligents. La philosophie de l'esprit aide ainsi l'Intelligence Artificielle à forger et à choisir ses modèles.

Le chapitre 3.1 présente brièvement les paradigmes qui ont émergés dans les années 80, en réaction au computationnalisme, et qui reposent sur des principes philosophiques anticartésiens, notamment défendus par Dreyfus. Le chapitre 3.2 généralise la démarche en élargissant le champ des interventions philosophies. Ici, c'est une théorie métaphysique concernant la structure de la réalité et la nature de la connaissance qui est adaptée aux travaux pratiques de l'Intelligence Artificielle, pour aider à la simulation de phénomènes complexes. Nous montrons que cet exemple de collaboration repose également sur la relation « IA forte → IA faible ».

Chapitre 3.1. La « nouvelle IA »

Par opposition à la *Good Old Fashioned AI* (GOFAI)[124], la « nouvelle IA » est une constellation de paradigmes apparus pour la plupart dans les années 80. Ces approches ne sont pas nécessairement compatibles deux-à-deux, mais elles s'accordent sur un point : la volonté de dépasser le modèle computationnel de la cognition. Sur le plan conceptuel, il est aisé de voir que ces nouveaux modèles s'appuient sur les critiques du computationnalisme, notamment celles formulées par Searle et par Dreyfus. Ce travail d'« analyse philosophique » est alors un moteur conceptuel pour l'Intelligence Artificielle. Sur le plan disciplinaire, il faudrait montrer qu'il y a bien une filiation entre les critiques du computationnalisme en philosophie et les nouveaux paradigmes de l'Intelligence Artificielle. Dans ce chapitre, nous nous appliquons essentiellement à mettre en évidence des relations *conceptuelles*. Nous nous contenterons de remarquer que les articles référencés dans ce chapitre sont représentatifs des nouveaux paradigmes dans la mesure où ils sont régulièrement cités en introductions d'articles de fond des paradigmes en question. De plus, ces articles font référence à de nombreux travaux de philosophie, notamment pour expliciter et justifier des travaux en Intelligence Artificielle. Ainsi, les articles des nouveaux paradigmes citent, de manière directe ou indirecte, des travaux de nature philosophique.

Dans la section 2.4.3, nous avons vu assez sommairement que Dreyfus identifie deux principes qui font défaut à la GOFAI, mais qui sont essentiels pour simuler l'intelligence : (1) le corps tient un rôle important dans l'exécution de fonctions cognitives de « bas-niveaux », telles que la perception et la motricité, et (2) la cognition humaine agit en fonction d'un contexte, d'une situation globale, qui ne peut être formalisé en terme de règles. Ce *corps* et cette *situation* constituent donc les Graals de la « nouvelle IA ».

- Nous avons vu que le *connexionnisme*[125] représente les connaissances à l'aide de structures non-arbitraires au sein de « réseaux subsymboliques » (*cf.* section 2.4.3). L'objectif est de définir des représentations implicites de l'environnement via une procédure d'apprentissage, au lieu de représenter explicitement les paramètres de l'environnement par des symboles arbitraires (comme le fait le computationnalisme). Le connexionnisme donne ainsi une

[124] Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*, Cambridge, MA : MIT Press.

[125] Dreyfus, H.L., Dreyfus, S.E. 1988. « Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 309-333 ; Smolensky, P. 1988. « Le traitement approprié du connexionnisme. » [« On the Proper Treatment of Connectionism. »] *In* Fisette, D. (éd.), Poirier, P. (éd.). 2003. *Philosophie de l'esprit : problèmes et perspectives*. Paris : Vrin, p. 197-215.

première approche pour dépasser la formalisation strictement symbolique du computationnalisme et définir un niveau de représentations émergentes. Il répond donc en partie au besoin d'une situation globale pour engendrer des comportements intelligents.

- La *robotique incarnée* et la *robotique évolutive*[126] construisent des robots qui agissent intelligemment sans utiliser de représentations de leur environnement. Tout comme le connexionnisme donc, ces nouvelles robotiques ne cherchent pas à formaliser explicitement la situation dans laquelle se trouve le robot, mais à définir des modes de réactions non-symboliques qui lui permettent, par composition et émergence, d'interagir avec l'environnement. La *robotique incarnée* s'intéresse ainsi à une intelligence *réactive* de « bas-niveau », et non à une intelligence *cognitive* de « haut-niveau ». Le corps y prend une place essentielle puisque c'est lui qui implémente ces modes de réactions à l'aide de senseurs et d'effecteurs reliés directement. Elle répond donc au second besoin mis en évidence par Dreyfus, concernant le rôle du corps dans la cognition.

- Les *systèmes dynamiques*[127] (que nous présenterons plus en détails dans le chapitre 4.2) constituent un large domaine de recherche dont se réclament parfois le connexionnisme et les nouvelles robotiques. Les systèmes dynamiques, eux-aussi, se passent de représentations. Ils sont *physiquement couplés* avec l'environnement, c'est-à-dire que leur état interne change en fonction de perturbations extérieures de manière continue. Ce couplage dépend directement de la conception physique du système et de son immersion dans l'environnement. Ainsi, le corps de la machine a également une importance essentielle dans la production des comportements. De plus, le couplage physique n'utilise pas de représentations symboliques implicites. Il est le résultat de principes purement physiques. Les systèmes dynamiques répondent ainsi aux deux défauts mis en évidence par Dreyfus.

Ces trois paradigmes exploitent de nouvelles conceptions de la cognition, développées en la philosophie, pour construire des machines intelligences. Ces nouveaux modèles s'appuient tous sur

---

[126] Brooks, R.A. 1991. « Intelligence without representation. » *Artificial Intelligence*, vol. 47, p. 139-159 ; Kaplan, F., Oudeyer, P.-Y. 2008. « Le Corps comme variable expérimentale. » *Revue philosophique de la France et de l'étranger*, vol. 133, n°3.

[127] McGeer, T. 1990. « Passive dynamic walking. » *International Journal of Robotics Research (IJRR'90)*, vol. 9, n°2, p. 62-82 ; Beer, R.D., Quinn, A.D., Chiel, H.J., Ritzmann, R.E. 1997. « Biologically Inspired Approaches to Robotics: What can we learn from insects? » *Communications of the ACM*, vol. 40, n°3, p. 30-38 ; van Gelder, T. 1998. « Dynamique et cognition. » [« Dynamics and Cognition. »] Lapointe, S. (trad.). *In* Fisette, D. (éd.), Poirier, P. (éd.). 2003. *Philosophie de l'esprit : problèmes et perspectives*. Paris : Vrin, p. 329-369.

un principe philosophique fort : pour de nombreuses tâches cognitives, les hommes n'utilisent pas de représentations formelles de leur environnement, mais des mécanismes corporels réactifs. Ce modèle de la cognition hérite notamment d'une position antireprésentationnaliste qui s'oppose directement au principe cartésien selon lequel l'esprit est un « miroir de la nature »[128]. Plus largement, le « modèle énactif » de Francisco J. Varela (qui sera également présenté dans la section 4.3.3) retrace bien le cheminement philosophique qui est parti du computationnalisme et est arrivé aux modèles incarnés de la cognition, en passant également par le connexionnisme[129]. On parle également du « tournant pragmatique »[130] des sciences cognitives pour désigner la place importante de l'action dans les nouveaux travaux d'Intelligence Artificielle, issues des conceptions pragmatistes de l'esprit.

Pour conclure, de la même manière que la tradition cartésienne a été un vecteur conceptuel important pour la GOFAI, de nombreux autres courants philosophiques ont été appliqués à la production de comportements intelligents. Dès lors qu'on postule le bien-fondé de ces courants philosophiques, leur application relève de la démarche collaborative « IA forte → IA faible »[131].

---

[128] *Cf.* notamment la critique de Rorty, R. 1980. *L'homme spéculaire*. [*Philosophy and the Mirror of Nature*.] Marchaisse, T. (trad.). Paris : Seuil, 1990.

[129] Varela, F.J., Thompson, E.T., Rosch, E. 1991. *L'Inscription corporelle de l'esprit : sciences cognitives et expériences humaine*. [*The Embodied Mind: Cognitive Science and Human Experience*.] Havelange, V. (trad.). Paris : Seuil, 1993. *Cf.* également le résumé de ce cheminement dans Varela, F.J. 1988. *Invitation aux sciences cognitives*. [*Cognitive Science. A Cartography of Current Ideas*.] Lavoie, P. (trad.). Paris : Seuil, 1996.

[130] Engel, A.K. 2010. « Directive Minds: How Dynamics Shapes Cognition. » *In* Stewart, J.R. (éd.), Gapenne, O. (éd.), Di Paolo, A.E. (éd.). 2010. *Enaction: Toward a New Paradigm for Cognitive Science*. Cambridge, MA: MIT Press, p. 219-243.

[131] Pour une évaluation trente ans plus tard, par Dreyfus, de ces nouveaux paradigmes et de leurs applications, voir Dreyfus, H.L. 2007. « Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian. » *Philosophical Psychology*, vol. 20, n°2, p. 247-268.

Chapitre 3.2. La simulation des phénomènes émergents

Dans ce chapitre, nous donnons un cadre plus général à la stratégie collaborative présentée ci-dessus. Nous ne nous intéressons plus seulement à la simulation de comportements intelligents mais, plus largement, à la simulation de « phénomènes complexes ». En effet, l'IA faible a été définie dans le chapitre 1.3 comme un ensemble d'outils permettant d'assister les recherches scientifiques concernant la nature de la cognition. Cependant, cela est également valable pour d'autres domaines scientifiques tels que la sociologie, l'économie, la biologie, la physique, *etc.* Dans ce chapitre, l'IA faible a donc pour objectif de reproduire et d'expliquer des phénomènes sociaux, biologiques ou physiques en simulant leurs dynamiques fondamentales. Plus particulièrement, nous nous intéressons à la notion de « phénomènes émergents ».

L'IA forte désigne alors les positions philosophiques concernant la nature de tels phénomènes. Nous verrons que celles-ci ne sont pas toujours en adéquation avec les méthodes de l'IA faible et que la *simulation* des « phénomènes émergents » et la *réalité* de ces phénomènes peuvent diverger. La relation « IA forte → IA faible » invite justement à emprunter aux théories philosophiques des modèles pour simuler, en pratique, des phénomènes complexes. Ce mode de collaboration ne fait plus intervenir le champ restreint de la philosophie de l'esprit, mais également des positions métaphysiques concernant la nature des « phénomènes émergents ». Nous voulons ainsi montrer que la philosophie, dans son intégralité, peut servir aux recherches en Intelligence Artificielle.

Ce chapitre enfin reprend des travaux réalisés dans le cadre d'une thèse en Intelligence Artificielle et présentés dans un atelier de recherche et un colloque sur les Systèmes Multi-Agents[132].

Section 3.2.1. Un problème pratique : les phénomènes émergents

Le paradigme multi-agents

Les Systèmes Multi-Agents (SMA) constituent une méthodologie récente de l'Intelligence Artificielle, également apparue dans les années 80. Elle fait partie de l'« Intelligence Artificielle distribuée » portant sur des systèmes de raisonnements constitués de processus parallèles, dépassant ainsi le cadre classique du raisonnement séquentiel. Tout comme les réseaux

[132] Lamarche-Perrin, R. 2011a. « Conceptualisation de l'émergence : dynamiques microscopiques et analyse macroscopique des SMA. » *Atelier Futur des Agents et des Multi-Agents (FUTURAMA'11)*. Chambéry : Plateforme AFIA 2011, mai 2011 ; Lamarche-Perrin, R., Demazeau, Y., Vincent, J.-M. 2011. « Observation macroscopique et émergence dans les SMA de très grande taille. » *Journées Francophones sur les Systèmes Multi-Agents* (JFSMA'11)*,* oct. 2011, Valenciennes : Cépaduès, p. 53-62.

connexionnistes, les SMA font intervenir deux niveaux : un niveau *microscopique*, où interagissent un ensemble d'agents, et un niveau *macroscopique*, comprenant les dynamiques du système pris dans sa globalité et au sein duquel s'inscrivent les « phénomènes émergents ». La relation entre ces deux niveaux de modélisation constituent alors un problèmes scientifique délicat.

Le terme « SMA » désigne ainsi des systèmes distribués, artificiels ou naturels, constitués d'un ensemble d'agents. Un « agent » est « une entité autonome, réelle ou abstraite, qui est capable d'agir sur elle-même et sur son environnement, qui, dans un univers multi-agent, peut communiquer avec d'autres agents, et dont le comportement est la conséquence de ses observations, de ses connaissances et de ses interactions avec les autres agents. »[133] Les SMA reposent donc sur cette abstraction, pouvant désigner tour-à-tour des entités physiques (*e.g.*, processus d'un système informatique, constituants élémentaires d'un réseaux de neurones formels, services Web) ou virtuelles (*e.g.*, particules d'un système gazeux, cellules d'un organisme, individus dans un système social), ayant toujours pour caractéristique principale une relative autonomie de comportement vis-à-vis du système et des autres agents.

La notion d'émergence au sein des SMA

Les SMA, et plus généralement l'Intelligence Artificielle distribuée, font ainsi intervenir au moins deux niveaux d'abstraction. Au niveau microscopique, les agents participent *localement* aux dynamiques du système. Ils sont notamment *situés* dans leur environnement et disposent de ce fait de connaissances et de moyens d'actions limités à leur *voisinage*. Au niveau macroscopique, le SMA lui-même, c'est-à-dire la somme des agents et de leurs interactions, fournit un second niveau d'abstraction. Le système pris dans sa globalité offre un comportement bien plus riche et bien plus complexe que celui de ses agents. C'est cette dynamique globale, parfois nommée « intelligence collective »[134], qui fait la particularité des SMA. On parle de « comportements émergents » dans la mesure où leurs mécanismes ne sont pas explicitement formalisés au niveau des entités microscopiques, mais qu'ils sont révélés par la mise en relation des agents, en tant que structures

---

[133] Chaib-Draa, B., Jarras, I., Moulin, B. 2001. « Systèmes multiagents : principes généraux et applications. » *In* Briot, J.-P. (éd.), Demazeau, Y. (éd.). 2001. *Principes et architecture des systèmes multiagents*. Paris : Hermes, p. 28. Cette définition est synthétisée à partir des propriétés énoncée par Ferber, J. 1995. *Les Systèmes multi-agents : vers une intelligence collective*. Paris : InterEditions. Pour une introduction plus détaillée à la notion d'agent en informatique, voir également Russell, S.J., Norvig, P. 2003. *Artificial Intelligence: A Modern Approach,* 3rd ed. Upper Saddle River, NJ : Prentice Hall, p. 34-63 et Wooldridge, M. 2009. *An Introduction to MultiAgent Systems* (2nd ed.). Chichester, WS: John Wiley & Sons.

[134] Ferber, J. 1995. *Op. cit.*

distribuées au sein du système. Les « phénomènes émergents » sont donc des propriétés ou des processus globaux induits par les propriétés et les processus locaux des parties du système.

Deux objectifs scientifiques bénéficient de cette notion de « comportements émergents ». Premièrement, ces comportements constituent une abstraction utile à la *modélisation* des systèmes. Ils sont notamment utilisés pour décrire, simuler et prédire le comportement de systèmes complexes naturellement distribués (systèmes physiques, biologiques, sociaux, *etc.*). L'enjeu épistémologique consiste alors à expliciter les relations de causalité entre le niveau microscopique et le niveau macroscopique. Par exemple : quels modes d'interaction locale permettent à un organisme de conserver son intégrité globale au cours du temps ? Comment le comportement des acteurs économiques influence la dynamique globale des marchés financiers ? Y a-t-il également une « causalité descendante » du système sur ses agents, de l'organisme sur ses organes et des marchés financiers sur ses acteurs économiques ?

La seconde utilisation des « comportements émergents » concerne la *résolution* de problèmes distribués. Contrairement à l'Intelligence Artificielle classique, la solution engendrée par un SMA n'est pas le fait d'un algorithme séquentiel, centralisant toutes les connaissances nécessaires à la résolution d'un problème, mais par un ensemble distribué d'agents autonomes. La méthode de résolution n'est donc pas explicitement implémentée au niveau des agents, mais résulte de leurs interactions. Elle *émerge* au niveau macroscopique. Une telle approche est particulièrement utile dans le cas de problèmes naturellement distribués (*e.g.*, répartition de ressources entre les nœuds d'un réseau, recherche d'informations sur le Web) ou dans le cas de problèmes complexes (*e.g.*, exploration dans un environnement dynamique de très grande taille, désambiguation sémantique). L'approche SMA offre par ailleurs, grâce à la notion d'émergence, des qualités de « résistance aux pannes » et d'« adaptation au contexte » qu'il est très difficile d'obtenir à partir d'un programme séquentiel classique, notamment grâce au fait que l'information y est distribuée et redondante.

Problématique de recherche

Pour ces deux contextes de recherches (*simulation* de systèmes complexes et *résolution* de problèmes distribués), la notion d'émergence a donc un rôle essentiel. Elle permet de définir le niveau macroscopique où réside la complexité du système ou du problème abordé. Il convient donc de définir précisément ce que l'on entend par « phénomènes émergents » et quelles sont les relations que ceux-ci entretiennent avec le niveau microscopique. De plus, lorsqu'on s'intéresse à des systèmes de très grande taille (évolution d'un grand nombre d'agents sur de longues périodes de temps), complexes (structures organisationnelles et interactionnelles non-triviales), hétérogènes (agents et protocoles d'interaction très différents les uns des autres) et ouverts (des agents peuvent

entrer ou sortir du système), la compréhension du niveau microscopique ne suffit pas à la compréhension globale du système. En effet, les descriptions microscopiques de tels systèmes sont extrêmement complexes et coûteuse en termes d'analyse. En pratique, les spécialistes de l'IA ont besoin d'abstractions solides pour concevoir, décrire et optimiser leurs systèmes en dépassant le seul niveau microscopique. La notion d'émergence devient alors nécessaire pour engendrer des modèles macroscopiques et appliquer les méthodes classiques de l'Intelligence Artificielle aux systèmes complexes de très grande taille.

Seulement, les SMA développés actuellement sont, bien souvent, entièrement conçus au niveau des agents. Les processus en présence sont donc distribués (aucun agent ne dispose d'une connaissance complète du système), décentralisés (aucun espace de mémoire commune n'est accessible à l'ensemble des agents) et asynchrones (aucune horloge globale ne donne de référentiel temporel commun). Nous parlons alors de « SMADAG »[135] (Systèmes Multi-Agents Décentralisés, Asynchrones et de très Grande taille). Dans ce contexte particulier, les outils classiques pour élaborer des abstractions spatiales et temporelles manquent aux spécialistes (absence de représentations globales centralisées et d'horloges globales synchronisées). Il est alors difficile de fonder le niveau macroscopique à partir du niveau microscopique. La problématique générale des SMADAG peut alors être formulée de la manière suivante : « comment obtenir une description macroscopique de processus entièrement microscopiques ? »

Stratégie collaborative

Nous soutenons qu'une conceptualisation propre et solide de la notion d'émergence permet de résoudre la problématique de recherche des SMADAG et ainsi d'aider à la simulation de systèmes complexes et à la résolution de problèmes distribués. De nombreuses acceptions de l'émergence, plus ou moins formalisées, ont été proposées dans des travaux d'Intelligence Artificielle[136]. Cependant, ces conceptualisations divergent sur les qualités fondamentales des « phénomènes émergents ». Pour démêler le nœud conceptuel et pour fournir une base solide à la conception et à l'analyse des SMA, nous allons emprunter des concepts à la philosophie.

---

[135] Lamarche-Perrin, R., Demazeau, Y., Vincent, J.-M. 2011. *Op. cit.*

[136] Pour une analyse non-exhaustive des différentes conceptualisations de l'émergence en Intelligence Artificielle, voir Lamarche-Perrin, R. 2011a. *Op. cit.* ; Deguet, J., Demazeau, Y., Magnin, L. 2006. « Element about the Emergence Issue: A Survey of Emergence Definitions. » *ComPlexUs*, vol. 3, p. 24-31 ; Picard, G. 2004. *Méthodologie de développement de SMA adaptatifs et conception de logiciels à fonctionnalité émergente*. Thèse de doctorat, Gleizes, M.-P. (dir.). Toulouse : Université Paul Sabatier, déc. 2004.

La section 3.2.2 reprend les termes du débat philosophique au sein duquel la notion d'émergence est apparue au tournant du XIX^e^ siècle. La notion particulière que nous retenons pour la problématique des SMADAG est celle d'« émergence épistémique ». Dans la section 3.2.3, nous adaptons les termes du débat philosophique au contexte des SMA. Deux contraintes d'ordre méthodologique sont alors formulées : le « monisme » et le « non-éliminativisme ». Nous montrons ainsi qu'une théorie philosophique adéquate et bien formulée peut, lorsqu'elle est appliquée sous forme de contraintes méthodologiques, résoudre un problème pratique de l'Intelligence Artificielle. Ici, il s'agit de la théorie de l'« émergence épistémique » pour la simulation de systèmes complexes et la résolution de problèmes distribuée. Cette démarche particulière exemplifie donc la stratégie collaborative basée sur la relation « IA forte → IA faible ».

Section 3.2.2. L'émergence épistémique en philosophique

Les limites du dualisme et du monisme

Cette section résume et schématise le cadre conceptuel au sein duquel la notion d'émergence a été développée, par la philosophie britannique, à la fin du XIX^e^ siècle[137]. Il s'agit du débat opposant deux courants philosophiques concernant la nature du vivant : le *vitalisme* et le *mécanisme*, deux positions ontologiques concernant la structure de la réalité et la nature de la connaissance. En ce sens, il s'agit d'un débat métaphysique. La controverse commence ainsi : philosophes et scientifiques sont confrontés à deux catégories d'objets, les « êtres vivants » et les « êtres inanimés », chacune d'elles ayant des spécificités empiriques. Il est important de rendre compte de ces spécificité, d'expliquer *en quoi* et *pourquoi* les êtres vivants et les êtres inanimés se présentent différemment à l'examen scientifique. Existe-t-il une différence de nature entre les deux catégories d'objets ou sont-elles au contraire relativement similaires ? Schématiquement, deux positions ontologiques s'affrontent :

1. Le *vitalisme* postule l'existence de deux substances distinctes (la matière inanimée et un principe de force vitale) pour rendre compte des deux catégories d'objets (êtres inanimés et êtres vivants). Il s'agit donc d'un *dualisme de substances* et, plus généralement, d'une position *non-réductionniste*. Pour le vitalisme en effet, le principe de force vital est ontologiquement indépendant des principes régissant la matière inanimée. Les phénomènes

[137] Pour une analyse historique plus détaillée de cette controverse, à laquelle participent notamment J.S. Mill, C.D. Broad et S. Alexander, voir notamment O'Connor, T., Wong, H.Y. 2006. « Emergent Properties. » *In Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/properties-emergent/, mis en ligne le 24 sept. 2002, révisé le 23 oct. 2006, consulté le 1 mai 2011.

du vivant ne peuvent donc être entièrement expliqués par l'analyse de leurs propriétés matérielles. En termes plus modernes, le vitalisme affirme que les lois de la biologie ne peuvent être « déduites de » ou « réduites aux » lois de la physique et de la chimie.

2. Le *mécanisme* affirme au contraire qu'il y a une seule substance : la matière inanimée. Les êtres vivants résultent simplement d'une disposition particulière de cette substance. Il s'agit donc d'un *monisme* et, plus généralement, d'une position *réductionniste*. Pour le mécanisme, les phénomènes du vivant sont entièrement déterminés par les principes agissants sur la matière inanimée. Ils doivent faire partie de la même théorie scientifique. Autrement dit, pour le mécanisme les lois de la biologie sont réductibles à celles de la physique et de la chimie.

L'opposition classique entre dualisme et monisme témoigne de positions ontologiques tranchées en faveur ou en défaveur du réductionnisme. Ici, les termes sont appliqués à la biologie et aux êtres vivants : sont-ils ou non de même nature que les êtres inanimés ? Cependant, aucune de ces deux positions ontologiques n'est entièrement satisfaisante pour rendre compte des phénomènes complexes du vivant. Le dualisme, puisqu'il multiplie les hypothèses concernant l'existence de substances indépendantes, est *ontologiquement coûteux*. Il ne répond pas au principe de parcimonie selon lequel « *pluralitas non est ponenda sine necessitate* »[138]. C'est le fameux « rasoir d'Ockham » qui, à pouvoir explicatif égal, privilégie les théories simples aux théories complexes. Par ailleurs, le dualisme met en péril l'idée chère aux scientifiques d'une possible unification de la science. S'il y a deux substances, alors il ne peut y avoir une science fondamentale unique. Dans le cas du vitalisme, c'est la biologie qui est également une science fondamentale, de la même manière que la physique des particules (par exemple).

Le monisme est préférable au dualisme puisqu'il réalise une « économie ontologique » en faisant l'hypothèse d'une seule substance. Il donne ainsi une base explicative unique à l'ensemble des phénomènes observables (vivants et inanimés) et vérifie le principe de parcimonie. Cependant, au motif d'une telle réduction vers la physique fondamentale, le monisme incite à éliminer les sciences spéciales, c'est-à-dire à en supprimer la pertinence scientifique. Dans le cas du mécanisme, les lois de la biologie peuvent être exprimées à partir de celles de la physique. Une dérive consiste alors à affirmer que la biologie est une science dont on peut se passer et que, à terme, les explications biologiques devraient être réduites à des explications physiques et ainsi éliminées de la méthode scientifique. Dès lors, un seul mode de connaissance est privilégié. Dans ce contexte, il est

[138] « Les multiples ne doivent pas être utilisés sans nécessité. »

difficile de rendre compte des différences essentielles entre êtres inanimés et êtres vivants puisque ces deux catégories d'objets sont régies par les mêmes lois et expliquées par la même science. Le monisme réductionniste, par sa tendance éliminativiste, peine à expliciter les distinctions empiriques entre les phénomènes. Il perd en pouvoir explicatif et, en pratique, il est difficile de décrire les phénomènes complexes tels que les systèmes biologiques à partir de concepts physiques uniquement. On dit alors que le monisme est *épistémiquement faible*, c'est-à-dire qu'il explique en soi moins de choses que le dualisme[139].

Il est apparu nécessaire à certains philosophes britanniques de dépasser ces difficultés en proposant une « voie moyenne » entre le dualisme et le monisme. Comment défendre une position ontologique qui est économe sur le plan ontologique, mais qui permet d'expliquer aisément les différences empiriques entre êtres vivants et êtres inanimés ? Comment conserver le pouvoir explicatif du vitalisme et la sobriété du mécanisme ? En d'autres termes, comment fonder un *monisme non-éliminativiste* ? C'est justement le projet de la philosophie émergentiste.

La voie moyenne de la philosophie émergentiste

Selon O'Connor & Wong, la position émergentiste a été développée dans le cadre de cette controverse sur la nature du vivant, pour proposer une alternative au dualisme de substance et au monisme réductionniste[140]. Elle consiste à faire une distinction entre la nature d'un objet et le cadre scientifique qui en fait l'analyse, c'est à dire une distinction entre ontologie et épistémie. Ainsi, les êtres vivants sont *ontologiquement* similaires aux êtres inanimés (ils sont faits de la même substance), mais on les distingue sur le plan *épistémique* pour rendre compte de leurs particularités (ils ne sont pas étudiés par les mêmes sciences). Pour être efficace, la méthode scientifique ne doit pas suivre le modèle ontologique. Ainsi, même si les phénomènes du vivant sont *en principe* explicables à partir des seules lois de la physique, ils doivent *en pratique* être étudiés à un niveau de description adéquat. Devant la complexité de certains phénomènes, les sciences spéciales fournissent des abstractions utiles, voire nécessaires, mais elles ne constituent *en aucun cas* un engagement quant à la structure de la réalité. Dès lors, la distinction entre êtres inanimés et êtres vivants est « dans l'œil du scientifique. »

Cette position est compatible avec le monisme puisque les lois de la biologie peuvent *en principe* être réduites à celles de la physique. Les êtres inanimés et les êtres vivants sont faits de la

[139] Pour une réflexion plus complète sur les conséquences du dualisme et du monisme sur l'unité de la science, sur son pouvoir explicatif et sur la notion d'élimination, voir Kistler, M. 2007. « La Réduction, l'émergence, l'unité de la science et les niveaux de réalité. » *Matière Première*, vol. 2, p. 67-97.

[140] O'Connor, T., Wong, H.Y. 2006. *Op. cit.*

même substance. Cette position est également *non-éliminativiste*, dans la mesure où la biologie est utile *en pratique* au scientifique et qu'elle obtient ainsi une responsabilité épistémique indéniable. Ce *monisme non-éliminativiste* est nommé « émergentisme épistémique »[141]. Les trois positions analysées dans cette section sont représentées dans le schéma ci-dessous (figure 4), en fonction de leurs positionnements ontologiques et épistémiques. L'émergentisme y est vu comme une synthèse du dualisme et du monisme.

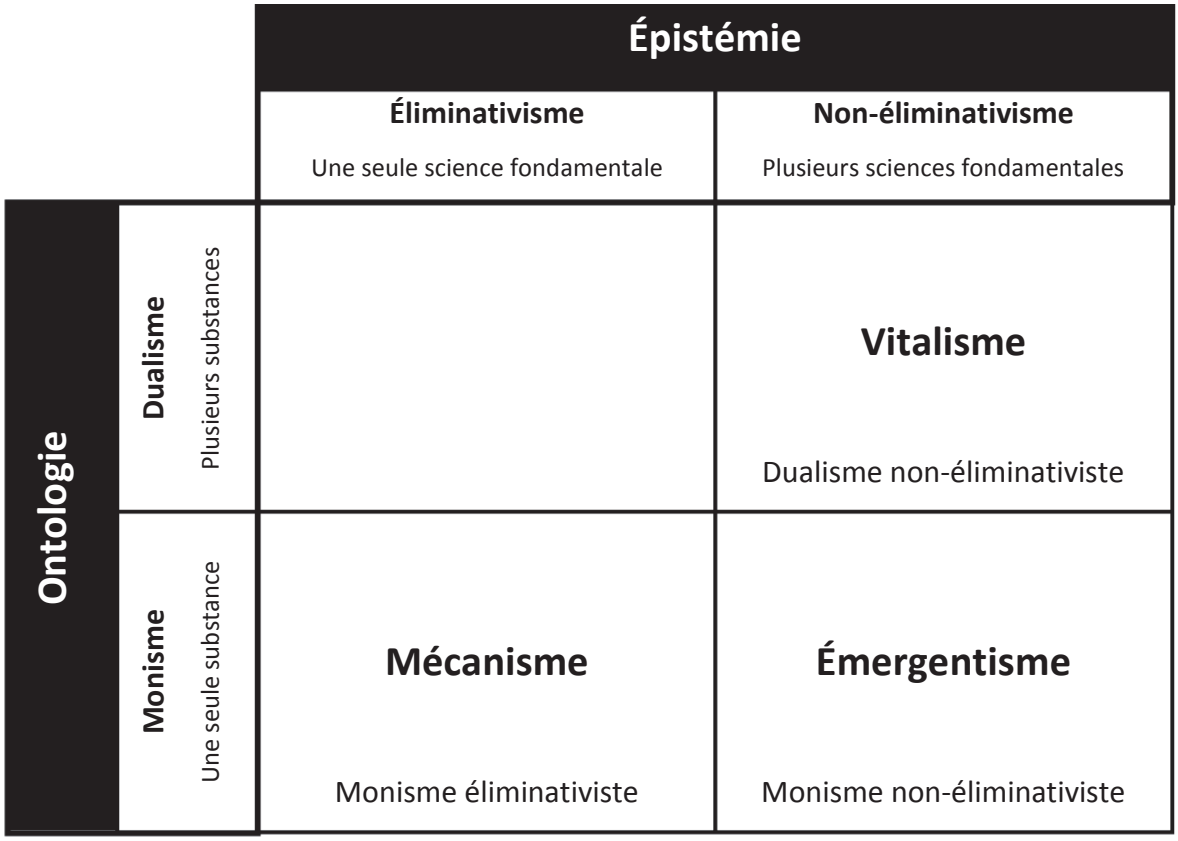

Figure 4 : représentation schématique des positions métaphysiques

[141] Il existe plusieurs notions d'émergence, dont l'« émergence ontologique » qui est souvent assimilée à une forme sophistiquée de dualisme (notamment parce qu'elle présuppose une causalité descendante du niveau émergent vers le niveau physique. *Ibid.*) La distinction que nous faisons entre « émergence épistémique » et « émergence ontologique » est similaire à la distinction entre « émergence faible » et « émergence forte » proposées par plusieurs auteurs. Voir par exemple Bedau, M.A. 1997. « Weak Emergence. » *Philosophical Perspectives*, vol. 11, p. 375-399 et Stephan, A. 1999. « Varieties of Emergentism. » *Evolution and Cognition*, vol. 5, n°1, p. 49-59.

Section 3.2.3. L'émergence épistémique en Intelligence Artificielle

Une analogie pour appliquer le concept

Afin d'appliquer le concept d'émergence épistémique au contexte des SMA nous utilisons l'analogie suivante :

1. L'*ontologie* d'un système informatique désigne tout ce qui est relatif à sa *conception*. C'est-à-dire tout ce qui est en amont de son exécution : ses spécifications, son modèle abstrait, son code, son implémentation matérielle, *etc.*

2. L'*épistémie* du système désigne tout ce qui est relatif à son *analyse*, en aval de son exécution. Elle comprend les phases d'observation, de description et d'analyse de l'exécution et des résultats qu'elle engendre.

De la même manière qu'en philosophie, l'ontologie d'un système informatique relève de l'objet *per se*, tandis que l'épistémie se rapporte à l'objet relativement à *un mode de connaissance donné*. La notion d'exécution est centrale pour faire cette distinction dans la mesure où elle définit une étape importante dans l'élaboration d'un programme, entre le moment où il est un programme *en puissance* (l'algorithme est alors un *objet mathématique*) et le moment où il réalise véritablement un calcul (l'algorithme est alors un *processus physique*). La nature de l'objet étudié change radicalement lors de cette étape. Elle nous paraît constituer une bonne limite entre l'ontologie (ce qu'*est* réellement le programme au niveau mathématique) et l'épistémie (la manière dont il se donne empiriquement : ses réalisations physiques et les méthodes d'analyse que nous leurs appliquons). Dans ce qui suit, les termes du débat philosophique exposé précédemment sont traduits sur la base de cette analogie : l'*épistémie* d'un objet est à son *ontologie* ce que l'*analyse* d'un programme est à sa *conception*.

Le dualisme en informatique

Par analogie, que sont les « approches dualistes » dans le cas des SMA ? Il s'agit de systèmes dont la *conception* comprend au moins deux niveaux de modélisations (par analogie : deux substances). On dit parfois que le niveau macroscopique est réifié[142], c'est-à-dire qu'il apparaît comme un objet *au sein même du système* : dans sa description mathématique, dans son code et dans son implémentation. Si on prend l'exemple de la simulation d'une ville, un SMA *dualiste*

[142] Voir par exemple David, D., Payet, D., Courdier, R. 2011. « Réification de zones urbaines émergentes dans un modèle simulant l'évolution de la population à La Réunion. » *Journées Francophones sur les Systèmes Multi-Agents* (JFSMA'11), oct. 2011, Valenciennes : Cépaduès, p. 63-72.

modélise à la fois le comportement local des acteurs (individus, agences immobilières, administrations, *etc.*) et le comportement global du système (par exemple son marché immobilier, sa politique d'urbanisation, ses impôts locaux, *etc.*). Par exemple, dans le cas de « SMA multi-niveaux »[143], deux niveaux de modélisation sont exécutés simultanément. Leurs états sont régulièrement confrontés et synchronisés afin de maintenir la cohérence du modèle général[144]. Par exemple, on fait en sorte que les variables globales du marché immobilier soient cohérentes avec le comportement local des acteurs et, réciproquement, on met à jour le niveau macroscopique en fonction des actions qui ont eu lieu au niveau microscopique. Dans le cas de la résolution de problèmes, le dualisme consiste à introduire dans les SMA des entités macroscopiques chargées de centraliser et d'agréger les informations produites localement par les agents. Pour les « systèmes à tableaux noirs »[145] par exemple, ces entités permettent de définir un espace partagé par tous les agents, qui peuvent ainsi enregistrer et récupérer de l'information. Elles peuvent également être utilisées pour assembler les résultats distribués et ainsi synthétiser la réponse globale du système.

Dans ces deux exemples de dualisme (simulation de systèmes et résolution de problèmes), les comportements globaux du SMA sont directement disponibles, lors de l'analyse, au niveau des entités macroscopique introduites lors de la conception. Pour R. Keith Sawyer, l'émergence apparaît lorsque les entités des deux niveaux entretiennent des rapports causaux bilatéraux, c'est-à-dire lorsque (1) le comportement des entités microscopiques influence celui des entités macroscopiques (« causalité ascendante ») et (2) les entités macroscopiques ont elles-mêmes un pouvoir causal sur le niveau microscopique (« causalité descendante » propre au dualisme). Cette notion de « causalité descendante » est souvent interprétée comme un trait essentiel des phénomènes émergents[146] : les objets macroscopiques organisent alors le comportement des individus qui les composent. Par

---

[143] Gil-Quijano, J., Hutzler, G., Louail, T. 2010. « Accroche-toi au niveau, j'enlève l'échelle. Éléments d'analyse des aspects multiniveaux dans la simulation à base d'agents. » *Revue d'Intelligence Artificielle*, vol. 24, p. 625-648.

[144] L'intérêt des « SMA multi-niveaux » réside dans l'économie réalisée (en termes de calcul) lorsqu'on simule le niveau macroscopique à la place du niveau microscopique sur certaines zones spatio-temporelles où la simulation ne requiert pas une grande précision. On se contente alors d'une « approximation » qu'il faut maintenir cohérente. Ce compromis réduit cependant l'intérêt d'une simulation orientée SMA, reposant avant tout sur les dynamiques locales des agents.

[145] « *Blackboard systems* » en anglais. Sawyer, R.K. 2001. « Simulating Emergence and Downward Causation in Small Groups. » *Multi-Agent-Based Simulation*, vol. 1979, p. 49-67.

[146] Voir à ce sujet l'analyse de la notion de « *downward causation* » dans Deguet, J., Demazeau, Y., Magnin, L. 2006. *Op. cit.*

exemple, le trafic routier influence le comportement des usagers, les marchés financiers contraignent l'activité des agents économiques, les organismes influencent le comportement de leurs organes et de leurs cellules, *etc*.

Cependant, dans le cas de SMA entièrement conçus au niveau microscopique (c'est le cas des SMADAG), une telle conceptualisation de l'émergence n'est pas opérante. En effet, le niveau macroscopique ne fait pas partie intégrante du système. Il n'est pas explicitement formalisé lors de sa conception et il n'est pas exécuté, contrairement au niveau microscopique. Les dynamiques causales des SMADAG sont donc entièrement supportées par les agents (il ne s'agit pas de SMA dualistes). En particulier, la décentralisation et l'asynchronisme du niveau microscopique rend extrêmement difficile l'intégration d'entités macroscopiques et donc la mise en place de « causalités descendantes ». Selon le dualisme, il n'y aurait donc aucune émergence possible au sein des SMADAG puisqu'ils sont « conçus uniquement au niveau des agents ». Dans un tel contexte, il faut alors privilégier une acception moniste de l'émergence, c'est-à-dire une acception qui ne présuppose pas l'existence de plusieurs niveaux de modélisation.

Le monisme en informatique

Dans le cas des approches monistes, les phénomènes émergents ne sont pas *directement* modélisés au sein du programme. Ils peuvent être anticipés lors de la conception, mais ils ne sont pas explicitement implémentés. Le programme en lui-même est uniquement constitué d'agents. Le niveau macroscopique apparaît *lors de l'exécution*, par le jeu des interactions entre les agents, mais il n'est pas lui-même *exécuté*.

Contrairement au dualisme, les phénomènes émergents ne sont donc pas réifiés : ils ne sont pas des « morceaux de programmes » à part entière et, de fait, ils n'ont pas de pouvoir causal lors de l'exécution. Ainsi, il ne peut y avoir de « causalité descendante » du niveau macroscopique sur les agents. Pour le monisme, les comportements émergents sont donc traités comme des *épiphénomènes* : ils sont causés par l'exécution, mais ils n'ont pas d'influence rétroactive. Ils sont des *aspects* de l'exécution, et non des manifestations indépendantes. Cette acception de l'émergence permet d'importer la notion d'épiphénomène de la philosophie de l'esprit[147] vers l'Intelligence Artificielle. Elle permet ainsi de rappeler aux scientifiques que certains objets qu'ils manipulent ne sont pas des entités physiques à proprement parler, mais seulement les aspects particuliers des

[147] Voir par exemple la position de Frank Jackson concernant la nature épiphénoménale des *qualia*. Jackson, F. 1982. « Epiphenomenal Qualia. » *Philosophical Quarterly*, vol. 32, p. 127-136.

objets microscopiques[148]. Les phénomènes émergents sont donc des *propriétés* du système et ils doivent être traités comme tels. En outre, cette « clôture causale » du monisme encourage une véritable démarche ascendante pour la conception des SMA : le système doit être conçu en commençant « par le bas » et les comportements émergents doivent être synthétisés à partir des niveaux inférieurs. On parle également d'approche *bottom-up*, par opposition aux approches *top-down* (comme les « SMA multi-niveaux ») qui construisent les systèmes par décomposition d'objets macroscopiques en éléments plus fins. Dans le cas d'une démarche ascendante, il ne s'agit pas seulement de *simuler* les comportements émergents à partir de modèles macroscopiques *ad hoc* (par exemple en définissant un modèle d'évolution des marchés immobiliers indépendant du modèle des acteurs économiques), mais de les *émuler* véritablement, c'est-à-dire de reproduire les dynamiques émergentes dans leurs intégralité, à partir de leurs fondements microscopiques. Pour les approches ascendantes donc, seuls les agents ont un pouvoir causal. Le niveau macroscopique n'est alors qu'un aspect particulier de l'ensemble de leurs actions.

Dans le cas de la résolution de problèmes, le monisme assure également une approche ascendante. Notamment, l'algorithme doit pouvoir être exécuté uniquement à partir d'agents décentralisés et asynchrones, comme dans le cas des SMADAG. De tels programmes ne s'appuient sur aucune entité macroscopique pour fonctionner. Contrairement aux « systèmes à tableau noir », ils n'ont pas besoin d'espaces de mémoire partagée ou d'horloges globales pour centraliser et synchroniser l'information. Ainsi, la solution engendrée est le résultat d'un véritable « calcul émergent »[149], contrairement aux méthodes descendantes qui décomposent des algorithmes centralisés en sous-tâches parallélisables. Les « systèmes auto-organisés »[150] constituent un très bon exemple de SMA monistes. Des fonctions globales y sont synthétisées à partir de fonctions locales uniquement. Dans le cas des SMADAG, du fait de leur composition microscopique, il faut également préférer une acception moniste de l'émergence, présupposant que *la conception du système est uniquement microscopique*. Les phénomènes émergents sont alors traités comme des

[148] Les spécialistes ont notamment tendance à multiplier les morceaux de code pour modéliser des objets de natures très différentes (objets physiques, abstractions, groupes d'objets, propriétés des objets, *etc*.). Le monisme invite au contraire à la parcimonie lors de l'implémentation, et à bien distinguer ce qui doit et ce qui ne doit pas être programmé explicitement.

[149] Forrest, S. 1990. « Emergent computation: Self-organizing Collective and Cooperative Phenomena in Natural and Artificial Computing Networks. » *Physica D: Nonlinear Phenomena*, vol. 42, n°1-3, p. 1-11.

[150] Voir par exemple l'état de l'art réalisé par Gauthier Picard dans sa thèse sur les « SMA adaptatifs ». Picard, G. 2004. *Op. cit.*

épiphénomènes. Ils doivent êtres identifiés du côté de l'analyse du système, et non du côté de sa conception.

L'éliminativisme en informatique

La version informatique du monisme ne doit pas succomber aux travers éliminativistes de son analogue philosophique. Sous prétexte d'une conception purement microscopique des SMA, il ne faut pas limiter l'analyse de leurs exécutions au niveau des agents (tout comme en science, il ne faut pas limiter l'analyse des phénomènes biologiques à leurs descriptions physiques). Certaines approches fondent en effet l'émergence sur une description complète des dynamiques microscopiques. C'est par exemple le cas de la conceptualisation de Darley et de celle de Bedau qui identifient chacun l'émergence à une propriété *computationnelle* du système, c'est-à-dire une propriété du calcul effectué au niveau microscopique[151]. L'objectif de telles approches est de définir une propriété globale du système (donc une propriété émergente) indépendante du modèle d'observation utilisé. Il faut pour cela se limiter à ce qui existe effectivement dans le système, c'est-à-dire le niveau microscopique. Les propriétés macroscopiques sont alors réductibles aux descriptions microscopiques et, par suite, elles ne fournissent pas d'informations supplémentaires concernant le système. Le monisme éliminativiste en informatique, en voulant définir l'émergence de manière *objective*, amène à privilégier un seul mode d'analyse des SMA, fondé sur le niveau microscopique, tout comme le monisme philosophique qui, en voulant unifier les lois de la nature, conduit à l'élimination des sciences spéciales.

Dans le cas de SMA complexes et de très grandes tailles, la première difficulté réside dans le calcul de telles propriétés objectives. En effet, le traitement de la description microscopique est très coûteux. Les phénomènes émergents sont par suite difficiles à détecter *en pratique*. De plus, certaines acceptions éliminativistes de l'émergence sont *indécidables*, c'est-à-dire qu'il n'existe

[151] Dans ces travaux l'émergence est associée à la complexité de l'exécution. Il y a « émergence » lorsque les états futurs du système ne peuvent être prédits autrement qu'en faisant une simulation complète. Il n'y a alors aucun « raccourci computationnel » pour prédire la dynamique du système. Darley, V. 1994. « Emergent Phenomena and Complexity. » *Artificial Life*, vol. 4, p. 411-416 ; Bedau, M.A. 1997. *Op. cit.* Cette acception de l'émergence est basée sur la notion de « complexité algorithmique » développée par Kolmogorov, A.N. 1965. « Three Approaches to the Quantitative Definition of Information. » *Problems Information Transmission*, vol. 1, n°1, p. 1-7.

aucun algorithme générique qui prend un système en entrée et qui détermine en sortie si celui-ci « est » ou « n'est pas » émergent[152]. De telles propriétés sont donc inutilisables en pratique.

Par ailleurs, nous soutenons qu'une caractéristique intéressante des phénomènes émergents réside dans les rapports qu'ils entretiennent avec des modes de connaissance particuliers. De la même manière que l'« émergence épistémique » en philosophie réside « dans l'œil du scientifique », les phénomènes émergents en informatique devraient être définis relativement à une méthode d'observation donnée. Cette acception *subjectiviste* de l'émergence invite à apposer au système un « modèle de l'observateur » responsable de l'élaboration du niveau macroscopique. Il y a alors autant de descriptions macroscopiques qu'il y a d'observateurs différents. La démarche sous-jacente est la suivante : même si un SMA peut *en principe* être entièrement analysé à partir de sa description microscopique, sa compréhension *en pratique* – et particulièrement dans le cas des SMADAG – nécessite un processus d'abstraction. Ce processus n'est pas neutre. Il doit donc être pris en compte dans l'édification de la sémantique macroscopique. C'est ce que nous appelons un *monisme non-éliminativiste*.

Le non-éliminativisme en informatique

Nous définissons alors une contrainte méthodologique complétant le monisme : le non-éliminativisme. Il garantit que *l'analyse du système n'est pas uniquement microscopique*. De nombreuses acceptions subjectivistes de l'émergence, défendant des conceptions non-éliminativistes, peuvent être trouvées dans la littérature spécialisée de l'Intelligence Artificielle. L'émergence peut ainsi être définie en fonction d'appareils de détection hiérarchisés[153], à travers la notion de « surprise » entre la conception du système et son observation[154], en fonction de grammaires de modèles qui permettent de décrire l'exécution microscopique de différentes manières[155] ou selon des agrégations réalisées à partir de sondes d'observation macroscopique[156]. Ces approches font toutes intervenir des instruments d'observation ou de description pour définir la notion d'émergence à partir du niveau microscopique. Elles sont à ce titre *non-éliminativistes*.

---

[152] C'est le cas de l'émergence computationnelle de Darley, V. 1994. *Op. cit.* et de celle de Bedau, M.A. 1997. *Op. cit.*

[153] Bonabeau, É., Dessalles, J.-L. 1997. « Detection and Emergence. » *Intellectica*, vol. 25, n°2, p. 85-94.

[154] Ronald, E.M.A., Sipper, M. 2001. « Surprise versus unsurprise: Implications of emergence in robotics. » *Robotics and Autonomous Systems*, vol. 37, n°1, p. 19-24.

[155] Kubík, A. 2003. « Toward a Formalization of Emergence. » *Artificial Life*, vol. 9, p. 41-65.

[156] Lamarche-Perrin, R., Demazeau, Y., Vincent, J.-M. 2011. *Op. cit.*

		Analyse	
		Éliminativisme Un seul niveau de description	**Non-éliminativisme** Plusieurs niveaux de description
Conception	**Dualisme** Plusieurs niveaux de modélisation		**SMA multi-niveaux** Gil-Quijano, J., *et al.* 2010. **Systèmes à « tableau noir »** Sawyer, R.K. 2001. Dualisme non-éliminativiste
	Monisme Un seul niveau de modélisation	**Émergence « computationnelle »** Darley, V. 1994. Bedau, M.A. 1997. Monisme éliminativiste	**Détection hiérarchique** Bonabeau, É., Dessalles, J.-L. 1997. **Grammaires formelles** Kubík, A. 2003. **Observation macroscopique** Lamarche-Perrin, R., *et al.* 2011. Monisme non-éliminativiste

Figure 5 : représentation schématique de travaux en Intelligence Artificielle

En rappelant que l'émergence est avant tout « dans l'œil de l'observateur », le non-éliminativisme permet également de fonder une approche *pragmatiste* des phénomènes émergents. En effet, contrairement aux approches éliminativistes, le niveau macroscopique ne peut être réduit en pratique aux descriptions microscopiques. Il n'est pas une propriété du système *per se* (ontologie), mais engage un mode de description subjective (épistémie). Par conséquent, l'activité du scientifique ne repose pas sur l'observation de phénomènes émergents préexistants, mais sur un processus créatif d'abstraction. Le scientifique doit *faire-émerger*[157] les phénomènes macroscopiques utiles à la compréhension globale du système. La pertinence de telles abstractions doit toujours être évaluée *en contexte*, c'est-à-dire en fonction du SMA étudié et des objectifs scientifiques de l'analyse. En d'autres termes, il n'y a pas de « bons phénomènes émergents » *per se*, mais seulement relativement à ce que l'on veut en faire. Le non-éliminativisme introduit donc une pensée

[157] Terme emprunté à Varela, F.J., Thompson, E.T., Rosch, E. 1991. *Op. cit.*, p. 210.

pragmatiste en Intelligence Artificielle[158], selon laquelle le spécialiste a la responsabilité de créer les phénomènes et les abstractions qui seront utiles à la conception et à l'analyse de ses programmes.

Dans le cas de la simulation de systèmes complexes, le non-éliminativisme permet de confronter et de choisir des abstractions pour la compréhension du système. Par exemple, les notions d'« organes » et de « marchés financiers » sont-elles pertinentes pour décrire respectivement le fonctionnement d'un organisme et celui d'un système économique ? Sont-elles les seules abstractions viables ? Quelles sont les autres niveaux macroscopiques que l'on peut construire à partir des niveaux microscopiques ? Ces problématiques amènent également à s'interroger sur le rôle de l'observateur et de l'analyste dans la composition même du système. Les phénomènes émergents témoignent en effet de la présence d'un sujet connaissant, d'un scientifique. Le système microscopique est alors *augmenté* d'un niveau macroscopique qui lui donne une valeur épistémique nouvelle.

Section 3.2.4. Bilan de la collaboration

Deux contraintes méthodologiques ont été exprimées dans la section précédente. Le *monisme* garantit que *la conception du système est uniquement microscopique* et le *non-éliminativisme* garantit que *l'analyse du système n'est pas uniquement microscopique*. La conjonction de ces deux contraintes détermine une position cohérente avec un monisme non-éliminatif. Selon l'analogie présentée en début de section, il s'agit donc de positions *émergentistes*. Mais quel est le rapport entre cet émergentisme informatique et l'émergentisme défendu par la philosophie britannique au XIX^{e} siècle ? La nature des phénomènes complexes y est abordée de la même manière. Dans le cas des SMA, les systèmes étudiés n'ont qu'un seul niveau de réalité, le niveau des agents (monisme). Donc *en principe*, ils peuvent être entièrement décrits et expliqués à ce niveau-là, tout comme les êtres vivants peuvent *en principe* être expliqués par les lois de la physique. Mais *en pratique*, il est nécessaire d'utiliser des abstractions de « haut-niveau » pour aborder la complexité des phénomènes (non-éliminativisme). En philosophie, on parle de sciences spéciales. En Intelligence Artificielle, il s'agit de développer des modes d'analyse adaptés aux systèmes que l'on étudie.

[158] Gertrudis van de Vijver souligne l'importance d'une telle démarche pragmatiste lorsqu'il s'agit de penser l'émergence en Intelligence Artificielle. van de Vijver, G. 1997. « Émergence et explication. » *Intellectica*, vol. 25, n°2, p. 7-23.

Bilan : « IA forte → IA faible »

Les deux contraintes garantissent donc une démarche en accord avec l'« émergentisme épistémique » présenté en section 3.2.2. En outre, elles permettent d'importer des concepts chers à la philosophie dans le domaine technique des SMA :

- Principe de parcimonie : À capacités égales, on doit préférer un modèle simple à un modèle compliqué. En particulier, il est préférable de simuler un système à partir d'un seul niveau de modélisation.

- Épiphénoménisme : Les phénomènes émergents ne sont pas des entités indépendantes, mais des aspects particuliers du niveau microscopique.

- Approche ascendante : Les phénomènes émergents sont produits par le niveau microscopique. Méthodologiquement, il faut donc partir de ce niveau et non l'inverse.

- Subjectivisme : Les phénomènes émergents dépendent d'un mode de connaissance donné. Ils sont « dans l'œil de l'observateur. »

- Pragmatisme : Les phénomènes émergents n'existent pas en soi. Ils sont des abstractions créées par le scientifique et doivent donc être évalués en fonction du contexte général de l'analyse.

Ces exigences ont des répercussions techniques sur la simulation des phénomènes. Les modèles choisis (monistes ou dualistes), les outils d'analyse (éliminativistes ou non-éliminativistes) et la méthode générale (*bottom-up* ou *top-down*, décentralisée ou asynchrone, *etc.*) dépend en effet de la conception sous-jacente de l'émergence. Dans de nombreux cas, cette position est implicite. Mais une analyse philosophique préliminaire permet d'expliciter les principes qui guident les différentes démarches scientifiques de l'Intelligence Artificielle. La philosophie permet alors de catégoriser les approches et les systèmes en fonction de leurs acceptions de l'émergence (*cf.* figure 5). Elle éclaircit les positions de chacun et permet de déterminer quels travaux satisfont nos critères. Dans le cas des SMADAG, c'est la démarche que nous adoptons[159]. La conceptualisation retenue alors est celle de Bonabeau et Dessalles[160] qui satisfait nos deux critères : le monisme et le non-éliminativisme.

Nous défendons donc ici la relation « IA forte → IA faible ». En choisissant une position philosophique adéquate (IA forte, ici empruntée à la métaphysique), nous mettons toutes les

[159] Lamarche-Perrin, R., Demazeau, Y., Vincent, J.-M. 2011. *Op. cit.*

[160] Bonabeau, É., Dessalles, J.-L. 1997. *Op. cit.*

chances de notre côté pour produire les comportements qui nous intéressent (IA faible, ici liée à la simulation de systèmes complexes et à la résolution de problèmes distribués). Ainsi, à la façon de la stratégie collaborative présentée dans le chapitre 3.1 pour l'édification d'une « nouvelle IA », nous pensons qu'un bon socle philosophique permet à l'Intelligence Artificielle d'avancer sur le plan technique.

Bilan : « non IA forte → IA faible peu probable »

Notons cependant que ces critères méthodologiques n'imposent aucune contrainte *de principe*. En effet, il est possible de simuler un système de manière dualiste, même si cela est moins pertinent pour la compréhension des phénomènes macroscopiques (on utilise alors des modèles *ad hoc* des phénomènes émergents). On peut également analyser l'exécution d'un système complexe à partir de son niveau microscopique, même si *en pratique* cette procédure est extrêmement coûteuse. Ces approches sont donc plus ou moins bien adaptées aux problèmes rencontrés par l'IA faible, mais elles ne sont jamais impossibles. Dans le cas des SMADAG par exemple, le monisme et le non-éliminativisme garantissent une conception de l'émergence efficace *en pratique* : elle s'applique aux systèmes décentralisés et asynchrones (monisme) et elle permet un traitement rapide des grandes quantités de données grâce à des abstractions de « haut-niveau » (non-éliminativisme).

Nous avons vu que Dreyfus ne s'oppose pas de manière systématique au computationnalisme puisqu'il pense que celui-ci peut réussir localement, notamment sur certains « micro-mondes » (*cf.* section 2.4.1). En revanche, il affirme que les programmes computationnalistes auront énormément de mal à résoudre des problèmes moins contrôlés et plus complexes, même si cela n'est pas complétement impossible (*cf.* section 2.4.2). De la même manière, nous défendons la relation « non IA forte → IA faible peu probable » dans le cas des phénomènes émergents. Ce n'est pas la réciproque stricte de la relation « IA forte → IA faible ». Il s'agit de dire qu'une conceptualisation erronée des phénomènes émergents entraîne de nombreuses difficultés pratiques concernant leur simulation. Comme dans le cas du computationnalisme, la philosophie peut donc identifier les erreurs des présupposés implicites aux méthodes de l'IA faible et ainsi expliquer ses difficultés pratiques.

Partie 4. L'Intelligence Artificielle au service de la philosophie

Dans cette partie, nous revenons sur le renversement entrevu par Andler dans l'ouvrage de Dreyfus et faisant de l'Intelligence Artificielle une « science de la philosophie » (*cf.* chapitre 1.1). Dans la partie précédente, nous avons exposé des exemples de collaboration au sein desquels la philosophie a un rôle fondationnel. Elle aide notamment à définir un socle conceptuel pertinent pour les recherches en Intelligence Artificielle. À ce titre, elle tient dans ces travaux un rôle parfois réservé à la « philosophie des sciences » et concernant l'évaluation de concepts scientifiques. Le renversement consiste donc à faire de l'Intelligence Artificielle une « science de la philosophie », c'est-à-dire une discipline chargée à son tour d'évaluer la pertinence des concepts philosophiques ou de fournir un socle aux recherches en philosophie. Cette partie expose des exemples de collaborations où l'Intelligence Artificielle, de cette manière, se met au service de la philosophie.

Dans le chapitre 4.1, la contraposée « non IA faible → non IA forte » de la relation utilisée dans la partie précédente est exploitée pour expliquer en quoi l'Intelligence Artificielle peut servir de « banc d'essai » à la philosophique. Elle permet notamment de confronter et de falsifier ses théories. Nous présentons des auteurs qui ont pressenti cette forme de collaboration et qui invitent à s'y engager. Le chapitre 4.2 se penche sur le travail de Timothy van Gelder. Sa démarche est interprétée comme une collaboration *positive* dans la mesure où l'Intelligence Artificielle n'a plus un rôle de falsification, mais un rôle d'émulation. Les machines concrètes de l'IA faible permettent notamment de formuler des hypothèses nouvelles, de révéler des pistes inexplorées, bref, d'insuffler des idées à la philosophie. Cette démarche repose sur la relation « IA faible → IA forte probable » notamment défendue par Hector J. Levesque (*cf.* section 2.2.3). Dans le chapitre 4.3, enfin, nous élaborons une démarche pour évaluer des théories philosophiques à partir de machines qui contredisent les deux relations « IA faible → IA forte » et « non IA faible → non IA forte ». Les conséquences d'un tel conflit intéressent alors la philosophie. La démarche est appliquée à la « chambre chinoise » de Searle et aux « chatons aveugles » de Held & Hein.

Chapitre 4.1. L'IA comme « science de la philosophie »

Dans la première partie de *What Computer's Can't Do*, Hubert L. Dreyfus constate l'échec du GPS et des autres tentatives d'application du computationnalisme. Nous avons interprété ce résultat comme une limite de l'Intelligence Artificielle en ce qui concerne la simulation de l'intelligence à partir de systèmes symboliques physiques (« non IA faible (SSP) », *cf.* section 2.4.1). Mais ce constat ne présente pas seulement un échec de l'IA faible : il témoigne également d'un échec de l'IA forte. En effet, si les nombreuses tentatives du computationnalisme n'aboutissent pas, peut-être qu'il ne s'agit pas seulement de difficultés techniques relatives à la formalisation, à l'implémentation et à l'exécution des programmes (IA faible). Peut-être qu'il s'agit, plus profondément, d'une erreur du paradigme scientifique dans lequel il s'inscrit, c'est-à-dire une erreur du computationnalisme lui-même ou de son socle philosophique (IA forte).

Les échecs des programmes computationnalistes révèlent ainsi les difficultés de la tradition cartésienne. Sans affirmer que l'échec d'un programme tel que le GPS permet à lui seul de falsifier une théorie philosophique, il constitue néanmoins un argument *contre* cette théorie. Ce raisonnement exploite la relation « non IA faible → non IA forte » ou, de manière moins radicale, « non IA faible → IA forte peu probable » : si un programme peine à simuler l'intelligence, et si celui-ci a été réalisé correctement, c'est que son modèle cognitif est inadéquat. En d'autres termes, il ne modélise pas de manière satisfaisante ce qu'est *réellement* la cognition. Remarquons qu'il s'agit de la contraposée de la relation « IA forte → IA faible » que nous avons exploitée dans la partie 3. En effet, si on est prêt à dire qu'un modèle adéquat de la cognition permet, lorsqu'il est bien implémenté, de simuler l'intelligence (« IA forte → IA faible »), alors il faut être prêt à dire qu'un programme bien implémenté, qui ne parvient à simuler l'intelligence, ne dispose pas d'un tel modèle adéquat (« non IA faible → non IA forte »). Avec ce raisonnement, les arguments contre les théories de l'esprit ne sont plus seulement *philosophiques*, ils sont également *expérimentaux* : « nous avons appliqué à de multiples reprises les principes de telle théorie de la cognition, et ceux-là n'ont pas fonctionné en pratique ! » Bien évidemment, l'Intelligence Artificielle n'est pas la première science à apporter des arguments empiriques à la philosophie de l'esprit. Par exemple, la psychologie, les neurosciences et la linguistique participent également aux débats concernant la nature de la cognition. Mais ces disciplines ont moins de « possibilités expérimentales » dans la mesure où leurs expériences portent sur des objets préexistants, alors que l'Intelligence Artificielle construit elle-même ses objets : les machines *sont* des expérimentations. C'est pourquoi nous parlons ici d'arguments *expérimentaux* et

non pas simplement d'arguments *empiriques*[161]. On retrouve l'idée que l'Intelligence Artificielle est un « banc d'essai »[162] pour la philosophie qui devient, du même coup, une *philosophie expérimentale*. Inman Harvey donne plus particulièrement cette fonction expérimentale à la robotique :

> Lorsqu'on construit un robot situé, notre position philosophique affecte les choix de conception. Elle est ainsi testée dans le monde réel – « faire de la philosophie de l'esprit avec un tournevis ».[163]

Cette démarche repose d'abord sur l'idée que « la philosophie induit une différence en pratique »[164] et qu'ainsi deux théories philosophiques différentes n'engendrent pas les mêmes programmes. On peut alors évaluer les théories à partir de leurs applications.

À l'origine, la démarche de Dreyfus n'est pas une démarche de falsification des théories philosophiques. En effet, nous avons vu dans la section 2.4.2 que Dreyfus explique les échecs du computationnalisme *à partir* de son socle philosophique : « non IA forte → IA faible peu probable ». Les théories sont alors évaluées en amont de l'expérimentation et elles servent à prédire ou à expliquer les difficultés pratiques de l'Intelligence Artificielle. La contraposée « IA faible → IA forte probable » consiste à dire que la réussite pratique d'une machine rend crédible le modèle de la cognition à son origine. Il s'agit d'une évaluation *positive* dans le sens où les arguments expérimentaux permettent de conforter une théorie, plutôt que de la réfuter dans le cas de la falsification « non IA faible → IA forte peu probable » qui constitue donc une évaluation *négative*. Une démarche d'évaluation *positive* est également développée dans le chapitre suivant à partir du

---

[161] Les sciences cognitives ont néanmoins des possibilités expérimentales très développées. L'expérience des « chatons aveugles », présenté un peu plus loin, en est un bon exemple. Seulement, l'Intelligence Artificielle procède « à partir de rien », sur des objets entièrement artificiels. C'est ce qui fonde à la fois sa liberté d'expérimentation et ses difficultés techniques.

[162] *Cf.* l'avant-propos de Dreyfus, H.L. 1979. *Intelligence Artificielle : mythes et limites*. [*What Computers Can't Do: The Limits of Artificial Intelligence*, 2nd ed.] Vassallo-Villaneau, R.-M. (trad.), Andler, D. (pref.), Perriault, J. (pref.). Paris : Flammarion, 1984, p. XIV.

[163] « In a project of building situated robots one's philosophical position affects design decisions and is then tested in the real world – "doing philosophy of mind with a screwdriver". » [Notre traduction] Harvey, I. 2000. « Robotics: Philosophy of Mind Using a Screwdriver. » *Evolutionary Robotics: From Intelligent Robots to Artificial Life*, vol. III, p. 208.

[164] « Philosophy does make a practical difference. » [Notre traduction] *Ibid.*, p. 228.

travail de Timothy van Gelder. Mais en ce qui concerne la falsification, certains passages de l'ouvrage de Dreyfus tendent à diriger sa critique dans ce sens :

> Si […] l'intelligence artificielle se révèle une chimère impossible à atteindre, alors nous devrons chercher à distinguer la manière dont raisonnent les machines de celle dont nous raisonnons, nous. Et cela aussi modifiera radicalement notre perception de nous-même. Par conséquent, l'heure a donc sonné ou bien d'accorder à la tradition philosophique le bien-fondé de son intuition centrale, ou bien d'abandonner cette conception de la nature humaine comme une mécanique, attitude qui s'est progressivement répandue en Occident au cours des vingt derniers siècles.[165]

Pour Dreyfus donc, un échec généralisé de l'Intelligence Artificielle aurait un impact philosophique sans précédent. Il nous amènerait notamment à abandonner la conception classique de la nature humaine, vieille de 2000 ans. De cette manière, les machines de l'Intelligence Artificielle entrent dans le jeu des controverses philosophiques. L'échec du computationnalisme permet notamment de systématiser la critique de la tradition cartésienne. Enfin, on peut imaginer que Dreyfus s'intéresse aux résultats pratiques de l'Intelligence Artificielle pour ajouter de nouveaux arguments à sa critique véhémente du cartésianisme.

[165] Dreyfus, H.L. 1979. *Op. cit.*, p. 18.

Chapitre 4.2. Les machines de van Gelder

Ce chapitre se penche sur l'analyse par Timothy van Gelder de dispositifs de régulation pour machines à vapeur et sur l'intérêt d'un tel travail pour la philosophie de l'esprit. Notre analyse repose sur la version remaniée de son article de 1995, malicieusement intitulé « What might Cognition be if not Computation? »[166]

Section 4.2.1. Contre le computationnalisme

La première phrase de l'article précise sans détour le domaine de recherche de van Gelder. « Qu'est-ce que la cognition ? »[167] Il s'agit de définir la nature de la faculté, ses propriétés intrinsèques et donc de répondre à la question fondamentale de la « philosophie de la cognition », comme il la nomme lui-même. Ainsi, le travail de van Gelder semble vouloir participer avant tout à l'édification d'un modèle viable de la cognition et, en ce sens, il peut être utile à la résolution du problème de l'IA forte. Sur ce point, van Gelder poursuit notamment l'analyse critique de Dreyfus en condamnant la mainmise du paradigme computationnaliste sur les sciences cognitives (*cf.* section 2.4.3). L'objectif principal de son article consiste à affirmer qu'il existe des alternatives conceptuelles à la théorie computo-représentationnelle et de réfuter l'argument du « *de-quoi-d'autre-pourrait-il-s'agir ?* », avancé notamment par les fervents défenseurs du computationnalisme Jerry A. Fodor et Allan Newell[168]. La position de van Gelder est à ce titre profondément antireprésentationnaliste (l'esprit se passe de représentations symboliques) et il milite pour un *modèle dynamique* de la cognition (que nous exposerons dans le chapitre suivant). Une fois que ce modèle alternatif est accepté comme hypothèse de travail, une démarche empirique de validation peut être mise en place. Mais l'objectif premier de l'article de van Gelder se limite à l'acceptation d'une telle hypothèse face à l'emprise du computationnalisme.

Comme Dreyfus, van Gelder explique que la domination exercée par le computationnalisme provient de son socle philosophique. La tradition cartésienne, à laquelle souscrivent implicitement de nombreux scientifiques, est en effet un attrait majeur du modèle computo-représentationnel. Celui-

---

[166] van Gelder, T. 1998. « Dynamique et cognition. » [« Dynamics and Cognition. »] Lapointe, S. (trad.). *In* Fisette, D. (éd.), Poirier, P. (éd.). 2003. *Philosophie de l'esprit : problèmes et perspectives*. Paris : Vrin, p. 329-369, version remaniée de van Gelder, T. 1995. « What might Cognition be if not Computation? » *Journal of Philosophy*, vol. 92, n°7, p. 345-381.

[167] van Gelder, T. 1998. *Op. cit.*, p. 329.

[168] *Ibid.*, p. 330 et introduction de Fisette, D. (éd.), Poirier, P. (éd.). 2003. *Philosophie de l'esprit : problèmes et perspectives*. Paris : Vrin, p. 39. *Cf.* également la section 2.3.1 concernant la position computationnaliste de Fodor et de Newell.

ci hérite de principes philosophiques qui nous sont chers (*cf.* section 2.4.2). C'est donc en suivant implicitement la relation « IA forte → IA faible », selon laquelle l'implémentation d'un modèle correct de la cognition permet au projet de l'IA faible d'aboutir, que philosophes et spécialistes de l'IA ont fondé le computationnalisme. Cependant, les erreurs de la tradition cartésienne (IA forte) se sont répercutées sur ses applications (IA faible). Puisque les prémisses sont mises en questions[169], il faut envisager d'autres modèles de la cognition et d'autres bases philosophiques. Van Gelder prend ainsi note de l'explosion paradigmatique des sciences cognitives qui a suivi la remise en cause du computationnalisme, engendrant de nouvelles approches telles que « le connexionnisme, les approches neurocomputationnelles, la psychologie écologique, la robotique située et la vie artificielle. »[170]

Section 4.2.2. Techniques de régulation et philosophie de la cognition

Régulateurs computationnels et régulateurs centrifuges

La stratégie de van Gelder est pour le moins étonnante. Cinq pages de son article, initialement consacré aux modèles de la cognition, sont réservées à un problème technique qui apparaît lors de la révolution industrielle[171]. Dans la seconde moitié du XVIII^{e} siècle, l'industrie textile voulait bénéficier des progrès récents concernant la conception des machines à vapeur. Celles-ci pouvaient efficacement remplacer les sources d'énergies alors utilisées pour les machines à tisser, telles que les chevaux et les moulins à eau. Cependant, le tissage nécessite une source d'énergie hautement uniforme. Les à-coups et les variations de vitesse, engendrés pas de brusques changements de pression dans les chaudières, devaient donc être contrôlés afin de garantir un tissage de qualité. La mise en place d'un système de régulation automatique, permettant de maintenir constante la vitesse de rotation de la machine, constituait donc un véritable défi technique pour l'ingénierie du XVIII^{e} siècle.

---

[169] Van Gelder cite notamment les travaux critiques de Ryle et de Heidegger, ainsi que ceux de Dreyfus. Ryle, G. 1949. *La Notion d'esprit*. [*The Concept of Mind*.] Stern-Gillet, S. (trad.), Tanney, J. (pref.). Paris : Payot & Rivages, 2005 ; Dreyfus, H.L. 1979. *Op. cit.*

[170] van Gelder, T. 1998. *Op. cit.*, p. 329.

[171] *Ibid.*, « 1. Le problème du régulateur », p. 331-335.

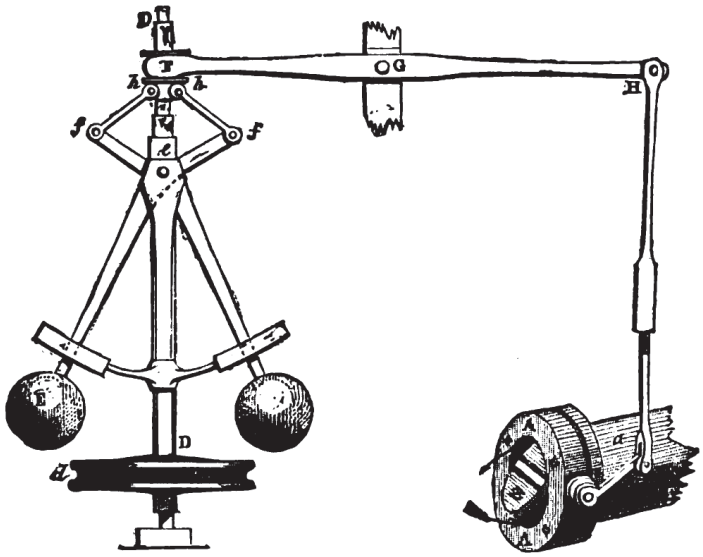

Figure 6 : le régulateur centrifuge de Watt.
van Gelder, T. 1998. *Op. cit.*, p. 335.

Van Gelder présente deux solutions à ce problème de régulation. La première, élaborée par James Watt, consiste en un appareillage mécanique utilisant la force centrifuge issue de la rotation du volant principal de la machine à vapeur pour ajuster la hauteur de bras pivotants. Ces bras sont à leur tour reliés à une soupape d'étranglement permettant de réguler la quantité de vapeur transmise au piston (*cf.* figure 6 ci-dessus). Ainsi, lorsque la vitesse de la machine augmente, la rotation du volant entraine les bras vers le haut, réduisant ainsi l'ouverture de la valve. Au contraire, lorsque la force centrifuge diminue, les bras, en se relâchant, libère la vapeur et relancent la machine. Van Gelder nomme ce dispositif « régulateur centrifuge ». La seconde solution, que nous utiliserions sûrement si le problème se présentait à nous aujourd'hui, consiste à décomposer la tâche à effectuer en sous-tâches plus simples : mesure de la vitesse de rotation du volant, mesure de l'écart avec la vitesse exigée pour le tissage, calcul du changement de pression nécessaire au changement de vitesse et ajustement de la soupape d'étranglement en fonction. Ces diverses opérations peuvent être réalisées par des appareils électroniques classiques adéquatement reliés entre eux : tachymètre (appareil de mesure), ordinateur (appareil de calcul) et bras mécanique (appareil effecteur). Van Gelder nomme ce type de dispositif « régulateur computationnel », dans la mesure où il réalise un calcul pour configurer une sortie en fonction d'une entrée.

Retour sur la philosophie de la cognition

Le régulateur centrifuge de Watt et le régulateur computationnel du XX^{e} siècle constituent deux objets techniques intéressants. « Mais quel intérêt cela peut-il bien avoir pour la philosophie

des sciences cognitives ? »[172] Van Gelder, après avoir observé des différences essentielles dans la conception des deux régulateurs et après avoir présenté les cadres conceptuels généralement utilisés pour rendre compte de leurs fonctionnements respectifs, revient sur sa question initiale : « qu'est-ce que la cognition ? » À la lumière de ce qui vient d'être exposé, une alternative apparaît clairement : la cognition fonctionne-t-elle plutôt comme un régulateur centrifuge ou comme un régulateur computationnel ?

Selon le computationnalisme, la cognition est semblable au second type de régulateur dans la mesure où celui-ci procède par un calcul. En vérité, il partage bien d'autres points communs avec le modèle computationnel de la cognition : il utilise des représentations concernant l'état de son environnement (vitesse du volant, pression dans la chaudière, *etc.*), il réalise une computation à partir de ces représentations, il obéit à des règles logiques séquentielles, *etc.* Au contraire, selon van Gelder, le régulateur centrifuge ne fait intervenir aucune sorte de représentation. Il est uniquement régit par les lois de la mécaniques et fait intervenir des « variables couplées »[173] (vitesse de rotation du volant, hauteur des bras emportés par la force centrifuge, position de la soupape). Celles-ci sont liées par un processus physique immédiat, continu, plutôt que par un calcul séquentiel, par étapes. Ces propriétés, qui sont fondamentales à la notion de *systèmes dynamiques*, diffèrent en tout point du modèle orthodoxe computationnaliste. Rien à voir donc *a priori* avec la cognition. Cependant, comme le défend van Gelder, « il se peut que les systèmes cognitifs soient des systèmes dynamiques. »[174]

L'intérêt pour la philosophie de la cognition d'un tel détour par la révolution industrielle réside donc dans l'introduction d'une hypothèse alternative. En partant d'un problème technique bien plus simple que la question initiale concernant la nature de la cognition, van Gelder établit une nouvelle hypothèse de travail. (1) Nous avons l'habitude de considérer que le cerveau fonctionne comme un système computationnel. (2) Pourtant, il existe des systèmes, appelés « systèmes dynamiques », qui peuvent accomplir les mêmes choses que les systèmes computationnels, mais qui diffèrent sur bon nombre de qualités essentielles. (3) Pourquoi les cerveaux ne pourraient-ils pas appartenir à cette seconde catégorie de systèmes ?

---

[172] *Ibid.*, p. 335.

[173] Voir *Ibid.*, p. 337 concernant la notion de « variable couplée », fondamentale à la compréhension des systèmes dynamiques.

[174] *Ibid.*, p. 334.

Section 4.2.3. Bilan de la collaboration

Bilan : « IA faible → IA forte possible »

L'hypothèse des systèmes dynamiques est la première position profondément antireprésentationnaliste depuis le béhaviourisme, lequel était justement rejeté par le computationnalisme[175]. Pourtant, celle-ci ne doit pas être comprise comme une nouvelle forme de béhaviourisme. En effet, van Gelder, lorsqu'il prétend que la cognition n'utilise pas de représentations, défend néanmoins une théorie sophistiquée concernant ses qualités internes. Il ne s'agit pas de dire que de telles qualités « psychologiques » sont inaccessibles à l'examen scientifique (béhaviourisme méthodologique, *cf.* section 2.1.2) et encore moins d'affirmer qu'elles n'ont pas de signification en dehors des comportements engendrés (béhaviourisme logique, *cf.* section 2.2.1). En particulier, le fait que les deux régulateurs ont des comportements similaires n'implique pas qu'ils doivent être considérés de manière identique. De plus, ce n'est pas parce que le dispositif computationnel parvient à résoudre le défi technique de la régulation (en ce sens il réussit à « simuler l'intelligence » : « IA faible (régulateur computationnel) »), qu'il constitue la meilleure façon de procéder. Dans le cas de la philosophie de la cognition, cela signifie que la réussite du régulateur computationnel *n'implique pas* qu'il soit un bon modèle de la cognition (« IA forte (régulateur computationnel) »). On a donc « non (IA faible → IA forte) ». Le béhaviourisme et les stratégies présentées en chapitre 2.2 sont ainsi rejetés par van Gelder.

Pourtant, l'examen de régulateurs permet de réfléchir sur la nature de la cognition ou, au moins, d'envisager de nouvelles possibilités. Le fait qu'un système dynamique parvienne également à résoudre le problème de la régulation (« IA faible (régulateur centrifuge) ») amène à s'interroger sur son fonctionnement. En philosophie de la cognition, on se demande alors s'il peut constituer un bon modèle pour les systèmes cognitifs (« IA forte (régulateur centrifuge) »). On a donc « IA faible → IA forte possible », le comportement intelligent fait naître l'hypothèse d'un fonctionnement « véritablement intelligent ». C'est ainsi que van Gelder oppose à l'argument du « *de-quoi-d'autre-pourrait-il-s'agir ?* » une hypothèse alternative issue de l'examen *en pratique* d'un système physique. Par la suite, la validation ou la réfutation de cette hypothèse n'est plus du ressort de l'ingénierie des systèmes de régulation. Les critères finaux, permettant de choisir parmi les deux hypothèses, dépendent du travail empirique des sciences cognitives. L'IA faible est ici un simple moteur pour la philosophie de la cognition et non une source d'arguments.

---

[175] Fisette, D. (éd.), Poirier, P. (éd.). 2003. *Op. cit.*, p. 39. *Cf.* également section 2.3.1 sur le computationnalisme et le rejet du béhaviourisme pour fonder les sciences cognitives.

Cette stratégie peut être rapprochée de celle d'Hector J. Levesque consistant à dire qu'*en pratique* une machine qui parvient à résoudre un problème donné a toujours un fonctionnement à la hauteur de la complexité de ce problème. Autrement dit, elle implémente nécessairement une méthode intelligente pour résoudre le problème en question. Il s'agit de *la nécessité en pratique de l'IA forte*, qui découle de la relation « IA faible → IA forte » (*cf.* section 2.2.3). Dans le cas de van Gelder, il s'agit d'une version plus faible de la relation : si une machine parvient *en pratique* à résoudre un problème complexe, alors son fonctionnement mérite d'être examiné, même s'il ne correspond pas obligatoirement à la façon optimale de procéder. De manière générale, la relation « IA faible → IA forte possible » amène à s'intéresser aux dispositifs engendrant des comportements complexes, dans le but de faire progresser l'IA forte. L'élégance de la solution technique et son adéquation avec le problème présagent un modèle fécond pour l'IA forte, sans constituer pour autant un argument définitif.

L'IA au service de la philosophie de l'esprit

Avec la stratégie de van Gelder, l'Intelligence Artificielle sort du domaine restreint de l'informatique pour rejoindre le champ plus large de l'ingénierie des systèmes. En effet, l'examen des dispositifs techniques ne se limite pas aux machines électroniques, telles que les ordinateurs, mais s'applique également aux machines mécaniques. Il ne s'agit pas de dire que les machines numériques sont toutes des systèmes computationnels[176] et les machines mécaniques des systèmes dynamiques[177], mais tout simplement d'élargir le champ d'investigation : si un artefact quelconque résout un problème complexe, son fonctionnement peut être intéressant pour la philosophie de la cognition. Ainsi, d'autres machines mécaniques ont été élaborées ou examinées par le courant antireprésentationnaliste de l'Intelligence Artificielle : le « marcheur dynamique » de Tad McGeer, les créatures de Rodney A. Brooks, l'insecte à six pattes de Randall D. Beer, *etc.*[178] En dehors de ce

[176] Certains programmes informatiques, tels que les réseaux connexionnistes, sont, selon van Gelder lui-même, très proches du modèle dynamique. van Gelder, T. 1998. *Op. cit.*, p. 348 et 355.

[177] Par exemple, le joueur de flûte traversière de Vaucanson et l'écrivain de la famille Jaquet-Droz (deux automates respectivement capables de jouer onze airs de flûte et d'écrire à la plume une combinaison de quarante lettres) sont des machines entièrement mécaniques qui fonctionnent plus sur le modèle computationnaliste que sur le modèle dynamique dans la mesure où la partition et le texte à écrire sont encodés sur un cylindre (à la manière d'un orgue de Barbarie) ou sur une roue crantée. Ainsi, l'exécution de ces deux automates est en partie régie par une suite de symboles (implémentés par les picots du cylindre et les crans de la roue).

[178] McGeer, T. 1990. « Passive dynamic walking. » *International Journal of Robotics Research (IJRR'90)*, vol. 9, n°2, p. 62-82 ; Brooks, R.A. 1991. « Intelligence without representation. » *Artificial Intelligence*, vol. 47,

courant de la robotique, orientée vers les systèmes dynamiques, d'autres systèmes physiques peuvent servir de point de départ au débat philosophique. Ainsi, on trouve dans l'article de van Gelder l'exemple d'une radio, pour expliquer le profil temporel des systèmes dynamiques[179], Varela étudie le système digestif pour savoir s'il peut être considéré comme « manipulant des représentations »[180], Searle utilise la métaphore de l'orage pour rappeler que personne ne considère qu'une simulation informatique d'un orage est un « véritable orage »[181], Russell & Norvig se demandent alors si « les processus mentaux sont plus semblable aux orages ou aux additions. »[182] De manière générale, les sciences physiques et la biologie peuvent, au même titre que l'informatique et la robotique, fournir des dispositifs techniques à partir desquels la stratégie de van Gelder peut s'appliquer. Ainsi, la philosophie peut se demander par exemple si la cognition ne serait pas similaire à un système digestif, plutôt qu'à un ordinateur. Une telle hypothèse de travail change radicalement notre conception des processus cognitifs et, même si elle n'est pas validée par la suite, elle a le mérite de ne pas restreindre les recherches philosophiques aux seules théories classiques.

Cette démarche amène la philosophie à travailler sur les archétypes concrets des théories qu'elle développe. L'Intelligence Artificielle, en proposant de nouvelles machines, ouvre la porte à de nouvelles théories qui, par la suite, sont évaluées par la philosophie de la cognition. Cependant, l'analyse des modes de conception de ces machines offre aussi des pistes pour la validation théorique. Pour van Gelder, par exemple, le régulateur centrifuge a un rapport au temps très différent du régulateur computationnel. Cette différence de « profil temporel »[183] offre un axe

---

p. 139-159 ; Beer, R.D., Quinn, A.D., Chiel, H.J., Ritzmann, R.E. 1997. « Biologically Inspired Approaches to Robotics: What can we learn from insects? » *Communications of the ACM*, vol. 40, n°3, p. 30-38. Pour une mise en situation de ces machines, ainsi que du travail de van Gelder, vis-à-vis de l'hypothèse des systèmes dynamiques, voir également Harvey, I. 2000. *Op. cit.* et Clark, A., Toribio, J. 1994. « Doing Without Representing? » *Synthese: Connectionism and the Frontiers of Artificial Intelligence*, vol. 101, n°3, p. 401-431.

[179] van Gelder, T. 1998. *Op. cit.*, p. 338.

[180] Anspach, M.R., Varela, F.J. 1992. « Le système immunitaire : un "soi" cognitive autonome. » *In* Andler, D. (éd.). 2004. *Introduction aux sciences cognitives* (édition augmentée). Paris : Gallimard, p. 585-605.

[181] « No one supposes […] that computer simulation of a rainstorm will leave us drenched. Why on earth would anyone suppose that a computer simulation of understanding actually understood something? » Searle, J.R. 1980. « Minds, Brains, and Programs. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 84.

[182] « Are mental processes more like storms, or more like addition? » [Notre traduction] Russell, S.J., Norvig, P. 2003. *Artificial Intelligence: A Modern Approach,* 3rd ed. Upper Saddle River, NJ : Prentice Hall, p. 1027.

[183] van Gelder, T. 1998. *Op. cit.*, p. 339.

d'analyse pour distinguer le modèle dynamique du modèle computationnaliste, et pour évaluer leurs pertinences respectives vis-à-vis des propriétés temporelles empiriques de la cognition. Ainsi, pour van Gelder, la cognition est un processus continu, ce qui tend à montrer que « le régulateur de Watt est préférable à la machine de Turing comme archétype des modèles de la cognition. »[184]

[184] *Ibid.*, p. 369.

Chapitre 4.3. Évaluer les modèles de la cognition

Dans un précédent travail de recherche, nous avons proposé une méthode pour confronter et évaluer les théories de l'esprit[185]. Elle consiste à travailler à partir de cas concrets de machines intelligentes pour discuter du bienfondé des positions philosophiques. Ce chapitre reprend les résultats principaux de ce travail pour montrer comment l'Intelligence Artificielle peut aider à évaluer les modèles de la cognition.

Section 4.3.1. Des individus concrets aux modèles de la cognition

Définitions en intension et définitions en extension

Pour un modèle de la cognition donné, nous distinguons sa *définition en intension*, qui décrit la nature de l'esprit et donne une définition compréhensive de ce qu'est la cognition, et sa *définition en extension*, qui désigne l'ensemble des individus satisfaisant les critères énoncés par le modèle. En d'autres termes, la *définition en intension* définit les propriétés de la « véritable intelligence » et la *définition en extension* précise quelles machines sont effectivement intelligentes. Quand deux modèles divergent sur leurs *définitions en intension*, cela se répercute éventuellement sur leurs *définitions en extension*. Par exemple, le modèle cartésien, d'une part, soutient que les animaux ne sont pas doués de conscience et des travaux d'éthologie cognitive récents, d'autre part, montrent que les primates développent un certain niveau de conscience[186]. Les divergences théoriques sur la nature de la conscience entre Descartes et l'éthologie moderne conduisent alors à des expérimentations sur des individus particuliers (ici, des primates). On utilise ainsi les différences *extensionnelles* pour trancher à propos des différences *intensionnelles*. La démarche consiste donc à utiliser « les *définitions en extension* pour discuter des différents modèles de l'esprit, les confirmer ou les infirmer. »[187]

Dans ce chapitre, nous retenons l'idée, similaire à l'approche de van Gelder, que l'étude d'exemples concrets permet d'éclaircir et d'enrichir les débats philosophiques. Mais plus que de

---

[185] Lamarche-Perrin, R. 2010. *Le Test de Turing pour évaluer les théories de l'esprit*. Mémoire de Master, Kistler, M. (dir.). Grenoble : Université Pierre-Mendès-France, sept. 2010.

[186] Denton, D. 1993. *L'Émergence de la conscience : de l'animal à l'homme*. [*The Pinnacle of life. Consciousness and Self-Awareness in Humans and Animals*.] Mourlon, J.-P. (trad.). Paris : Flammarion, 1995.

[187] Lamarche-Perrin, R. 2010. *Op. cit.*, p. 11.

concevoir de nouvelles hypothèses de travail, que la philosophie doit ensuite évaluer[188], nous soutenons que ces exemples concrets peuvent constituer de véritables arguments en faveur ou en défaveur des modèles de la cognition.

Évaluer le test de Turing

Étant donné un test comportemental (tel que le test de Turing, *cf.* section 2.1.2) et un modèle de la cognition, nous avons deux moyens de répondre à la question générale de l'intelligence artificielle : « les machines sont-elles intelligentes ? » À propos d'une machine donnée, le test comportemental répond plus précisément à la question de l'IA faible : « cette machine est-elle capable de *simuler* l'intelligence ? ». Le modèle de la cognition, quant à lui, répond à la question de l'IA forte : « cette machine est-elle *véritablement* intelligence ? » Pour tout individu, nous avons donc deux réponses concernant son intelligence, l'une venant de critères comportementaux formulés par l'IA faible et l'autre de critères cognitifs formulés par l'IA forte.

En supposant que le test de Turing est une méthode adéquate pour répondre au problème de l'IA faible[189], nous pouvons évaluer ses résultats relativement aux critères *intensionnels* d'un modèle donné. Ainsi, comme pour l'évaluation d'un test en médecine, il faut distinguer quatre cas :

- Les *vrais positifs* sont les individus qui passent le test avec succès et qui sont effectivement considérés comme intelligents par le modèle de la cognition : « IA faible (x) et IA forte (x) »
- Les *vrais négatifs* sont, à l'inverse, les individus déclarés non-intelligents par le test *et* par le modèle de la cognition : « non IA faible (x) et non IA forte (x) »
- Les *faux positifs* sont les individus qui passent avec succès le test de Turing, mais qui ne répondent pas aux critères de la théorie cognitive concernée. Ils ne sont pas censés être

[188] Comme dans le cas de l'hypothèse dynamique, rendue viable par l'étude du régulateur centrifuge de Watt, mais dont la confirmation dépend d'un travail empirique des sciences cognitives (*cf.* chapitre précédent).

[189] De nombreuses variantes du test peuvent être envisagées pour améliorer la définition comportementale de l'intelligence. Notamment, nous pouvons imaginer un test qui ne soit pas seulement dialogique, mais également moteur, perceptif, corporel, social, *etc.* Tout dépend des propriétés que l'on souhaite détecter.

« véritablement intelligents », et pourtant ils se comportent comme tel[190] : « IA faible(x) et non IA forte(x) »

- Les *faux négatifs* sont les individus qui échouent au test de Turing, mais qui pourtant, selon le modèle concerné, sont « véritablement intelligents »[191] : « IA forte (x) et non IA faible (x) »

Ces quatre cas sont résumés dans le tableau ci-dessous (figure 7). Les cas de *faux positifs* et les cas de *faux négatifs* nous intéressent alors puisqu'ils mettent en évidence des contradictions entre le test comportemental et le modèle cognitif. Ils nous servent alors à évaluer ce modèle.

		Test de Turing	
		« IA faible (x) »	**« non IA faible (x) »**
Modèle de la cognition	**« IA forte (x) »**	Vrai positif	Faux négatif
	« non IA forte (x) »	Faux positif	Vrai négatif

Figure 7 : l'évaluation du test de Turing

[190] Dans Lamarche-Perrin, R. 2010. *Op. cit.*, p. 31-33, ces *faux positifs* sont nommés « zombies » en référence à l'argument de Robert Kirk contre le physicalisme où l'on imagine des individus se comportant exactement comme nous, mais n'ayant pas de vies conscientes. Kirk, R., Squires, J.E. 1974. « Zombies v. Materialists. » *The Aristotelian Society*, vol. 48, p. 135-163. *Cf.* également la note 71 à ce sujet.

[191] Dans Lamarche-Perrin, R. 2010. *Op. cit.*, p. 29-31, ces *faux négatifs* sont nommés « fous » en référence à l'article de David Lewis où l'on imagine un individu qui partage nos états mentaux (par exemple celui de la douleur), mais qui se comporte très différemment lorsqu'il est dans cet état (par exemple il « croise les jambes et claque des doigts »). Lewis, D. 1978. « Douleur de fou et douleur de martien. » [« Mad Pain and Martian Pain. »]. *In* Fisette, D. (éd.), Poirier, P. (éd.). 2002. *Philosophie de l'esprit : psychologie du sens commun et sciences de l'esprit*. Paris : Vrin, p. 289-306.

Évaluer les modèles de la cognition

Les *faux positifs* et les *faux négatifs* constituent l'ensemble des cas où le test de Turing commet une erreur *vis-à-vis d'un modèle de la cognition donné*. Cependant, il faut considérer pour cela qu'il devrait y avoir une relation d'équivalence entre le comportement intelligent et la « véritable intelligence », c'est-à-dire entre l'IA faible et l'IA forte. Il existe ainsi deux sortes d'erreurs[192] :

- Si l'on souscrit à la relation « IA faible → IA forte », les *faux positifs* « IA faible (x) et non IA forte (x) » sont problématiques. En effet, selon cette relation, un individu se comportant intelligemment devrait toujours être « véritablement intelligent ». La découverte d'un *faux positif* peut avoir l'une des trois conséquences suivantes :
    - **Le test est mis en cause : « non IA faible (x) ».** Contrairement à ce que l'on pensait, l'individu ne se comporte pas de manière intelligente, soit qu'il y ait eu une erreur lors de la réalisation du test, soit que sa conception doit être reconsidérée.
    - **Le modèle est mis en cause : « IA forte (x) ».** Contrairement à ce que l'on pensait, l'individu est bien « véritablement intelligent ». Cela signifie que le modèle de la cognition se trompe quant à cet individu et qu'il doit être reconsidéré.
    - **La relation est mise en cause : « non (IA faible → IA forte) ».** La « véritable intelligence » n'est pas toujours nécessaire aux comportements intelligents, ce qui explique le cas de *faux positif*.
- Si l'on souscrit à la relation « IA forte → IA faible », ce sont les *faux négatifs* « IA forte (x) et non IA faible (x) » qui constituent à leur tour des cas problématiques. Selon cette relation, un individu « véritablement intelligent » devrait toujours passer le test de Turing avec succès. La découverte d'un *faux négatif* peut avoir une des trois conséquences suivantes :
    - **Le test est mis en cause : « IA faible (x) ».** Contrairement à ce que l'on pensait, l'individu se comporte de manière intelligente. Comme précédemment, la réalisation ou la conception du test doit être reconsidérée.

---

[192] Ces deux erreurs et leurs conséquences possibles sont en partie issues de Lamarche-Perrin, R. 2010. « Partie 3. Évaluer les théories de l'esprit. » *Op. cit.*, p. 25-38. Elles ont été reformulées à partir des termes de ce mémoire. La troisième conséquence, concernant la remise en cause de la relation entre les deux problèmes, a été ajoutée pour compléter notre démarche. En effet, les *faux positifs* et les *faux négatifs* peuvent être expliqués par l'indépendance des problèmes de l'IA faible et de l'IA forte.

- **Le modèle est mis en cause : « non IA forte (x) ».** Contrairement à ce que l'on pensait, l'individu n'est pas « véritablement intelligent ». Le modèle de la cognition doit être reconsidéré.

- **La relation est mise en cause : « non (IA forte → IA faible) ».** La « véritable intelligence » n'entraîne pas nécessairement des comportements intelligents, ce qui explique le cas de *faux négatif*.

Dès lors qu'on souscrit à l'une ou l'autre des relations, les cas de *faux positifs* ou de *faux négatifs* révèlent donc des contradictions entre les critères comportementaux et les critères cognitifs. Ils conduisent ainsi à ajuster notre jugement et à remettre en cause un des trois composants de la relation considérée : le test comportemental, le modèle de la cognition ou la relation elle-même. Dans ce qui suit, nous considérons que le test comportemental ne peut être mis en cause dans la mesure où il donne une définition communément acceptée de ce que sont les comportements intelligents[193]. Il s'agit en effet de la partie triviale de la relation, celle que l'on peut difficilement modifier. Ainsi, si l'on défend une relation particulière entre les deux sous-problèmes de l'intelligence artificielle, alors c'est le modèle de la cognition qui doit être reconsidéré. En d'autres termes, celui-ci est falsifié par la machine étudiée. D'autres modèles, au contraire, peuvent être ainsi validés. Nous retrouvons ici la démarche d'Andler et de Harvey faisant de l'Intelligence Artificielle un atelier pour la philosophie de l'esprit (*cf.* chapitre 4.1).

Les deux sections suivantes présentent des exemples de de validation et de falsification à partir de l'étude d'individus concrets. Il s'agit du computationnalisme, validé par l'expérience de la « chambre chinoise » de Searle (*cf.* également section 2.3.2), et du connexionnisme, remis en cause par une expérience réalisée sur des chatons dans les années 50 par Held & Hein.

Section 4.3.2. L'expérience de la « chambre chinoise »

Nous avons vu dans la section 2.2.4 que l'expérience de pensée de la « chambre chinoise » de Searle présente un dispositif se comportant de manière intelligente (il parvient notamment à tenir une conversation en chinois, « IA faible (cc) »), alors que, lorsqu'on regarde de plus près, il n'est pas « véritablement intelligent » (il ne comprend pas un mot de chinois, « non IA forte (cc) »)[194]. De

[193] Éventuellement, on peut considérer que le test de Turing n'est pas adéquat puisqu'il se limite aux seules capacités dialogiques. Mais il est facile de concevoir des extensions, pour détecter toutes sortes de comportements observables (*cf.* note 189 ci-dessus).

[194] Searle, J.R. 1980. « Minds, Brains, and Programs. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 67-88.

même pour le « zombie chanceux » (section 2.2.4) qui passe avec succès le test de Turing en engendrant des comportements aléatoires. Ces deux cas de *faux positifs* (« IA faible (cc) et non IA forte (cc) ») peuvent donc être interprétés de trois manières différentes :

1. **Critique du test de Turing : « non IA faible (cc) ».** Les comportements observables de la « chambre chinoise » ne sont pas des comportements intelligents, contrairement à ce que laisse penser le test de Turing. Cette interprétation est difficilement tenable dans la mesure où les deux dispositifs engendrent *par construction* des réponses similaires à l'homme. Il faudrait donc dire, du même coup, que les hommes ne se comportent pas de manière intelligente.

2. **Critique de la relation « non (IA faible → IA forte) ».** Ce n'est pas parce qu'un individu se comporte de manière intelligente qu'il est « véritablement intelligent ». Ici, ce n'est pas le test en tant que tel qui est critiqué, mais l'utilisation du test pour définir la cognition[195].

3. **Découverte d'une nouvelle forme de cognition : « IA forte (cc) ».** On peut conclure au contraire que les fonctionnements internes de la « chambre chinoise » et du « zombie chanceux » relèvent bien de la « véritable intelligence ». Pour Levesque par exemple, le fonctionnement intelligent est nécessaire *en pratique* à la production de comportements intelligents (*cf.* section 2.2.3). Même si le « zombie chanceux » n'a aucune viabilité en pratique et qu'on peut douter de la réalisabilité d'une « chambre chinoise » telle que Searle la décrit[196], Levesque dirait d'un programme computationnaliste qui parvient réellement à parler chinois qu'il « comprend véritablement » le chinois. Le modèle de la cognition défendu par Searle est alors mis en cause. Si l'on souscrit à cette troisième conséquence, il faut se rendre à l'évidence : le calcul, même lorsqu'il est réalisé physiquement par des dispositifs étranges, doit bien être compris comme un phénomène cognitif.

---

[195] Nous avons vu dans la section 2.3.2 que l'argumentation de Searle était mal dirigée. En effet, en proposant l'expérience de la « chambre chinoise », Searle remet en cause la relation « non (IA faible → IA forte) », défendue notamment par le béhaviourisme, et non le modèle computationnaliste (« IA forte (cc) »). L'argument anti-computationnaliste selon lequel la cognition n'est pas un simple calcul (« non IA forte (cc) ») précède logiquement l'argument anti-béhaviourisme (« IA faible (cc) *et* non IA forte (cc) »). La véritable opposition de Searle au computationnalisme doit donc être trouvée dans son objection *de principe* contre le pouvoir causal des machines numériques, et non dans l'argument de la « chambre chinoise » lui-même.

[196] *Cf.* note 88 à ce sujet.

L'expérience de la « chambre chinoise » conduit donc à prendre position entre un rejet de la relation « IA faible → IA forte » et une validation du computationnalisme « IA forte (cc) »[197]. En accord avec la critique du computationnalisme par Dreyfus, on répondrait aujourd'hui que la relation « IA faible → IA forte » n'est pas valide. En effet, il est possible que des programmes computationnalistes parviennent à produire des comportements intelligents, même si cela leur est très difficile en pratique (version moins radicale, notamment défendue par van Gelder : « IA forte → IA faible possible »). Sans cela, l'expérience de la « chambre chinoise » argumente en faveur du computationnalisme, contrairement à l'objectif initial de Searle.

Section 4.3.3. L'expérience des chatons aveugles

La seconde expérience que nous considérons fait intervenir des animaux et des capacités cognitives de « bas-niveau » (capacités sensori-motrices). La « véritable intelligence » désigne alors le fait d'avoir les états cognitifs suffisants pour percevoir le monde et pour s'y déplacer sans encombre. « Simuler l'intelligence » signifie « avoir des comportements similaires aux individus dotés de tels états cognitifs » et donc se déplacer effectivement sans encombre dans l'environnement. L'expérience qui nous intéresse a été réalisée par Richard Held et Alan Hein[198] et elle est reportée par Varela dans son ouvrage sur la « cognition incarnée »[199] :

> Dans le cadre d'une étude devenue classique, Held et Hein élevèrent des chatons dans l'obscurité et les exposèrent à la lumière seulement dans des conditions contrôlées. Un premier groupe d'animaux furent autorisés à circuler normalement, mais ils étaient attelés à une voiture et à un panier contenant le second groupe d'animaux. Les deux groupes partageaient donc la même expérience visuelle, mais le second groupe était entièrement passif. Quand les animaux furent relâchés après quelques semaines de ce traitement, les chatons du premier groupe se comportaient normalement, mais ceux qui avaient été

[197] Les défenseurs du computationnalisme, et plus largement ceux du fonctionnalisme, ont critiqués la « pétition de principe » formulée par Searle à l'encontre de l'IA forte (*cf.* section 2.3.2). Il s'agit cependant de critiques directes de la forme « IA forte (cc) », ne faisant pas intervenir de machines concrètes pour argumenter, comme ici, à l'aide de la relation « IA faible → IA forte ».

[198] Held, R., Hein, A. 1958. « Adaptation of disarranged hand-eye coordination contingent upon re-afferent simulation. » *Perceptual-Motor Skills*, vol. 8, p. 87-90.

[199] Varela, F.J., Thompson, E.T., Rosch, E. 1991. *L'Inscription corporelle de l'esprit : sciences cognitives et expériences humaine*. [*The Embodied Mind: Cognitive Science and Human Experience*.] Havelange, V. (trad.). Paris : Seuil, 1993.

> véhiculés se conduisaient comme s'ils étaient aveugles : ils se cognaient contre les objets et tombaient par-dessus les bords.[200]

Cette expérience peu commune révèle un cas de *faux négatifs* pour de nombreux modèles de la cognition. En effet, à première vue, les deux groupes de chatons ne devraient pas présenter beaucoup de différences sur le plan cognitif. Même s'ils ont été élevés dans l'obscurité, ils ont ensuite été habitués à la lumière et devraient donc avoir des capacités cognitives similaires à des chatons élevés normalement, c'est-à-dire en pleine lumière. De ce point de vue, il n'y a pas de raisons pour lesquelles les chatons du second groupe, tout comme ceux du premier, aient des déficiences cognitives. Ils devraient donc disposer d'états cognitifs normaux et être ainsi considérés comme « véritablement intelligents »[201]. Selon les théories cognitives communes, on a donc « IA forte (ca) » où « ca » désigne les « chatons aveugles ».

Pourtant, les résultats de l'expérience montrent que les chatons du second groupe sont incapables de se comporter intelligemment (« non IA faible (ca) »). En effet, ils n'ont pas les comportements moteurs que l'on attend d'un chaton ayant été habitué à la lumière. Ils échouent

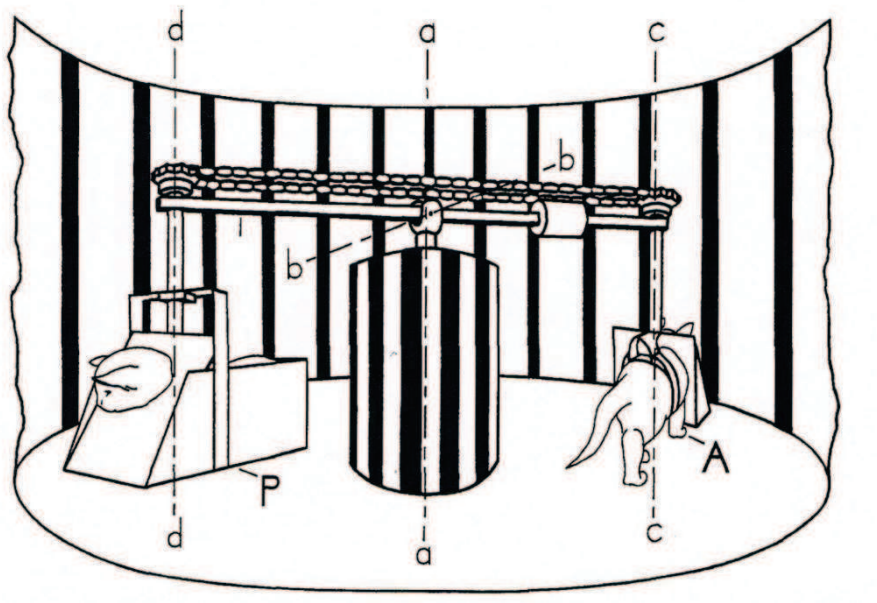

Figure 8 : le dispositif de Held & Hein
Held, R., Hein, A. 1958. *Op. cit.*

---

[200] *Ibid.*, p. 236-237. *Cf.* figure 8 ci-dessus.

[201] Il s'agit bien sûr de l'intelligence communément attribuée à cette catégorie d'individus. La « véritable intelligence » d'un chaton réside bien plus dans ses capacités sensori-motrices (percevoir, se déplacer, éviter les obstacles, *etc.*), que dans ses capacités cognitives de « haut-niveau » comme la faculté de gagner une partie d'échecs.

notamment au test du « déplacement des chatons », au cours duquel un observateur doit juger si les individus passant le test se déplacent à la manière de chatons. Ainsi, l'expérience de Held & Hein révèle un cas de *faux négatif* (« IA forte (ca) et non IA faible (ca) ») dans la mesure où les chatons échouent au test comportemental alors que les modèles classiques de la cognition affirment qu'ils ont les capacités cognitives suffisant à la réussite d'un tel test. Ce résultat peut recevoir trois interprétations différentes :

1. **Critique du test de Turing : « IA faible (ca) ».** Contrairement à ce qu'indique le test, les chatons du second groupe ont bien des comportements intelligents sur le plan sensori-moteur. Il faut élargir les résultats positifs du test pour y incorporer ce genre de comportements inhabituels. Cette interprétation conduit cependant à définir de manière *ad hoc* une catégorie de comportements « intelligents », alors qu'ils témoignent justement d'une difficulté à percevoir le monde et à s'y déplacer. Comme nous l'avons dit précédemment, le test de Turing peut difficilement être mis en cause.

2. **Critique de la relation « non (IA forte → IA faible) ».** Ce n'est pas parce qu'un individu possède les états cognitifs suffisant à la perception du monde et à la motricité, que celui-ci aura des comportements appropriés. Les chatons, même s'ils ont une activité cognitive saine, peuvent agir de manière déficiente. Ainsi, dans certains cas, la relation défendue par Dreyfus ne tient pas : il ne suffit pas d'implémenter le bon modèle pour simuler l'intelligence.

3. **Découverte d'une nouvelle forme de déficience cognitive : « non IA forte (ca) ».** Il faut bien se rendre à l'évidence : les chatons du second groupe ne sont pas « véritablement intelligents ». En d'autres termes, ils n'ont pas, contrairement à ce que prévoyaient nos modèles de la cognition, les états cognitifs suffisant à la perception et à la motricité. Cette déficience cognitive doit donc être expliquée et incorporée à nos théories cognitives. Ainsi, si cette interprétation est retenue, l'expérience des chatons aveugle conduit au perfectionnement des modèles classiques de la cognition.

Ce qui manque ici aux chatons du second groupe, c'est une boucle sensori-motrice lors de leur apprentissage. En effet, lors de leur premier contact avec le monde, ils n'agissent pas, mais sont déplacés de manière passive. Varela expose ainsi les limites de modèles tels que le connexionnisme (et à plus forte raison le computationnalisme) qui ne prennent pas en compte le fait que, lors de l'apprentissage des fonctions cognitives, l'action doit être couplée à la perception. Pour Varela, la cognition n'agit pas dans un environnement qui se donne à la perception de manière objective et passive. Au contraire, la cognition correspond à la création pro-active d'un monde subjectif dont

l'individu doit « *faire-émerger* »[202] les caractéristiques essentielles. En d'autres termes, la cognition ne peut pas se passer de l'action. C'est le *modèle énactif* de l'esprit, défendu par Varela[203] et rejoignant la position de Dreyfus concernant l'importance du corps dans l'exercice de l'intelligence[204].

Section 4.3.4. Bilan de la collaboration

Dans ce chapitre, les relations « IA faible → IA forte » et « IA forte → IA faible » sont exploitées pour confronter et évaluer des modèles de la cognition. La méthode présentée suppose que ces relations soient justifiées en amont du processus et qu'on ne mette pas en cause les résultats du test comportemental. Dès lors, nous avons deux résultats :

- L'expérience de la « chambre chinoise » argumente contre la « pétition de principe » de Searle et en faveur du computationnalisme (« IA forte (cc) »).

- L'expérience des « chatons aveugles » argumente contre les modèles classiques de la cognition (notamment contre le computationnalisme et le connexionnisme) et en faveur du modèle énactif de Varela (« non IA forte (ca) »).

Bien évidemment, si on rejette les relations responsables de ces contradictions, les cas de *faux positifs* et de *faux négatifs* ne sont plus problématiques. Il n'y a alors rien à expliquer. La « chambre chinoise », tout comme les « chantons aveugles », deviennent des cas réguliers où le test comportemental échoue parce qu'il ne présente pas d'équivalence avec les modèles cognitifs.

Cette méthode d'évaluation réalise le renversement épistémologique annoncé par Andler dans la mesure où des individus concrets (machines artificielles et machines biologiques) servent à tester des théories philosophiques. Les sciences *expérimentales*, comme l'Intelligence Artificielle, mais également les sciences cognitives, ont alors leur mot à dire quant aux développements théoriques de sciences *explicatives*, telle que la philosophie.

---

[202] Varela, F.J., Thompson, E.T., Rosch, E. 1991. *Op. cit.*, p. 210.

[203] Voir *Ibid.*, « Quatrième partie. Vers une voie moyenne », p. 189-288 et plus particulièrement *Ibid.*, « Chapitre 8. L'énaction : cognition incarnée », p. 207-248 pour une argumentation détaillée en faveur du modèle énactif. *Cf.* également Varela, F.J. 1988. *Invitation aux sciences cognitives*. [*Cognitive Science. A Cartography of Current Ideas*.] Lavoie, P. (trad.). Paris : Seuil, 1996 pour un résumé de la position de Varela et pour un exposé synthétique des oppositions conceptuelles entre le cognitivisme (dont fait partie le computationnalisme), le connexionnisme et le modèle énactif de l'esprit.

[204] *Cf.* également la section 2.4.3 et Dreyfus, H.L. 1979. « Chapitre 9. Le rôle du corps dans l'exercice de l'intelligence. » *Op. cit.*, p. 301-327.

Partie 5. Conclusion

De nombreuses formes de collaboration ont été présentées dans ce mémoire. Avant d'en faire la synthèse, il faut se pencher sur les positions non-collaboratives selon lesquelles les échanges entre la philosophie et l'Intelligence Artificielle sont soit impossibles, soit inutiles. Nous avons d'abord vu que, en pratique, les philosophes et les spécialistes de l'IA n'échangeaient qu'à de très rares occasions (section 2.1.1). Cela ne doit pas néanmoins constituer un argument *contre* la collaboration. Ce n'est pas parce qu'il y a peu d'échanges aujourd'hui qu'il ne peut y en avoir plus à l'avenir. Ce mémoire œuvre justement dans cette direction.

Le béhaviouriste méthodologique déclare que les théories de l'esprit sont *inutiles* au projet de l'Intelligence Artificielle (section 2.1.2). Le béhaviourisme logique les *élimine* carrément (section 2.2.1). Si on se restreint au problème des autres esprits et à la conscience des machines, ces positions sont peut-être justes, mais c'est limiter considérablement les questions couvertes par la philosophie. L'essor des sciences cognitives a montré que des modèles philosophiques concernant la nature de la cognition pouvaient outrepasser l'interdiction béhaviouriste. La question de la « véritable intelligence » ne se limite plus à la conscience des esprits, mais s'interroge sur leurs structures informationnelles, leurs fonctionnements cognitifs et d'autres propriétés qui peuvent être soumis à l'examen empirique. La critique de John R. Searle, enfin, qui interdit à l'Intelligence Artificielle de se préoccuper de questions philosophiques (section 2.3.2), s'inscrit dans cette compréhension étroite des problèmes de l'esprit. S'il est possible que l'Intelligence Artificielle ne soit d'aucune aide pour résoudre le problème de la conscience, elle n'en reste pas moins appréciable pour investiguer et comprendre de nombreuses autres facettes de l'esprit. En conclusion, pour qu'elle s'épanouisse, la collaboration ne doit pas se limiter à des controverses métaphysiques qui semblent rester indécidables. La *philosophie de l'Intelligence Artificielle*, telle qu'elle a été développée jusqu'aux années 80, a trop souvent borné l'Intelligence Artificielle à la conscience des machines et, réciproquement, les spécialistes de l'IA ont eu une vision trop restreinte des champs de la philosophie.

Dans le chapitre 5.1, nous faisons la synthèse des collaborations fructueuses présentées tout au long du mémoire. Le chapitre 5.2 indique enfin comment celles-ci doivent être utilisées pour opérer un rapprochement entre Intelligence Artificielle et philosophie.

Chapitre 5.1. Résumé des modes de collaboration

Classification des collaborations

Les modes de collaboration abordés dans ce mémoire sont synthétisés dans la figure 9 ci-dessous. Ils sont classifiés selon deux aspects : la direction et le signe de la collaboration. La *direction* indique laquelle des deux disciplines (« discipline source ») sert aux recherches de l'autre (« discipline cible »). Les exemples que nous avons présentés peuvent ainsi être analysés à partir d'éléments de collaboration unilatéraux. Cela ne signifie pas que, de manière plus générale, des collaborations bilatérales ne puissent être mises en place. Le chapitre 5.2 compose les éléments présentés dans la figure 9 pour préciser de telles démarches. Le *signe* de la collaboration indique le rôle de la « discipline source » quant aux recherches de la « discipline cible ». Il peut être *positif*, indiquant que la « discipline source » offre de nouveaux concepts à la « discipline cible » et lui sert d'inspiration. Il peut être *négatif*, indiquant que la « discipline source » met en évidence les erreurs conceptuelles de la « discipline cible » et a ainsi un objectif de falsification.

Les collaborations reposent sur l'une des deux relations réciproques « IA faible → IA forte » et « IA forte → IA faible ». En outre, nous montrons que chacune de ces deux relations peut servir à des collaborations bidirectionnelles de signes différents, en utilisant leurs contraposées. Ainsi, « IA faible → IA forte » permet un apport *positif* de l'Intelligence Artificielle vers la philosophie et sa contraposée « non IA forte → non IA faible » permet un apport *négatif* de la philosophie vers l'Intelligence Artificielle. Réciproquement, « IA forte → IA faible » permet un apport *positif* de la philosophie vers l'Intelligence Artificielle et « non IA faible → non IA forte » un apport *négatif* de l'Intelligence Artificielle vers la philosophie.

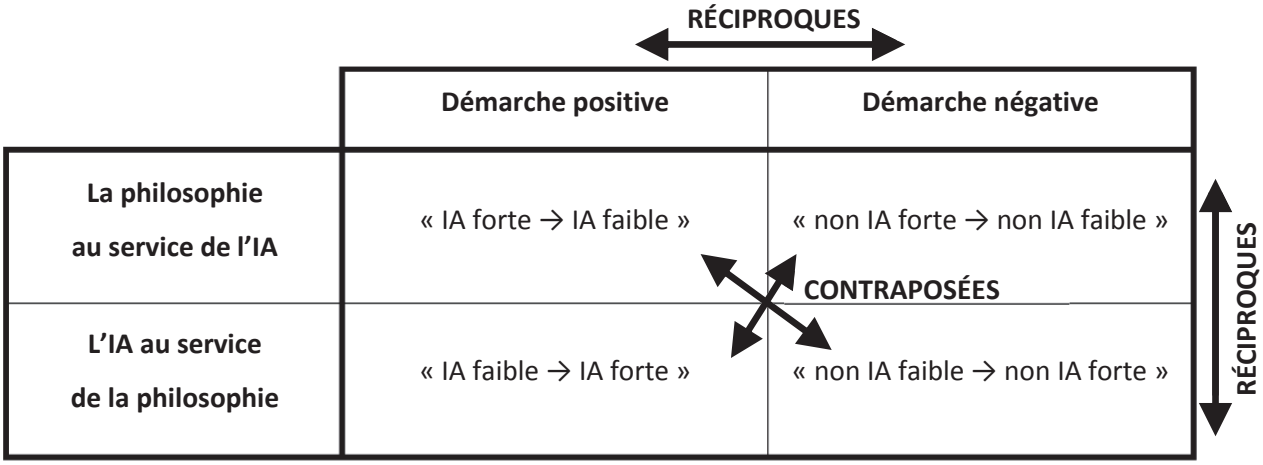

Figure 9 : classification des modes de collaboration

Relation « IA forte → IA faible »

Les collaborations reposant sur la relation « IA forte → IA faible » sont des démarches positives au sein desquelles les théories et les concepts de la philosophie sont appliqués à l'Intelligence Artificielle.

- Dans le chapitre 3.1, nous avons vu que la phénoménologie a bénéficié au développement d'une « nouvelle IA » en proposant une conception inhabituelle de l'esprit et de la cognition. La robotique incarnée, les systèmes dynamiques et, plus généralement, les courants antireprésentationnalistes de l'Intelligence Artificielle ont été rendus possibles par cet apport philosophique.

- Dans le chapitre 3.2, une théorie métaphysique concernant la structure de la réalité et la nature de la connaissance sert de base conceptuelle à l'Intelligence Artificielle. La notion d'« émergence épistémique », empruntée à la philosophie britannique, permet de définir des contraintes méthodologiques pertinentes pour la simulation de systèmes complexes et la résolution de problèmes distribués. La philosophie a ainsi un impact positif sur les méthodes pratiques de l'Intelligence Artificielle.

Relation « non IA forte → non IA faible »

Les collaborations reposant sur la relation « non IA forte → non IA faible » sont des démarches négatives au sein desquelles la philosophie se charge d'expliquer ou de prédire les difficultés pratiques de l'Intelligence Artificielle à partir de ses erreurs de conceptualisation.

- Dans la section 2.4.2, nous avons vu comment Dreyfus explique les échecs pratiques des applications du computationnalisme par une analyse de leurs postulats philosophiques implicites. Plus largement, c'est parce que ces postulats proviennent de la tradition cartésienne, qui défend elle-même une théorie erronée de la cognition, que le computationnalisme peine, en pratique, à produire des comportements intelligents.

- Dans le chapitre 3.2, nous montrons également qu'une mauvaise conceptualisation de l'émergence induit des démarches moins pertinentes. Le « dualisme » conduit à utiliser des modèles *ad hoc* pour simuler les phénomènes macroscopiques, au lieu d'en reproduire les dynamiques fondamentales à partir du niveau microscopique. L'« éliminativisme » interdit l'utilisation d'abstractions de « haut-niveau » pour l'analyse des systèmes, et peine ainsi à décrire les phénomènes complexes.

Relation « IA faible → IA forte »

Par contraposition, les collaborations reposant sur la relation « IA faible → IA forte » sont des démarches positives au sein desquelles les machines de l'Intelligence Artificielle offrent des perspectives nouvelles à la philosophie. Elles permettent de formuler des hypothèses et d'argumenter en faveur de certains modèles de la cognition.

- Dans la section 2.2.3, nous avons vu que Levesque soutient qu'un programme qui résout des problèmes complexes ne peut pas, en pratique, le faire de manière complètement « stupide ». Ainsi, si un dispositif tel que la « chambre chinoise » parvient effectivement à passer le test de Turing, alors son fonctionnement doit être considéré avec attention par la philosophie, parce qu'il implémente nécessairement des opérations cognitives complexes. Selon la démarche de Levesque, une réussite *en pratique* des programmes computationnalistes constitue donc un argument en faveur de son socle philosophique.

- Dans le chapitre 4.2, de manière moins radicale, van Gelder, en faisant l'analyse technique d'appareils de régulation, propose une nouvelle hypothèse de travail pour la philosophie de la cognition. Le fait qu'un système dynamique parvienne en pratique à résoudre un problème complexe ne constitue pas une preuve en faveur d'un modèle dynamique de la cognition. Cela constitue néanmoins une piste pertinente que la philosophie se doit de considérer avec attention et, à terme, elle est chargée de l'évaluer.

Relation « non faible → non IA forte »

Enfin, les collaborations reposant sur la relation « non IA faible → non IA forte » sont des démarches négatives au sein desquelles l'Intelligence Artificielle peut évaluer les théories de l'esprit à l'aide de machines concrètes. Elle a un rôle de falsification.

- Dans le chapitre 4.1, nous présentons des auteurs qui défendent cette démarche. Andler et Harvey font ainsi de l'Intelligence Artificielle un « banc d'essai » pour tester les théories philosophiques « dans le monde réel ». Pour Dreyfus, un échec généralisé du computationnalisme, et plus largement de l'Intelligence Artificielle, aurait de lourdes conséquences philosophiques. Il mettrait notamment en cause la tradition philosophique élaborée et défendue au cours des vingt derniers siècles.

- Dans la section 4.3.3, l'expérience des « chatons aveugles » de Held & Hein met en évidence les lacunes des modèles classiques de la cognition. Ceux-là ne rendent pas compte du comportement des chatons observés. Ces modèles sont donc mis en échec par l'étude de machines concrètes (ici, des machines organiques) et doivent être améliorés ou abandonnés.

Chapitre 5.2. Comment collaborer ?

Quelle relation utiliser ?

Les modes de collaboration présupposent l'une ou l'autre des deux relations « IA forte → IA faible » et « IA faible → IA forte », en utilisant éventuellement leurs contraposées. Il est donc nécessaire de soutenir l'une ou l'autre pour espérer collaborer. Nous avons vu néanmoins que ces relations n'avaient pas besoin d'être défendues strictement pour opérer le rapprochement des deux disciplines. En effet, des formes plus souples de relation sont possibles. Elles consistent à remplacer l'affirmation du second terme par une notion de « possibilité » ou de « forte probabilité ». Par exemple, « non IA forte → non IA faible » devient « non IA forte → IA faible peu probable » dans la démarche empruntée par Dreyfus (section 2.4.2) et « IA faible → IA forte » devient « IA faible → IA forte possible » dans le cas de van Gelder (chapitre 4.2). Il ne s'agit pas de dire que la « discipline source » détermine à coup sûr les résultats de la « discipline cible », mais qu'elle révèle des pistes d'analyse intéressantes, qui nécessitent toutefois un travail de validation de la part de la « discipline cible ».

Par ailleurs, il semble plus facile de défendre la relation « IA forte → IA faible » que la relation « IA faible → IA forte » dans la mesure où l'Intelligence Artificielle est aujourd'hui capable de produire toutes sortes de comportements intelligents à partir de méthodes très variées. Une réussite de l'Intelligence Artificielle peut ainsi être expliquée par le seul travail des spécialistes de l'IA, sans prendre en compte les théories philosophiques sous-jacentes. La relation stricte « IA faible → IA forte » est par exemple réfutée par de nombreux exemples de machines simulant l'intelligence sans pour autant avoir des propriétés cognitives satisfaisantes (dans ce mémoire, nous citons le cas du « zombie chanceux » et de la « chambre chinoise », *cf.* section 2.2.4). Les résultats positifs d'une science *applicative* doivent donc être utilisés avec prudence dans le cadre d'une science *explicative*. Ce mémoire montre néanmoins qu'une telle démarche est possible.

Au contraire, la méthode collaborative utilisant les résultats positifs d'une science *explicative* pour servir au développement d'une science *applicative* est plus communément admise. C'est le cas des démarches reposant sur la relation « IA forte → IA faible » au sein desquelles la philosophie sert de socle à l'Intelligence Artificielle. Cependant, de telles démarches ne sont pas inattaquables dans la mesure où il est parfois difficile d'identifier, en Intelligence Artificielle, ce qui relève de la construction des machines et ce qui relève des théories implicites. Ainsi, un échec de l'Intelligence Artificielle peut avoir de nombreuses causes, y compris des difficultés internes à la discipline. La relation « non IA faible → non IA forte » doit donc être également manipulée avec le plus grand soin.

Démarche positive ou démarche négative ?

Une fois que l'on a défini la relation que nous souhaitons exploiter, le signe de la collaboration détermine sa direction. Ainsi, le choix d'une démarche positive ou négative dépend de qui tiendra le rôle de « discipline source » et qui tiendra le rôle de « discipline cible ». Une communauté de recherche est alors placée en amont de l'autre. D'un point de vue épistémologique, cela signifie que la « discipline source » critique et oriente les recherches de la « discipline source ». C'est elle qui « mène » la collaboration.

Cependant, pour une relation donnée, l'utilisation simultanée d'une démarche positive et d'une démarche négative permet d'élaborer une collaboration bilatérale. L'une des disciplines est alors source d'inspiration, l'autre est source de falsifications. Par exemple, les relations « IA faible → IA forte » et sa contraposée « non IA forte → non IA faible » font de l'Intelligence Artificielle une source d'inspiration pour la philosophie qui, en retour, évalue négativement certaines approches de l'Intelligence Artificielle. Philosophes et spécialistes de l'IA ont alors chacun un rôle déterminé (inspiration ou falsification), qu'ils peuvent exercer au sein d'une collaboration bilatérale.

Enfin, si l'on argumente en faveur des deux relations, on définit alors une « collaboration maximale » qui s'appuie sur la relation d'équivalence « IA forte ↔ IA faible ». Chacune des deux communautés de recherche peut alors exercer un rôle positif *et* négatif. Elles peuvent être à la fois une source d'inspiration *et* une méthode de falsification pour sa coéquipière.

Annexe

Des collaborations possibles entre Intelligence Artificielle et philosophie de l'esprit

Résumé

Dans cet article, les relations logiques entre deux problèmes paradigmatiques concernant l'intelligence des machines (IA faible et IA forte) servent à exemplifier les rapports possibles entre Intelligence Artificielle et philosophie de l'esprit. Après avoir écarté la méthode béhavioriste de Turing (IA faible → IA forte) et l'hypothèse computationnaliste de Newell & Simon (IA faible ↔ IA forte), l'article s'intéresse à la démarche critique adoptée par Dreyfus. Son analyse du computationnalisme, appuyée sur l'examen de ses origines philosophiques, constitue une stratégie logique féconde pour penser les rapports entre Intelligence Artificielle et philosophie de l'esprit (IA forte → IA faible). Cette stratégie peut également être exploitée pour résoudre des difficultés conceptuelles relatives à la simulation informatique des phénomènes émergents. Le concept d'« émergence épistémique », emprunté aux controverses de la philosophie britannique, induit notamment des résultats méthodologiques intéressants pour l'Intelligence Artificielle. Ce rapprochement particulier (IA forte → IA faible) débouche sur un second mode de collaboration (non IA faible → non IA forte) au sein duquel l'Intelligence Artificielle devient à son tour une philosophie expérimentale, *i.e.* une science au service de la philosophie.

1. Introduction

Cet article s'intéresse aux collaborations possibles entre Intelligence Artificielle (IA) et philosophie. Une première approche évidente consiste à définir une *philosophie de l'IA*, au sens usuel de *philosophie des sciences*. Une épistémologie, donc, qui s'intéresse de manière extra-théorique aux fondements, aux objets et aux méthodes de l'IA. Nous soutenons cependant qu'il existe un endroit où IA et philosophie collaborent *à un même niveau*, c'est-à-dire au sein d'une même théorie et à propos des mêmes objets. Nous examinons pour cela le rôle particulier de la *philosophie de l'esprit* dans l'élaboration de l'IA moderne. Nous souhaitons également rendre compte du baptême administré par Daniel Andler dans l'avant-propos à l'édition française de l'ouvrage d'Hubert L. Dreyfus *What Computers Can't Do*. Andler y désigne l'IA comme « science de la philosophie »,

comme « *science philosophique* » par excellence[1]. Cet article vise également à comprendre les termes d'un tel baptême.

La distinction entre IA faible et IA forte, introduite à plusieurs reprises par John R. Searle[2], offre de bons jalons pour discuter des rapports possibles entre IA et philosophie de l'esprit. L'*IA faible* pose un problème pratique : peut-on réaliser des machines qui *simulent* les comportements intelligents et qui résolvent ainsi des problèmes techniques ou conceptuels habituellement appréhendés par l'homme ? L'*IA forte* répond à un problème de nature plus philosophique. Elle s'intéresse à l'ontologie même des machines : sont-elles *réellement* intelligentes ou ne font-elles que *simuler* l'intelligence ? Plusieurs critères peuvent être introduits pour distinguer ce qui est de l'ordre de la véritable intelligence et ce qui est de l'ordre de la simulation. Les discussions les plus classiques font intervenir les notions d'états mentaux, d'intentionnalité (est-ce qu'une machine est capable d'états intentionnels ?) ou de conscience (est-ce qu'une machine peut avoir une vie phénoménale, une conscience de soi, *etc.* ?).

À première vue, les chercheurs qui travaillent sur ces deux problèmes appartiennent à des communautés distinctes. Il y a d'une part les spécialistes de l'IA, dont l'objectif premier est la production de comportements intelligents, et pour lesquels la question d'une « véritable intelligence » a peu de sens[3]. Tout comme un ingénieur en aéronautique se demande rarement si un avion « simule le vol » ou s'il « vole réellement »[4], la plupart des chercheurs en IA considèrent qu'une machine est intelligente lorsqu'elle peut résoudre des problèmes difficiles, peu importe la similarité de fonctionnement avec l'intelligence naturelle. Ainsi, les travaux des spécialistes font très

---

[1] H. L. Dreyfus, *Intelligence Artificielle : mythes et limites* [*What Computers Can't Do: The Limits of Artificial Intelligence*, *2nd ed.*, 1979], traduit par R.-M. Vassallo-Villaneau, avant-propos de D. Andler et J. Perriault, Paris : Flammarion, 1984, p. XIV.

[2] Cette distinction apparaît pour la première fois dans J. R. Searle, « Minds, Brains, and Programs » [1980], *in* M. A. Boden éd., *The Philosophy of Artificial Intelligence*, Oxford, New York : Oxford University Press, 1990, p. 66. Elle est souvent réitérée par Searle et utilisée par des chercheurs en philosophie et en IA. Les acceptions retenues dans cet article sont celles de S. Russell et P. Norvig, *Artificial Intelligence: A Modern Approach*, *3rd ed.*, Upper Saddle River, New Jersey : Prentice Hall/Pearson Education, 2010, p. 1020.

[3] « *Most AI researchers take the Weak AI hypothesis for granted, and don't care about the strong AI hypothesis—as long as their program works, they don't care whether you call it a simulation of intelligence or real intelligence.* » S. Russell et P. Norvig, *op. cit.*, p. 1020.

[4] Cette analogie est empruntée à *ibid.*, p. 3 et 1021.

marginalement référence aux notions de conscience et d'états mentaux[5]. Les philosophes, d'autre part, lorsqu'ils s'intéressent au problème de la conscience des machines (un sous-problème de l'IA forte), n'ont pas d'avis sur les possibilités et les difficultés de l'IA faible. La plupart n'émettent simplement aucune objection de principe[6].

Aucune collaboration n'est possible, dès l'abord, entre spécialistes de l'IA et philosophes de l'esprit. Les chercheurs n'aspirent simplement pas à résoudre les mêmes problèmes. Cependant, l'indépendance de ces problèmes est à mettre en question. Dans cet article, nous envisageons les cas où une dépendance logique ou empirique apparaît entre les hypothèses de l'IA faible et de l'IA forte. De telles dépendances témoignent d'interconnections entre le problème pratique et le problème philosophique. Elles nous servent ainsi à exemplifier les rapports possibles entre IA et philosophie de l'esprit, comme un cas particulier des rapports entre science et philosophie. Dans la première section, les dépendances « IA faible → IA forte » et « IA faible ↔ IA forte »[7] sont abordées à partir de deux paradigmes classiques de l'IA : le behaviorisme d'Alan M. Turing, en 1950, et le computationnalisme, à l'origine de l'« IA classique » dans les années 1960. La deuxième section examine deux approches critiques de l'IA, argumentées par Searle et par Dreyfus. Le travail de Dreyfus est exploité pour établir une collaboration constructive entre les deux disciplines, sur la base de la dépendance « IA faible ← IA forte ». Les termes de cette collaboration sont développés et généralisés dans les deux dernières sections. Elles montrent comment les modèles philosophiques peuvent offrir une base saine au travail empirique et comment celui-ci peut en retour aider la philosophie à évaluer ses théories.

2. Les débuts de l'Intelligence Artificielle

Le behaviorisme. Dans son article de 1950, Alan M. Turing donne à l'intelligence une définition strictement comportementale : une machine est intelligente lorsqu'elle *se comporte*

---

[5] Par exemple, dans l'appel à la célèbre conférence de Dartmouth, où le terme même d'« Intelligence Artificielle » a été décidé, et dans l'ensemble des travaux de « simulation cognitive » qui lui ont succédé, les conjectures concernaient seulement la *simulation* de l'intelligence : « *every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.* » J. McCarthy, M. L. Minsky, N. Rochester et C. E. Shannon, « A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence » [1955], *in AI Magazine*, vol. 27, n°4, 2006, p. 12.

[6] *Cf.* par exemple la position de Searle : « *I have no objection to the claims of weak AI, at least as far as this article is concerned.* » J. R. Searle, *op. cit.*, p. 67.

[7] Les symboles → et ↔ désignent respectivement l'implication et l'équivalence logiques. Les expressions « IA faible → IA forte », « IA forte → IA faible » et « IA faible ↔ IA forte » sont explicitées ci-après.

comme un homme[8]. Le désormais célèbre « Test de Turing » constitue une méthode empirique pour détecter une faculté psychologique à partir de ses propriétés observables, *i.e.* les comportements qu'elle engendre. Il hérite ainsi du behaviorisme méthodologique, apparu en psychologie au début du siècle. Le test répond au problème de l'IA faible (simulation de l'intelligence). Mais, en évitant de s'adosser aux qualités internes de la faculté, il reste muet quant au problème de l'IA forte (intelligence réelle).

Il est intéressant de noter que Turing se positionne également vis-à-vis du problème de l'IA forte et notamment vis-à-vis du problème difficile de la conscience appliqué aux machines. Pour Turing, il est impossible de déterminer si une machine a *e.g.* une vie phénoménale, à moins bien sûr d'« *être* la machine et de se sentir penser soi-même. »[9] Face au *problème des autres esprits*, réputé parmi les philosophes, Turing recommande d'avoir « la convention polie (*the polite convention*) que tout le monde pense. »[10] C'est effectivement la solution que l'on adopte quotidiennement pour échapper au solipsisme. Cependant, en déployant toutes les conséquences de sa proposition, Turing aurait pu postuler en faveur d'un behaviorisme logique, plus radical que son homologue méthodologique. Ici en effet, le concept de conscience n'a aucune signification en dehors de ses propriétés observables. Il s'agit une propriété analytique que l'on réduit à une simple convention entièrement déterminée par l'implication « IA faible → IA forte » : le constat empirique de comportements intelligents définit *à lui seul* la notion de conscience. Les concepts de la philosophie de l'esprit sont dès lors inadéquats pour résoudre le problème de l'IA forte. Ils sont éliminés et remplacés par la « convention polie » de Turing qui récuse alors toute collaboration possible entre IA et philosophie de l'esprit[11].

Le computationnalisme. Les sciences cognitives se sont organisées autour d'une critique véhémente du behaviorisme. Dans les années 1960, parmi les hypothèses fondatrices du cognitivisme, le computationnalisme est directement influencé par les avancées de l'IA. Une acception technique en est donnée par Allen Newell et Herbert A. Simon sous le nom de « *Physical-*

---

[8] A. M. Turing, « Computing Machinery and Intelligence » [1950], *in* M. A. Boden éd., *The Philosophy of Artificial Intelligence*, Oxford, New York : Oxford University Press, 1990, p. 40-66.

[9] *Ibid.*, p. 52. [Notre traduction]

[10] *Ibid.* [Notre traduction]

[11] Turing n'ira pas jusque-là. Pour lui, plus simplement, le problème de l'IA faible peut être résolu sans qu'on ait résolu celui de l'IA forte. *Cf. ibid.*, p. 53. À l'indépendance des deux problèmes, postulée par Turing, nous appliquons ici la position plus stricte du béhaviorisme logique, visant à la réduction et à l'élimination de l'un des deux problèmes.

Symbol System Hypothesis: A physical symbol system[12] *has the necessary and sufficient means for general intelligent action.* »[13] Il s'agit d'une hypothèse double. Par « suffisant », on entend que les ordinateurs, en tant que systèmes symboliques physiques, peuvent en principe agir intelligemment. En ce sens, l'hypothèse de Newell & Simon est ni plus ni moins une généralisation de l'hypothèse de l'IA faible. Par « nécessaire », on entend que le cerveau humain, puisqu'il est capable d'engendrer des comportements intelligents, doit lui aussi être une sorte de système symbolique physique. En un sens, le cerveau est donc similaire à l'ordinateur et l'esprit aux programmes qu'il exécute. La théorie computo-représentationnelle de l'esprit, dont Jerry A. Fodor est sans doute le plus grand défenseur du côté de la philosophie[14], affirme que (1) l'esprit est un système symbolique implémenté par un cerveau et que (2) la cognition consiste en un calcul sur ces symboles, *i.e.* une *computation*. Par ailleurs, cette analogie apporte une solution au problème de l'IA forte : à l'instar de leurs analogues biologiques, les ordinateurs sont en principe capables d'engendrer la conscience.

L'hypothèse computationnaliste a participé à l'essor des sciences cognitives, notamment parce qu'elle amenait de nombreuses disciplines (psychologie, neurobiologie, linguistique, philosophie, Intelligence Artificielle, *etc.*) à étudier de concert une seule catégorie d'objets : les systèmes symboliques et leurs implémentations physiques. L'hypothèse double de Newell & Simon affirme en effet l'équivalence des objets et des problèmes : « IA faible ↔ IA forte ». Elle donne une possibilité de collaboration intra-théorique maximale entre IA et philosophie de l'esprit.

Pourtant, le computationnalisme – et plus généralement le cognitivisme – ont été très largement remis en cause à partir des années 1980. Nous arrêtons donc ici l'analyse de cette collaboration particulière pour nous concentrer sur les critiques qui lui ont été opposées.

3. Les critiques de l'Intelligence Artificielle

La critique searlienne. La critique du computationnalisme par John R. Searle, dont l'argument le plus célèbre est celui de la « Chambre Chinoise »[15], souffre de plusieurs défauts. Avant

---

[12] Un *système symbolique* est constitué d'un ensemble de symboles (le vocabulaire) et d'opérations (la syntaxe). Un *système symbolique physique* est la réalisation matérielle d'un tel système. Les ordinateurs sont des exemples canoniques de *systèmes symboliques physiques*.

[13] A. Newell et H. A. Simon, « Computer Science as Empirical Inquiry: Symbols and Search » [1976], *in* M. A. Boden éd., *The Philosophy of Artificial Intelligence*, Oxford, New York : Oxford University Press, 1990, p. 111.

[14] J. A. Fodor, *The Language of Thought*, Cambridge, Massachusetts : Harvard University Press, 1975.

[15] J. R. Searle, « Minds, Brains, and Programs » [1980], *in* M. A. Boden éd., *The Philosophy of Artificial Intelligence*, Oxford, New York : Oxford University Press, 1990, p. 67-88.

toute chose, elle est souvent mal renseignée sur les recherches en IA qui lui sont contemporaines. L'argumentation ne couvre qu'une catégorie limitée de programmes et n'atteint pas l'universalité qu'elle prétend avoir. Par exemple, dans le cas de la « Chambre Chinoise », seuls le behaviorisme de Turing et une forme « linéaire »[16] du computationnalisme sont réellement mis en cause. De plus, lorsqu'on l'examine avec précaution, l'objection majeure de Searle apparaît comme étant avant tout une objection *de principe*. Il est optimiste en ce qui concerne le rôle des neurosciences dans la résolution du problème difficile de la conscience *humaine*. Par contre, si les cerveaux ont la base biologique nécessaire à l'intentionnalité et à l'émergence d'une conscience, Searle soutient que les transistors de silicium ne disposent pas d'un tel pouvoir causal[17]. Irrémédiablement, il nie ainsi la possibilité d'une IA forte.

En ce qui concerne les rapports entre IA et philosophie, le défaut majeur de la critique searlienne est qu'elle s'intéresse uniquement au problème de l'IA forte (intentionnalité et conscience des machines). Searle ne fournit aucune analyse des possibilités de l'IA faible et son travail ne peut donc être exploité par les spécialistes de la seconde communauté.

La critique dreyfusienne. La critique d'Hubert L. Dreyfus ne présente pas les mêmes défauts. Premièrement, à l'inverse de Searle, Dreyfus est bien informé des avancées et des résultats de l'IA. En témoigne la première partie de *What Computers Can't Do*[18] dans laquelle il établit un bilan critique de vingt années de recherches, entre 1957 et 1977.

---

[16] Nous qualifions de « computationnaliste linéaire » tout programme qui associe de manière linéaire ses entrées (*inputs*) et ses une sorties (*outputs*) sans tenir compte d'éventuelles variables internes. Il semble que le premier programme pris en exemple dans l'expérience de Searle est bien de ce type : un simple dictionnaire d'associations linéaires. Daniel C. Dennett dénonce la simplicité de l'argument fondé sur « *the (unwarranted) supposition that the giant program would work by somehow simply "matching up" the input Chinese characters with some output Chinese characters*. » Les autres exemples présentés dans l'article de Searle, dont la complexité reflète mieux celle des programmes développés à l'époque, limitent cependant la clarté de son argument. Pour Dennett justement, « *complexity does matter.* » D. C. Dennett, *Consciousness Explained*, New York : Hachette Book, 1991, p. 431-455.

[17] « *Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena.* » J. R. Searle, « Minds, Brains, and Programs » [1980], *in* M. A. Boden éd., *The Philosophy of Artificial Intelligence*, Oxford, New York : Oxford University Press, 1990, p. 86-87. Dennett s'oppose à cette pétition de principe : D. C. Dennett, *op. cit.* et D. J. Chalmers, *The Conscious Mind*, New York : Oxford University Press, 1996, section IV.9.1.

[18] H. L. Dreyfus, *op. cit.*

Première partie. Parmi les projets que Dreyfus analyse, le *General Problem Solver*[19] (GPS) de Newell & Simon constitue une application canonique du computationnalisme. Le programme consiste en un ensemble d'algorithmes génériques et de fonctions heuristiques travaillant à la résolution de problèmes formalisés. Par exemple, à partir de l'ensemble des règles du jeu d'échecs (*base de règles*), de la position des pièces sur l'échiquier (*base de faits*) et d'un *objectif* à atteindre, le GPS peut en théorie proposer un coup régulier et viable sur le plan stratégique. Newell & Simon supposent de plus que tout problème peut être modélisé par un système symbolique (faits et règles) et que sa résolution peut être engendrée par un calcul sur ces symboles (exécution d'algorithmes génériques). Outre les difficultés spécifiques aux problèmes abordés par le GPS, notamment liées à l'explosion combinatoire des faits et des opérations possibles, cette approche computationnaliste rencontre deux difficultés de taille. Premièrement, il est très difficile d'agréger les problèmes modélisés pour augmenter la portée du GPS. Les « micromondes »[20] ne s'imbriquent pas aisément et le programme, perdu dans la multitude des problèmes possibles, n'atteint pas la généricité escomptée. Deuxièmement, certains problèmes sont très difficiles à formaliser en termes de faits et de règles, notamment ceux dont l'espace des états possibles n'est pas autant contrôlé que celui du jeu d'échecs. Exemple parmi d'autre, la motricité dans un environnement complexe, dynamique et incertain résiste à une description symbolique intégrale. Dreyfus montre comment, face à ces difficultés, l'idée d'un programme générique, capable de résoudre « tout type de problème », est soldée par un échec et finalement abandonnée.

Deuxième partie. La deuxième partie de l'ouvrage s'efforce d'expliciter les soubassements philosophiques des approches computationnalistes. Pour Dreyfus, trois postulats non-avoués sont à l'origine de l'échec du GPS et de toute autre tentative d'application du computationnalisme. Le *postulat psychologique* est l'hypothèse computo-représentationnelle elle-même, associant cognition et computation. Le *postulat épistémologique* affirme que « tout savoir peut être explicitement formulé. »[21] Le *postulat ontologique* affirme que « tout ce qui existe est un ensemble de faits, dont chacun est logiquement indépendant de tous les autres. »[22] Dreyfus montre comment ces

[19] A. Newell et H. A. Simon, « GPS, A Program that Simulates Human Thought », *in* E. A. Feigenbaum et J. Feldman éd., *Computers and Thought*, New York : McGraw-Hill, 1963.

[20] Les « micromondes » sont des modèles simplistes de problèmes réels. Il s'agit d'univers clos, certains, discrets, dont la complexité des variables d'état est largement atténuée. Le jeu d'échecs peut être considéré comme un micromonde dans la mesure où un état du plateau et ses coups réguliers peuvent être décrits intégralement et avec certitude.

[21] H. L. Dreyfus, *op. cit.*, p. 192.

[22] *Ibid.*, p. 193.

présupposés héritent d'une longue tradition philosophique. On peut citer par exemple Thomas Hobbes et son « "*reason*" […] *is nothing but "reckoning"* »[23], Descartes et l'idée que l'esprit est un « miroir de la nature », Leibniz et l'espoir de construire un langage algorithmique capable de venir à bout des problèmes philosophiques, l'idée kantienne que tout comportement humain est régi par des règles transcendantales qu'il faut s'efforcer d'expliciter, le premier Wittgenstein et l'atomisme logique, *etc.* Dreyfus montre également comment le poids de ces traditions a amené l'IA à se forger un paradigme dominant, lequel a méticuleusement étouffé les approches hétérodoxes qui faisaient pourtant l'économie de certains de ces présupposés[24].

Troisième partie. Dreyfus propose de dépasser la tradition philosophique et ouvre ainsi la voie à une « nouvelle IA ». Sa stratégie consiste à emprunter aux critiques de la tradition cartésienne – notamment du côté de la phénoménologie – des arguments qu'il oppose aux démarches computationnalistes. *E.g.*, la distinction entre *knowing-how* et *knowing-that* révèle les limites du modèle computationnel[25]. Le *knowing-that* (ou « *savoir-que* ») désigne notre aptitude à résoudre des problèmes rationnels de manière logique. Il nécessite une halte de l'esprit et une séquence d'attitudes propositionnelles (croyances, désirs, *etc.*). Le GPS de Newell & Simon implémente avec précision cette faculté épistémique de « haut-niveau ». Cependant, la phénoménologie rappelle qu'une seconde aptitude cognitive, le *knowing-how* (ou « *savoir-comment* »), distinct et irréductible au *knowing-that*, est responsable d'une grande part de notre activité quotidienne. Elle consiste en des processus continus, souvent inconscients, qui gèrent les facultés de « bas-niveau » telles que la perception, la motricité, les émotions, *etc.* Le *knowing-how*, contrairement à son analogue rationnel,

---

[23] T. Hobbes, *Of Man, Being the First Part of Leviathan* [1651], The Harvard Classics, vol. XXXIV, part. 5, New York : P. F. Collier & Son, 2001.

[24] C'est par exemple le cas du connexionnisme, véritablement écrasé par la recherche computationnaliste, alors qu'il répondait notamment aux critiques de Searle concernant l'impossibilité d'une sémantique non-arbitraire dans les systèmes symboliques. L. Dreyfus et S. E. Dreyfus, « Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint » [1988], *in* M. A. Boden éd., *The Philosophy of Artificial Intelligence*, Oxford, New York : Oxford University Press, 1990, p. 309-333.

[25] La discussion de Dreyfus se trouve principalement dans H. L. Dreyfus et S. E. Dreyfus, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*, Oxford : Blackwell, 1986. Les termes « *knowing-how* » et « *knowing-that* » sont introduits pour la première fois par Gilbert Ryle pour argumenter contre le mythe cartésien du dualisme. G. Ryle, *La Notion d'esprit* [*The Concept of mind*, 1949], traduit par S. Stern-Gillet, Paris : Éditions Payot & Rivages, 2005, p. 93-140. Cependant, la critique de Dreyfus est moins fondée sur les concepts de Ryle que sur la distinction heideggérienne entre « *Vorhandenheit* » et « *Zuhandenheit* » (resp. « être-sous-la-main » et « être-à-portée-de-la-main »).

n'utilise ni symbole, ni logique. De plus, il est sensible aux contextes : corporel, environnemental, social, *etc.* De fait, il est difficile (voire impossible) de modéliser chacun de ces contextes par un système symbolique approprié. Il est donc difficile (voire impossible) d'implémenter le *knowing-how* à l'aide d'un programme tel que le GPS. Cette faculté épistémique non-rationnelle échappe au computationnalisme.

Cependant, la critique dreyfusienne ne doit pas être entendue comme un rejet intégral du computationnalisme. Si on tire toutes les conséquences de son argumentation, il existe encore des domaines de l'intelligence pour lesquels les systèmes symboliques sont efficaces. Lorsqu'on cherche par exemple à résoudre un problème hautement rationnel, facilement formalisable, tel que celui du jeu d'échecs, les approches classiques peuvent encore faire leurs preuves – en témoigne la victoire de Deep Blue sur Garry K. Kasparov en mai 1997[26]. Cependant, la phénoménologie insiste sur le fait que la plus grande part de l'activité humaine, contrairement à ce qui y paraît, est de l'ordre du *knowing-how* : l'usage du langage, la motricité quotidienne, la reconnaissance et la classification de formes sont autant d'activités aux soubassements non-symboliques et non-rationnels. Si elles constituent les objectifs des spécialistes de l'IA, alors il faut trouver *une nouvelle voie*. Pour résumer :

- Le travail de Dreyfus est bien informé des recherches et des résultats de l'IA ;
- Il ne constitue pas une opposition de principe à l'IA, mais cible son paradigme dominant (le computationnalisme) et son utilisation inadéquate pour résoudre une large quantité de problèmes ;
- Il s'intéresse enfin à l'IA faible. L'enjeu de la critique ne porte pas sur l'ontologie des machines, mais sur leurs capacités concrètes de résolution.

Ce dernier point distingue fondamentalement les travaux de Searle et de Dreyfus. *What Computers Can't Do*, à partir de la position philosophique de l'auteur, opère une critique constructive des méthodes *pratiques* de l'IA. Les deux sections suivantes généralisent cette démarche pour définir les termes d'une collaboration véritable entre IA et philosophie.

[26] Cela-dit, même pour les grands joueurs d'échecs, une bonne part de la stratégie n'est ni explicite, ni rationnelle. La psychologie, le bluff, l'utilisation du temps forment un contexte global qui s'ajoute à la simple connaissance des positions des pièces et des règles de déplacement.

4. La philosophie au service de l'IA[27]

Avec le temps, le travail de Dreyfus est passé des mains des critiques à celles des praticiens. Les modèles phénoménologiques de la cognition, en réaction aux modèles de la tradition cartésienne, offrent l'opportunité d'un nouveau paradigme pour l'IA faible. Nous formulons ainsi la démarche de Dreyfus : « IA forte → IA faible ». Autrement dit, une théorie de l'esprit adéquate, si elle est correctement appliquée, donne de bons résultats en pratique. Ainsi, à partir des années 1980, suite aux nombreuses critiques du computationnalisme et aux sollicitations de la phénoménologie, l'IA connait une véritable crise fondationnelle. Des chercheurs en philosophie et en informatique œuvrent pour l'édification de nouveaux paradigmes. Le connexionnisme refait son apparition avec les débuts de l'Intelligence Artificielle Distribuée et de la Vie Artificielle. De nouveaux modèles voient le jour en robotique : *e.g.*, le modèle énactif de la cognition, la robotique incarnée, la robotique évolutionniste. Apparaissent également des modèles anti-représentationnalistes (systèmes réactifs, intelligence sans représentation, systèmes dynamiques, *etc.*). On parle aussi du « tournant pragmatique » de l'IA[28].

Afin de généraliser la démarche « IA forte → IA faible » inaugurée par Dreyfus, un autre exemple de collaboration est présenté dans cette section. L'objectif n'est plus ici la simulation de comportements intelligents, mais la simulation de « phénomènes émergents », c'est-à-dire de propriétés et processus globaux d'un système induits par les propriétés et processus locaux de ses parties. L'émergence distingue ainsi au moins deux niveaux de description. La difficulté réside dans la modélisation et simulation simultanées de ces deux niveaux au sein d'un même programme. Par exemple, comment simuler les dynamiques globales d'une ville ou d'un pays ? Comment rendre

---

[27] Le travail présenté dans cette section a fait l'objet d'une communication pour la plateforme *AFIA 2011* (Association Française pour l'Intelligence Artificielle). R. Lamarche-Perrin, « Conceptualisation de l'émergence : dynamiques microscopiques et analyse macroscopique des SMA *», in Plateforme AFIA 2011*, *atelier FUTURAMA*, 2011.

[28] Sur la nouvelle robotique et l'anti-représentationnalisme : R. A. Brooks, « Intelligence without representation ». *In Artificial Intelligence*, 1991, p. 139-159. Sur le connexionnisme et le modèle énactif de la cognition : F. J. Varela, E. Thompson et E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*, Cambridge, Massachusetts : MIT Press, 1991. Sur les systèmes dynamiques : T. van Gelder, « Dynamics and cognition », *in* J. Haugeland éd., *Mind Design II*, Bradford/MITP, 1996. Sur la notion de « tournant pragmatique » : A. K. Engel, « Directive Minds: How Dynamics Shapes Cognition », *in Enaction : Toward a New Paradigm for Cognitive Science*, MIT Press, 2010, p. 219-243. Pour un retour de Dreyfus sur ces nouvelles approches, une vingtaine d'années plus tard : H. L. Dreyfus, « Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian », *in Philosophical Psychology*, vol. 20, n°2, 2007, p. 247-268.

compte des relations entre la dynamique locale des individus et la taille de la ville, sa politique générale, son marché immobilier ? De quelle manière ces variables macroscopiques dépendent-elles des comportements microscopiques ? Ce type de problème est très répandu en IA[29]. Nous soutenons que la philosophie peut aider à clarifier la notion d'émergence et à résoudre certaines difficultés théoriques liées à la simulation des phénomènes macroscopiques. Dans ce qui suit, nous faisons l'analyse de la notion d'émergence telle qu'elle fut historiquement développée par la philosophie britannique au tournant du XIX^{e} siècle[30]. Les concepts et les enjeux du débat philosophique nous permettent d'élaborer une définition adéquate dans le cas des simulations informatiques. Ainsi, à partir de problématiques et de contraintes *philosophiques*, nous exprimons des problématiques et des contraintes d'ordre *méthodologique*.

Dualisme et monisme. Prenons un problème auquel sont confrontés scientifiques et philosophes : comment rendre compte des phénomènes du vivant ? Comment expliquer leurs spécificités ? Comment expliquer les différences essentielles qui existent entre un être inanimé et un être vivant ? Sommairement, au XIX^{e} siècle, deux positions s'affrontent : le *vitalisme* et le *mécanisme*. Le *vitalisme* est un « dualisme non-réductionniste » : il postule l'existence de deux substances (la matière inanimée et un principe de force vitale) pour rendre compte des deux catégories d'objets, la seconde substance ne pouvant être entièrement déterminée par la première. Autrement dit, le non-réductionnisme affirme que les lois de la biologie ne peuvent être « déduites de » ou « réduites aux » lois physico-chimiques. Le dualisme, parce qu'il multiplie les postulats d'existence, est *ontologiquement coûteux*. Le *mécanisme* au contraire fait une économie ontologique en affirmant que les phénomènes du vivant sont entièrement explicables par les lois de la physique et de la chimie. Cependant, en effectuant cette réduction, le monisme participe à l'élimination des sciences spéciales. Puisque le vocabulaire de la biologie peut être exprimé à partir de celui de la physique, son usage est rendu caduc. Dès lors, un seul mode de connaissance est privilégié : celui de la physique fondamentale. Il devient difficile de rendre compte et d'expliquer la distinction entre êtres inanimés et être vivants. Celle-ci est amenuisée, voire radicalement éliminée. On dira donc que ce « monisme éliminatif » est *épistémiquement faible*.

---

[29] Pour une analyse non-exhaustive des différentes conceptualisations de l'émergence en IA : R. Lamarche-Perrin, *op. cit.* et J. Deguet, Y. Demazeau et L. Magnin, « Element about the Emergence Issue : A Survey of Emergence Definitions », *in ComPlexUs*, vol. 3, 2006, p. 24-31.

[30] Pour une analyse historique détaillée de cette controverse, à laquelle participent notamment J. S. Mill, C. D. Broad et S. Alexander, voir : T. O'Connor et H. Y. Wong, « Emergent Properties », *in Stanford Encyclopedia of Philosophy*, 2006, [http://plato.stanford.edu/entries/properties-emergent/, mis en ligne le 24 sept. 2002, révisé le 23 oct. 2006, consulté le 1 mai 2011].

Émergence épistémique. La position émergentiste consiste à trouver une voie moyenne entre vitalisme et mécanisme, entre « dualisme non-réductionniste » et « monisme éliminatif ». Comment défendre une position à la fois économe sur le plan ontologique et forte sur le plan épistémique ? Comment concevoir un « monisme non-éliminatif » ? La notion d'*émergence épistémique*[31] répond justement à ces attentes. Elle soutient que le principe de force vitale, même s'il est en principe réductible à la matière inanimée, constitue une abstraction utile au scientifique. La distinction entre êtres inanimés et êtres vivants n'est pas d'ordre ontologique : elle est « dans l'œil du scientifique », elle est *épistémique*. Cette position est compatible avec le monisme puisque les lois de la biologie peuvent *en principe* être réduites à celles de la physique. Elle est de plus non-éliminative, puisque la biologie présente *en pratique* des lois et des modèles utiles et nécessaires à la compréhension des phénomènes complexes.

Pour appliquer la notion d'émergence épistémique aux simulations informatiques, nous exploitons l'analogie suivante : ce qui est *ontologique* pour un système, c'est ce qui est relatif à sa conception, *i.e.* en amont de son exécution (le modèle du programme, son code source, son implémentation matérielle, *etc.*) ; ce qui est *épistémique*, c'est ce qui est relatif à l'analyse *a posteriori* de l'exécution du programme (les outils d'observation, de description et d'analyse). L'ontologie relève du système *per se* ; l'épistémologie se rapporte au système relativement à *un point de vue donné*. Dans ce qui suit, le concept d'émergence épistémique est traduit sur la base de cette analogie. Deux contraintes méthodologiques sont engendrées : le *monisme microscopique* et le *non-éliminativisme*.

Monisme microscopique. Par analogie, qu'est-ce qu'une simulation « dualiste » en informatique ? Il s'agit d'un programme dont la conception comprend au moins deux niveaux de modélisation (niveau microscopique *et* niveau macroscopique). Pour la simulation des dynamiques urbaines, par exemple, le programme modélise à la fois le comportement des individus et celui des variables globales. Lors de l'exécution, ces deux modèles sont synchronisés pour simuler une activité urbaine multi-échelle[32]. Au contraire, une approche « moniste » limite la conception du système à

[31] On parle d'« émergence épistémique » par opposition à l'« émergence ontologique », notion que l'on associe parfois au dualisme. Pour une analyse conceptuelle plus détaillée : T. O'Connor et H. Y. Wong, *op. cit.*

[32] Les « systèmes multi-modèles » sont un bon exemple de systèmes dualistes. Plusieurs niveaux de modélisations sont synchronisés et maintenus cohérents pour simuler un système à plusieurs échelles. J. Gil-Quijano, G. Hutzler et T. Louail, « Accroche-toi au niveau, j'enlève l'échelle. Éléments d'analyse des aspects multiniveaux dans la simulation à base d'agents », *in Revue d'Intelligence Artificielle*, vol. 24, 2010, p. 625-648.

ses parties microscopiques. Le niveau macroscopique n'est pas modélisé *a priori*, mais élaboré *a posteriori* de l'exécution, lors de l'analyse[33]. Les phénomènes émergents, tels que le marché immobilier, sont alors conçus comme des épiphénomènes, *i.e.* des phénomènes sans puissance causale, qui existent relativement à un observateur ou à un processus d'abstraction. En interdisant la conception macroscopique des systèmes, le *monisme microscopique* rend possible une véritable « approche *bottom-up* » : il ne s'agit pas seulement de *simuler* les phénomènes émergents (en concevant un modèle *a priori* des variables globales), mais de les *émuler* véritablement, c'est-à-dire de reproduire leurs dynamiques émergentes dans leur intégralité à partir de leurs fondements microscopiques.

Non-éliminativisme. Le monisme méthodologique ne doit pas tomber dans les travers éliminativistes de son analogue philosophique. Ainsi, les méthodes d'analyse ne doivent pas se limiter à une description purement microscopique de l'exécution. Au contraire, nous souhaitons multiplier les points de vue sur le système[34]. En informatique, l'introduction de « modèles de l'observateur » permet d'engendrer des descriptions macroscopiques variées[35]. La tâche du scientifique n'est pas d'observer des phénomènes macroscopiques *per se*, mais de construire les abstractions qui seront utiles à la compréhension globale des systèmes. Ces abstractions *sont* les phénomènes émergents ; leur pertinence est toujours évaluée *en contexte*, c'est-à-dire relativement aux besoins particuliers de l'analyse et aux objectifs scientifiques préalablement fixés. Il n'y a pas de

---

Les « systèmes à tableau-noirs » (*blackboard systems*) présentent un autre cas de dualisme. Des entités macroscopiques, en interaction avec les entités microscopiques du système, sont conçues pour implémenter des variables globales. R.K. Sawyer, « Simulating Emergence and Downward Causation in Small Groups », *in Multi-Agent-Based Simulation*, vol. 1979, 2001, p. 49-67.

[33] Les « systèmes par auto-organisation » sont des systèmes monistes dans la mesure où leurs fonctionnalités émergentes ne sont pas explicitées lors de la conception. Celles-ci reposent uniquement sur l'implémentation de fonctions locales. G. Picard, *Méthodologie de développement de SMA adaptatifs et conception de logiciels à fonctionnalité émergente*, Thèse de doctorat de l'Université Paul Sabatier de Toulouse III, 2004.

[34] Par exemple, les travaux qui définissent l'émergence en fonction de qualités intrinsèques aux systèmes sont éliminativistes. Ils n'autorisent qu'un seul niveau de description des phénomènes. V. Darley, « Emergent Phenomena and Complexity », *in Artificial Life*, vol. 4, 1994, p. 411-416. M. Bedau, « Weak Emergence », *in Philosophical Perspectives*, vol. 11, 1997, p. 379-399.

[35] Par exemple, la conceptualisation de Bonabeau & Dessalles est non-éliminative. Elle définit l'émergence relativement à des hiérarchies de détecteurs. É. Bonabeau et J.-L. Dessalles, « Detection and Emergence », *in Intellectica*, vol. 25, n°2, 1997, p. 85-94.

« bons phénomènes émergents » en-soi, mais seulement en fonction de ce que l'on veut en faire. Le *non-éliminativisme* favorise ainsi une conception pragmatiste des phénomènes émergents.

Ces deux contraintes (*monisme microscopique* et *non-éliminativisme*) établissent une approche cohérente pour simuler des phénomènes émergents. Elles bénéficient d'une conceptualisation solide de l'émergence, héritée de la philosophie, et importent ainsi l'épiphénoménisme et le pragmatisme en informatique (IA forte → IA faible). En pratique, elles permettent également de trier les formalisations de l'émergence et de travailler à partir de celles qui satisfont les exigences des spécialistes. Il est important de préciser cependant qu'il s'agit d'exigences *méthodologiques*. Elles n'imposent aucune contrainte *de principe* aux systèmes de simulation. Selon la critique de Dreyfus, le computationnalisme reste pertinent pour résoudre des problèmes particuliers bien délimités. De la même manière, les simulations dualistes ou éliminativistes peuvent se révéler efficaces en fonction des contraintes et des objectifs techniques de la simulation. Cependant, nous affirmons que, dans le cas de la simulation de systèmes distribués, complexes et de très grande taille, ces deux contraintes se révèlent nécessaires *en pratique*[36].

5. L'IA au service de la philosophie

Une autre forme de collaboration entre IA et philosophie peut être dérivée du travail de Dreyfus. Sa critique de la mise en application du computationnalisme cristallise un point de désaccord avec la tradition philosophique. Elle a notamment permis de systématiser l'opposition au « Théâtre Cartésien » à partir de ses conséquences pratiques. Considérons la contraposée de l'implication précédente : « non IA faible → non IA forte ». Cela signifie que les échecs de l'IA faible témoignent de l'inadéquation des présupposés philosophiques dont elle hérite. *E.g.*, l'échec du GPS peut être interprété comme l'indice empirique des erreurs de la tradition cartésienne. Autrement dit, lorsque le computationnalisme échoue, son cadre philosophique est remis en cause.

Grâce au concours de l'IA, la philosophie de l'esprit peut ainsi devenir une *philosophie expérimentale*. L'IA est alors envisagée comme un banc d'essais, un atelier où le philosophe implémente ses modèles à partir de robots et de programmes. Il y observe ensuite le résultat de ses

[36] Pour plus de détails sur l'importance pratique de ces contraintes méthodologiques : R. Lamarche-Perrin, Y. Demazeau et J.-M. Vincent, « Observation macroscopique et émergence dans les SMA de très grande taille », *in* J.-P. Sansonnet et É. Adam éd., *19e Journées Francophones sur les Systèmes Multi-Agents,* Valenciennes : Cépaduès, oct. 2011, p. 53-62.

théories mises en pratique[37]. De cette manière, l'IA permet d'évaluer et de falsifier les modèles de l'esprit par des méthodes empiriques. On comprend enfin le renversement amorcé en introduction par Daniel Andler et qui fait de l'IA une véritable *science de la philosophie*[38]. Or, ce n'est pas la première fois que l'étude de machines concrètes illustre certaines options philosophiques. L'analyse comparée d'un régulateur symbolique et d'un régulateur dynamique amène par exemple Tim van Gelder à discuter de la pertinence philosophique des modèles qui leur sont associés[39] : de quel genre de processus mécaniques la cognition est-elle responsable ? À quel type de régulateur l'esprit est-il similaire ? De la même manière, les programmes de l'IA peuvent éclairer, à l'avenir, les modèles philosophiques et ainsi peser dans la balance des controverses.

[37] Inman Harvey a pour cela une formule amusante : « faire de la philosophie de l'esprit à l'aide d'un tournevis. » I. Harvey, « Robotics: Philosophy of Mind Using a Screwdriver », *in Evolutionary Robotics: From Intelligent Robots to Artificial Life*, vol. III, 2000, p. 207–230.

[38] H. L. Dreyfus, *op. cit.*, p. XIV.

[39] T. van Gelder, « Dynamics and cognition », *in* J. Haugeland éd., *Mind Design II*, Bradford/MITP, 1996.

Bibliographie

- Anspach, M.R., Varela, F.J. 1992. « Le système immunitaire : un "soi" cognitif autonome. » *In* Andler, D. (éd.). 2004. *Introduction aux sciences cognitives* (édition augmentée). Paris : Gallimard, p. 585-605.

- Bedau, M.A. 1997. « Weak Emergence. » *Philosophical Perspectives*, vol. 11, p. 375-399.

- Beer, R.D., Quinn, A.D., Chiel, H.J., Ritzmann, R.E. 1997. « Biologically Inspired Approaches to Robotics: What can we learn from insects? » *Communications of the ACM*, vol. 40, n°3, p. 30-38.

- Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press.

- Boden, M.A. 1995. « AI's Half-Century. » *AI Magazine*, vol. 16, n°4.

- Bonabeau, É., Dessalles, J.-L. 1997. « Detection and Emergence. » *Intellectica*, vol. 25, n°2, p. 85-94.

- Brooks, R.A. 1991. « Intelligence without representation. » *Artificial Intelligence*, vol. 47, p. 139-159.

- Buchanan, B.G., Shortliffe, E.H. 1984. *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA : Addison-Wesley.

- Buchanan, B.G., Sutherland, G.L., Feigenbaum, E.A. 1969. « Heuristic DENDRAL: a Program for Generating Explanatory Hypotheses in Organic Chemistry. » *In* Meltzer, B. (éd.), Michie, D. (éd.). 1969. *Machine Intelligence*, vol. 4. Edinburgh University Press, p. 209-254.

- Chaib-Draa, B., Jarras, I., Moulin, B. 2001. « Systèmes multiagents : principes généraux et applications. » *In* Briot, J.-P. (éd.), Demazeau, Y. (éd.). 2001. *Principes et architecture des systèmes multiagents*. Paris : Hermes, p. 27-70.

- Chalmers, D. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

- Clark, A., Toribio, J. 1994. « Doing Without Representing? » *Synthese: Connectionism and the Frontiers of Artificial Intelligence*, vol. 101, n°3, p. 401-431.

- Darley, V. 1994. « Emergent Phenomena and Complexity. » *Artificial Life*, vol. 4, p. 411-416.

- David, D., Payet, D., Courdier, R. 2011. « Réification de zones urbaines émergentes dans un modèle simulant l'évolution de la population à La Réunion. » *Journées Francophones sur les Systèmes Multi-Agents* (JFSMA'11)*,* oct. 2011, Valenciennes : Cépaduès, p. 63-72.

- Deguet, J., Demazeau, Y., Magnin, L. 2006. « Element about the Emergence Issue: A Survey of Emergence Definitions. » *ComPlexUs*, vol. 3, p. 24-31.

- Dennett, D.C. 1991. *Consciousness Explained*. New York : Hachette Book.

- Denton, D. 1993. *L'Émergence de la conscience : de l'animal à l'homme*. [*The Pinnacle of life. Consciousness and Self-Awareness in Humans and Animals*.] Mourlon, J.-P. (trad.). Paris : Flammarion, 1995.

- Dreyfus, H.L. 1979. *Intelligence Artificielle : mythes et limites*. [*What Computers Can't Do: The Limits of Artificial Intelligence*, 2nd ed.] Vassallo-Villaneau, R.-M. (trad.), Andler, D. (pref.), Perriault, J. (pref.). Paris : Flammarion, 1984.

- Dreyfus, H.L. 2007. « Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian. » *Philosophical Psychology*, vol. 20, n°2, p. 247-268.

- Dreyfus, H.L., Dreyfus, S.E. 1988. « Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 309-333.

- Engel, A.K. 2010. « Directive Minds: How Dynamics Shapes Cognition. » *In* Stewart, J.R. (éd.), Gapenne, O. (éd.), Di Paolo, A.E. (éd.). 2010. *Enaction: Toward a New Paradigm for Cognitive Science*. Cambridge, MA: MIT Press, p. 219-243.

- Ferber, J. 1995. *Les Systèmes multi-agents : vers une intelligence collective*. Paris : InterEditions.

- Fisette, D. (éd.), Poirier, P. (éd.). 2003. *Philosophie de l'esprit : problèmes et perspectives*. Paris : Vrin.

- Fodor, J.A. 1975. *The Language of Thought*. Cambridge, MA : Harvard University Press.

- Forrest, S. 1990. « Emergent computation: Self-organizing Collective and Cooperative Phenomena in Natural and Artificial Computing Networks. » *Physica D: Nonlinear Phenomena*, vol. 42, n°1-3, p. 1-11.

- Gil-Quijano, J., Hutzler, G., Louail, T. 2010. « Accroche-toi au niveau, j'enlève l'échelle. Éléments d'analyse des aspects multiniveaux dans la simulation à base d'agents. » *Revue d'Intelligence Artificielle*, vol. 24, p. 625-648.

- Harnad, S. 1990. « The Symbol Grounding Problem. » *Physica D*, vol. 42, p. 335-346.

- Harnad, S. 2001. « Minds, Machine and Seale II: What's Wrong and Right About Searle's Chinese Room Argument? » *In* Bishop, M. (éd.), Preston, J. (éd.). 2001. *Essays on Searle's Chinese Room Argument*. Oxford University Press.

- Harnad, S. 2003. « Can a Machine Be Conscious? How? » *Journal of Consciousness Studies*, vol. 10, n°4, p. 67-75.

- Harvey, I. 2000. « Robotics: Philosophy of Mind Using a Screwdriver. » *Evolutionary Robotics: From Intelligent Robots to Artificial Life*, vol. III, p. 207-230.

- Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*, Cambridge, MA : MIT Press.

- Held, R., Hein, A. 1958. « Adaptation of disarranged hand-eye coordination contingent upon re-afferent simulation. » *Perceptual-Motor Skills*, vol. 8, p. 87-90.

- Hempel, C. G. 1935. « L'analyse logique de la psychologie. » [« The Logical Analysis of Psychology. »] *In* Fisette, D. (éd.), Poirier, P. (éd.). 2002. *Philosophie de l'esprit : psychologie du sens commun et sciences de l'esprit*. Paris : Vrin, p. 197-215.

- Hyslop, A. 2009. « Other Minds. » *In Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/other-minds/, mis en ligne le 6 oct. 2005, révisé le 23 nov. 2009, consulté le 19 juillet 2011.

- Jackson, F. 1982. « Epiphenomenal Qualia. » *Philosophical Quarterly*, vol. 32, p. 127-136.

- Kaplan, F., Oudeyer, P.-Y. 2008. « Le Corps comme variable expérimentale. » *Revue philosophique de la France et de l'étranger*, vol. 133, n°3.

- Kirk, R., Squires, J.E. 1974. « Zombies v. Materialists. » *The Aristotelian Society*, vol. 48, p. 135-163.

- Kistler, M. 2007. « La Réduction, l'émergence, l'unité de la science et les niveaux de réalité. » *Matière Première*, vol. 2, p. 67-97.

- Kubík, A. 2003. « Toward a Formalization of Emergence. » *Artificial Life*, vol. 9, p. 41-65.

- Kolmogorov, A.N. 1965. « Three Approaches to the Quantitative Definition of Information. » *Problems Information Transmission*, vol. 1, n°1, p. 1-7.

- Lamarche-Perrin, R. 2010. *Le Test de Turing pour évaluer les théories de l'esprit*. Mémoire de Master, Kistler, M. (dir.). Grenoble : Université Pierre-Mendès-France, sept. 2010.

- Lamarche-Perrin, R. 2011a. « Conceptualisation de l'émergence : dynamiques microscopiques et analyse macroscopique des SMA. » *Atelier Futur des Agents et des Multi-Agents (FUTURAMA'11)*. Chambéry : Plateforme AFIA 2011, mai 2011.

- Lamarche-Perrin, R. 2011b. « Des collaborations possibles entre Intelligence Artificielle et philosophie de l'esprit. » *Colloque des doctorants et jeunes chercheurs du groupe de recherche PLC : Repenser les rapports entre science(s) et philosophie*. Grenoble : Philosophie, Langages et Cognition, juin 2011. En cours d'édition.

- Lamarche-Perrin, R., Demazeau, Y., Vincent, J.-M. 2011. « Observation macroscopique et émergence dans les SMA de très grande taille. » *Journées Francophones sur les Systèmes Multi-Agents* (JFSMA'11), oct. 2011, Valenciennes : Cépaduès, p. 53-62.

- Levesque, H.J. 2009. « Is It Enough to Get the Behavior Right? » *International Joint Conference on Artificial Intelligence (IJCAI'09)*, p. 1439-1444.

- Lewis, D. 1978. « Douleur de fou et douleur de martien. » [« Mad Pain and Martian Pain. »]. *In* Fisette, D. (éd.), Poirier, P. (éd.). 2002. *Philosophie de l'esprit : psychologie du sens commun et sciences de l'esprit*. Paris : Vrin, p. 289-306.

- McCarthy, J. 1995. « What has AI in Common with Philosophy? » *International Joint Conference on Artificial Intelligence (IJCAI'95)*, vol. 2, p. 2041-2042.

- McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E. 1955. « A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. » *AI Magazine*, 2006, vol. 27, n°4, p. 12-14.

- McCulloch, W.S., Pitts, W.H. 1943. « A Logical Calculus of the Ideas Immanent in Nervous Activity. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 22-39.

- McGeer, T. 1990. « Passive dynamic walking. » *International Journal of Robotics Research (IJRR'90)*, vol. 9, n°2, p. 62-82.

- Moravec, H. 1988. *Mind Children*. Cambridge, MA : Harvard University Press.

- Newell, A., Shaw, J.C., Simon, H.A. 1959. « Report on a general problem-solving program. » *International Conference on Information Processing*, p. 256-264.

- Newell, A., Simon, H.A. 1958. « Heuristic Problem Solving: The Next Advance in Operations Research. » *Operations Research*, vol. 6, n°1, p. 1-10.

- Newell, A., Simon, H.A. 1963. « GPS, A Program that Simulates Human Thought. » *In* Feigenbaum, E.A. (éd.), Feldman, J. (éd.). 1995. *Computers and Thought*. Cambridge, MA : MIT Press, p. 279-293.

- Newell, A., Simon, H.A. 1976. « Computer Science as Empirical Inquiry: Symbols and Search. » *Communications of the ACM*, vol. 19, n°3, p. 113-126.

- O'Connor, T., Wong, H.Y. 2006. « Emergent Properties. » *In Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/properties-emergent/, mis en ligne le 24 sept. 2002, révisé le 23 oct. 2006, consulté le 1 mai 2011.

- Picard, G. 2004. *Méthodologie de développement de SMA adaptatifs et conception de logiciels à fonctionnalité émergente*. Thèse de doctorat, Gleizes, M.-P. (dir.). Toulouse : Université Paul Sabatier, déc. 2004.

- Ronald, E.M.A., Sipper, M. 2001. « Surprise versus unsurprise: Implications of emergence in robotics. » *Robotics and Autonomous Systems*, vol. 37, n°1, p. 19-24.

- Rorty, R. 1980. *L'homme spéculaire*. [*Philosophy and the Mirror of Nature*.] Marchaisse, T. (trad.). Paris : Seuil, 1990.

- Russell, S.J., Norvig, P. 2003. *Artificial Intelligence: A Modern Approach,* 3rd ed. Upper Saddle River, NJ : Prentice Hall.

- Ryle, G. 1949. *La Notion d'esprit*. [*The Concept of Mind*.] Stern-Gillet, S. (trad.), Tanney, J. (pref.). Paris : Payot & Rivages, 2005.

- Sawyer, R.K. 2001. « Simulating Emergence and Downward Causation in Small Groups. » *Multi-Agent-Based Simulation*, vol. 1979, p. 49-67.

- Searle, J.R. 1980. « Minds, Brains, and Programs. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 67-88.

- Searle, J.R. 1984. *Du cerveau au savoir : conférences Reith 1984 de la BBC*. [*Minds, Brains and Science: The 1984 Reith Lectures.*] Chaleyssin, C. (trad.). Paris : Hermann, 1985.

- Searle, J.R. 1992. *The Rediscovery of the Mind*. Cambridge, MA : MIT Press.

- Searle, J.R. 1997. *The Mystery of Consciousness*. New York, NY : The New York Review of Books.

- Shannon, C. 1950. « Programming a Computer for Playing Chess. » *Philosophical Magazine*, vol. 41, n°314, p. 256-275.

- Smolensky, P. 1988. « Le traitement approprié du connexionnisme. » [« On the Proper Treatment of Connectionism. »] *In* Fisette, D. (éd.), Poirier, P. (éd.). 2003. *Philosophie de l'esprit : problèmes et perspectives*. Paris : Vrin, p. 197-215.

- Stephan, A. 1999. « Varieties of Emergentism. » *Evolution and Cognition*, vol. 5, n°1, p. 49-59.

- Turing, A.M. 1950. « Computing Machinery and Intelligence. » *In* Boden, M.A. (éd.). 1990. *The Philosophy of Artificial Intelligence*. Oxford University Press, p. 40-66.

- van de Vijver, G. 1997. « Émergence et explication. » *Intellectica*, vol. 25, n°2, p. 7-23.

- van Gelder, T. 1995. « What might Cognition be if not Computation? » *Journal of Philosophy*, vol. 92, n°7, p. 345-381.

- van Gelder, T. 1998. « Dynamique et cognition. » [« Dynamics and Cognition. »] Lapointe, S. (trad.). *In* Fisette, D. (éd.), Poirier, P. (éd.). 2003. *Philosophie de l'esprit : problèmes et perspectives*. Paris : Vrin, p. 329-369.

- Varela, F.J. 1988. *Invitation aux sciences cognitives*. [*Cognitive Science. A Cartography of Current Ideas*.] Lavoie, P. (trad.). Paris : Seuil, 1996.

- Varela, F.J., Thompson, E.T., Rosch, E. 1991. *L'Inscription corporelle de l'esprit : sciences cognitives et expériences humaine*. [*The Embodied Mind: Cognitive Science and Human Experience*.] Havelange, V. (trad.). Paris : Seuil, 1993.

- Watson, J.B. 1913. « Psychology as the Behaviorist Views it. » *Psychological Review*, vol. 20, p. 158-177.

- Wooldridge, M. 2009. *An Introduction to MultiAgent Systems* (2nd ed.). Chichester, WS : John Wiley & Sons.

Index

Des collaborations possibles entre philosophie et Intelligence Artificielle

Robin Lamarche-Perrin

Résumé

Ce mémoire s'intéresse aux collaborations possibles entre Intelligence Artificielle et philosophie. Il montre que les deux disciplines peuvent partager des objets, des théories et des résultats pour apprendre l'une de l'autre. La stratégie de ce mémoire consiste à expliciter des relations épistémologiques entre les problématiques propres aux deux disciplines (« IA faible » et « IA forte »), afin de définir des modes de collaboration sur le plan disciplinaire.

La deuxième partie de ce mémoire présente les travaux de philosophes et de spécialistes de l'IA, depuis les débuts de l'Intelligence Artificielle jusqu'aux années 80. Elle expose les démarches collaboratives exploitées par ces chercheurs, de manière implicite ou explicite. La troisième partie présente des travaux où la philosophie sert de socle conceptuel à l'Intelligence Artificielle, notamment en ce qui concerne la simulation de phénomènes émergents. La quatrième partie réalise un renversement des relations classiques entre les deux disciplines. C'est au tour de l'Intelligence Artificielle de se mettre au service de la philosophie, en formulant de nouvelles hypothèses de recherche ou en testant les théories philosophiques à partir de cas concrets.

Ce mémoire, enfin, espère œuvrer pour le rapprochement des deux disciplines et ainsi encourager philosophes et spécialistes de l'IA à collaborer sur les sujets qui leurs sont chers.

Mots-clés

Philosophie de l'esprit, Intelligence Artificielle, sciences cognitives, épistémologie, collaboration scientifique, interdisciplinarité.