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

Robin Lamarche-Perrin

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 qui s'intéresse donc, 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.* ?).

[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.

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 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

[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.

[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'ensembles 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.

peuvent offrir une base saine au travail empirique et comment celui-ci peut en retour aider la philosophie à évaluer ses théories.

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 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].

[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.

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-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.

[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.

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 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 20 années de recherches, entre 1957 et 1977.

---

[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.

[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 liés 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 soldé par un échec et finalement abandonné.

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 présupposés héritent d'une longue tradition philosophique. On peut citer par exemple Thomas

[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.

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, n'utilise ni symbole, ni logique. De plus, il est sensible aux contextes : contexte 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.

[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 introduit 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 »).

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-rationnelles. 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.

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

[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.

[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.

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 simultanée 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 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*.

[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.

[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].

Dualisme et monisme. Prenons un problème auxquels 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, au motif d'une telle réduction, le monisme conduit 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*.

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 être 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 (*e.g.*, le modèle du programme, son code source, son

[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.*

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 ceux 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 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

[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. 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.

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 « 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].

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

[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.

[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.

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 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 ? L'esprit est similaire à quel type de régulateur ? 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.