Joëlle Proust

L'intelligence artificielle comme philosophie

Ce qui, pour le philosophe, fait la difficulté de l'assimilation et de l'évaluation épistémologique du travail en intelligence artificielle ne réside pas seulement dans le caractère relativement récent de la discipline et dans l'abondance et la diversité des travaux et des méthodes. La difficulté provient, plus essentiellement, du statut double qui est le sien. Issue de l'informatique et de l'ingénierie, l'intelligence artificielle se donne pour objectif non seulement d'« *étudier* l'intelligence en utilisant les idées et les méthodes du calcul », mais de « *fabriquer des machines* qui font des choses qui, accomplies par l'homme, demanderaient de l'intelligence »[1]. L'intelligence artificielle, telle l'amphisbène serpentiforme de la légende, est ainsi pourvue de deux têtes qui n'en font chacune qu'à la sienne ; une tête spéculative, théorique, visant à comprendre la nature de l'intelligence, et encline à poser des questions de fondement ; une tête technologique, voire technocratique, qui n'est pas insensible aux vastes plans subventionnés par l'État, qui se détermine en fonction d'impératifs pragmatiques et vise moins à poser des problèmes qu'à trouver des solutions. La technologie de la connaissance s'intéresse avant tout aux résultats. L'intelligence artificielle théorique, dite aussi « abstraite », tente de dégager les conditions générales auxquelles un système peut être dit « intelligent ». Admettons provisoirement que les deux têtes du serpent soient nettement distinctes, l'une ignorant les impératifs de l'autre. La tête spéculative se pose donc une question fondamentale, une question qu'elle reprend, comme le remarque Dan Dennett, à la tradition kantienne : « Comment la connaissance en général est-elle possible ? » Il s'agit là, comme le rappelle Dennett, d'une question « *top-down* », c'est-à-dire régressive. Kant part de l'existence d'une connaissance de la nature capable d'énoncer des lois objectives, c'est-à-dire universelles et nécessaires ; pour pouvoir développer une métaphysique scientifique, le philosophe doit rechercher quelles conditions *a priori* fondent la nécessité dans les autres sciences.

L'intelligence artificielle, à dire vrai, n'a pas le projet kantien de fonder le caractère objectif des sciences de la nature, ni de démontrer que la recherche de ce fondement coïncide avec l'élaboration

1. La première citation est de P. H. Winston et Mike Brady, dans *Artificial Intelligence, an M.I. T. Perspective*, éd. par P. H. Winston et R. H. Brown, Cambridge, The M.I.T. Press, 1979, p. IX ; la seconde est de Marvin Minsky, *Semantic Information Processing*, Cambridge, The M.I.T. Press, 1968, p.V.

Joëlle Proust est l'auteur de *Questions de forme*, Paris, Fayard, 1986. Elle prépare un ouvrage sur l'intelligence artificielle et la philosophie dont le présent article est tiré.

Cet article est paru en novembre-décembre 1987 dans le n° 47 du *Débat* (pp. 88 à 102).

d'une nouvelle science, la métaphysique comme science des pouvoirs *a priori* du sujet de connaissance. En ce sens, on ne peut identifier, sans autre forme de procès, l'entreprise transcendantale kantienne et la recherche en I.A. En dépit de ce genre de divergences qu'il conviendra de préciser, il n'en reste pas moins une analogie générale de la démarche de la philosophie critique et de l'I.A. dans sa version classique, telle qu'elle a été développée dans les travaux de Newell et Simon. Dans des travaux récents, Newell tente de mettre à jour les hypothèses qui ont orienté implicitement ses recherches empiriques antérieures. Il se donne l'objectif de découvrir les caractéristiques les plus générales qu'un système doit posséder pour pouvoir être dit « intelligent ». Ces caractéristiques sont découvertes, comme chez Kant, régressivement, c'est-à-dire à partir du produit en direction de ses conditions de possibilité. Les deux types d'interrogation ont encore ceci de commun qu'elles recherchent des conditions formelles indépendantes d'un contenu donné d'expérience ou d'une région particulière de l'enquête sur les faits. De ce point de vue, elles ont l'une comme l'autre l'objectif d'atteindre des conditions *a priori* rendant possibles non seulement la connaissance, mais la forme générale des objets de cette connaissance.

Dans ce qui suit, je me propose de mettre à l'épreuve l'analogie qui apparaît à cette étape préliminaire de l'enquête et de clarifier le statut respectif de la philosophie de Kant et de l'I.A. relativement à la question de ce qui rend possibles (respectivement) la connaissance et le comportement intelligent. Hubert Dreyfus est l'un des premiers à avoir suggéré que la philosophie de Kant avait préparé le terrain à l'intelligence artificielle. Le présupposé d'après lequel « toute conduite ordonnée est nécessairement régie par des règles » aurait selon lui donné naissance à l'idée qu'un programme, c'est-à-dire une suite d'instructions hiérarchisées, puisse modéliser adéquatement le comportement cognitif humain. Notons que tandis que Dennett considère que l'I.A. reprend la question fondamentale de la philosophie transcendantale de Kant (à quelles conditions une connaissance est-elle possible ?), Dreyfus choisit d'esquisser l'analogie entre les deux types de recherches au niveau de la réponse spécifique qui est apportée de part et d'autre, en termes de règles *a priori* et de programme.

Dennett et Dreyfus donnent l'exemple de deux manières de penser le rapport possible entre philosophie et intelligence artificielle. Le propos de Dreyfus est de ramener l'I.A. aux dimensions d'une entreprise qu'on pourrait dire quasiment « idéologique », c'est-à-dire conditionnée par des présupposés inadéquats sur la nature de son objet. L'intervention du philosophe consiste de ce point de vue à mettre au jour les thèses – proprement philosophiques – qui organisent le jeu de langage appelé I.A. pour en montrer les limites sinon l'irrémédiable incohérence. Celui de Dennett est en revanche œcuménique : il consiste à supposer qu'un certain nombre de questions « épistémologiques » fondamentales se posent, de manière incontournable, au psychologue, au philosophe aussi bien qu'à l'informaticien converti à l'I.A. La discussion systématique, menée dans ce qui suit, de la pertinence d'une comparaison entre philosophie kantienne et intelligence artificielle, devrait permettre de clarifier le mode d'intervention de ces deux types de lectures philosophiques de l'I.A., et peut-être suggérer de nouvelles modalités de la communication entre philosophie et intelligence artificielle.

Questions de méthode.

L'idée de « comparer » un ensemble de recherches en I.A. avec un ensemble de textes philosophiques n'implique pas qu'il s'agisse de réflexions homogènes. Nous avons vu plus haut que l'I.A. dite « abstraite » pose comme la philosophie des questions générales « régressives » *(top-down)*. Cepen-

dant, tandis que la philosophie fournit sa réponse en termes conceptuels généraux, c'est-à-dire par l'intervention d'un certain nombre de thèses interdépendantes, voire, dans le cas limite d'une réponse complètement circonstanciée, par l'édification d'un système, l'intelligence artificielle répond en construisant « *in concreto* » un objet, nommé lui aussi « système », mais en un sens tout différent. Le « système » particulier qui résulte typiquement du travail en I.A. consiste en une unité de traitement de l'information capable d'exécuter un programme et d'obtenir, à partir de données d'entrée, un certain résultat. En tant qu'il livre une réponse à la question posée plus haut, ce système particulier instancie implicitement une théorie, en sorte que seules certaines de ses caractéristiques sont capitales du point de vue de l'I.A. abstraite : ce sont celles qui, selon les termes de Dennett, « sont nécessaires non seulement à un système particulier, mais à tout système de ce genre » (*Brain-storms*, p. 112), Kant avait déjà pensé à cette opposition, à propos de la manière dont travaillent respectivement le philosophe et le mathématicien : le philosophe, disait-il, « s'en tient à des concepts généraux ». La philosophie se meut dans les abstractions ; c'est une « connaissance rationnelle par concepts ». De son côté, l'intelligence artificielle ressemble plus à la mathématique telle que la concevait Kant. C'est une discipline qui examine son objet *in concreto*, c'est-à-dire en se donnant une « intuition », soit ici un système particulier, présentant les caractéristiques générales de la solution. L'I.A. considère donc « le général dans le particulier », en laissant un certain nombre de déterminations de côté dans cet examen. Certaines des caractéristiques du système seront en effet manifestement inessentielles, comme, par exemple, les caractéristiques propres à telle manière de câbler, etc. Seulement ce en quoi l'I.A. se distingue des mathématiques vues par Kant, c'est que, faute de disposer encore d'une théorie parfaitement maîtrisée, elle doit se livrer à un travail critique à partir des objets ainsi élaborés afin de découvrir celles de leurs caractéristiques qui sont des traits nécessaires de la cognition.

L'une des difficultés posées par une réflexion comparative entre philosophie et intelligence artificielle consiste donc dans la différence du medium de recherche. Les contraintes qui portent sur les réponses de l'I.A. sont évidemment non seulement celles de cohérence et d'adéquation que doivent respecter, de manière générale, les constructions conceptuelles. S'y ajoutent les exigences de complétude et de cohérence *techniques*, en entendant par là la nécessité d'expliciter le programme dans ses moindres détails et de supprimer l'absence d'interférences indésirables entre les instructions. Quoique cette différence soit souvent masquée par le caractère apparemment purement conceptuel de nombreux textes d'I.A., elle n'en demeure pas moins toujours présente, ne serait-ce qu'à titre virtuel. Réciproquement, une hypothèse « philosophique » ne peut recevoir de sens en I.A. que s'il lui correspond un choix entre des procédures concurrentes de conception de programme ou d'implémentation.

Système symbolique physique et sujet transcendantal.

Ces remarques méthodologiques préliminaires étant faites, revenons à la question fondamentale dont toute l'I.A. procède : « Comment la connaissance est-elle *en général* possible ? » Ce qui distingue l'I.A. « abstraite » des essais de modélisation de la cognition humaine naturelle consiste dans le fait que s'y trouvent explorés *tous les modes possibles* d'« intelligence », qu'ils soient ou non instanciés par l'homme. Le support du comportement intelligent n'est plus l'homme concret, comme dans les recherches en « simulation cognitive » axées sur la constitution de systèmes experts, mais le medium objectif qui est constitutif des capacités opératoires de tel organisme ou de telle machine : le système symbolique physique. Que faut-il entendre par là ?

On sait qu'un système symbolique est un ensemble de signes ayant la propriété d'être clos. Mais ce n'est pas à ce concept général que songent Newell et Simon ; c'est au concept plus précis de système symbolique *formel*, seul capable de garantir l'universalité opératoire requise. Ce concept suppose en effet que trois contraintes soient satisfaites. Tout d'abord, un tel système doit s'appuyer sur des règles qui déterminent le sens opératoire des symboles, permettant ainsi de différencier dans le signe ce qui est accessoire et ce qui est essentiel. En second lieu, l'ensemble des éléments constituants du système doivent pouvoir être ramenés par décomposition à une liste finie de constituants élémentaires. Enfin, les règles gouvernant la construction des expressions du système doivent être de type concaténatoire, c'est-à-dire qu'elles doivent expliciter les associations légitimes de symboles permettant d'obtenir de nouvelles expressions bien formées. La conjonction de ces propriétés habilite les systèmes formels à permettre le *calcul.*

N'oublions pas toutefois qu'il ne s'agit pas en I.A. d'instruments de calcul, mais d'agent capable de calculer. L'adjectif « physique » est destiné à renvoyer à une réalisation physique quelconque des systèmes symboliques. L'hypothèse sous-jacente à cette précision capitale est que seul un agent capable d'intervenir causalement dans le monde physique – capable, par exemple, de déplacer des symboles dans l'espace, et de produire au fil du temps de nouvelles combinaisons de symboles, peut effectuer le calcul dont le système formel fournit seulement les conditions abstraites de possibilité.

En substituant à la notion de « sujet » celle de « système symbolique physique », Newell et Simon déterminent un niveau d'objectivation propre à l'intelligence artificielle tel qu'aucun privilège particulier ne s'attache désormais à l'activité symbolique spécifiquement *humaine.* Cette manière de généraliser la question des conditions de possibilité qui consiste à rechercher, en deçà des circonstances concrètes d'acquisition des connaissances par l'homme, les structures formelles nécessaires au déroulement du processus cognitif comme tel est bien connue des philosophes. C'est l'enquête transcendantale de Kant qui en a donné le premier exemple, ouvrant ainsi un champ d'enquête inaccessible au psychologue comme tel, limité comme il l'est au champ positif de sa discipline (qu'on le désigne comme la nature humaine, l'ensemble des comportements humains, etc. : l'important est que, dans tous les cas, son objet lui soit empiriquement donné). Ce qui précisément disqualifie aux yeux de Kant un type d'enquête purement psychologique sur la cognition mérite d'autant plus d'être souligné que les meilleurs auteurs, Dennett par exemple, ont tendance à minimiser la différence entre les deux disciplines et donc entre les deux types d'enquête sur ce qui rend la connaissance possible, l'une étant simplement considérée comme le pôle abstrait ou le cas limite de la seconde.

Philosophie, psychologie et intelligence artificielle.

Si nous nous situons du point de vue de la philosophie transcendantale, comme paraît le faire Dan Dennett dans le texte cité plus haut, la différence d'approche des phénomènes de la cognition entre psychologues et philosophes peut être caractérisée en disant que les premiers s'intéressent à la manière dont nous sommes effectivement parvenus à nos représentations tandis que les seconds cherchent à *justifier* l'usage que nous faisons des concepts. Ainsi, Locke et Hume ont du point de vue de Kant mené une enquête de type psychologique, dans la mesure où ils ont tenté de produire une « dérivation physiologique des concepts », c'est-à-dire de montrer comment les concepts de substance, de causalité, etc., ont pu trouver une genèse dans l'expérience humaine. Mais un autre point de vue sur la connaissance

est possible, et même requis pour s'assurer de la validité objective des sciences : c'est l'enquête sur la « déduction transcendantale » des concepts. Celle-ci s'interroge sur la légitimité d'une connaissance de la nature, c'est-à-dire qu'elle cherche à expliquer la manière dont les concepts *a priori* – ceux qui précisément ne sont pas tirés de l'expérience, c'est-à-dire ne sont pas représentables en termes empiriques – se rapportent à leurs objets.

Il est donc impossible, sans déformer ou, à tout le moins, réformer le projet kantien, de soutenir comme le suggère Dennett que la philosophie de type transcendantal soit un cas limite de la psychologie : ce serait faire de l'enquête en légitimation une interrogation sur les faits, réduire le *quid juris* à un *quid facti* suffisamment pointilleux ; ignorer, de ce fait, que, pour le philosophe, ce n'est pas en observant la manière dont nous parvenons à telle ou telle connaissance que nous saurons ce qui la rend objectivement possible. D'une part, parce que cette enquête risque de s'égarer dans les causes occasionnelles de la cognition ; d'autre part, parce que la psychologie est du point de vue propre à la réflexion transcendantale une science incomplète ; tout le domaine des concepts purs *a priori*, qui ne peuvent pas être à proprement parler observés dans une genèse empirique, lui échappe en effet.

On pourrait objecter ici que l'intelligence artificielle n'est nullement concernée par la querelle de frontière entre philosophie et psychologie. Après tout, ne peut-on modéliser l'intelligence humaine en laissant de côté le débat sur les niveaux d'analyse ? Ne peut-on, tout simplement, construire des programmes simulant l'activité intelligente d'apprentissage, de raisonnement, de recherche heuristique, sans avoir à se demander si l'on modélise les conditions contingentes de l'acquisition du savoir ou les conditions les plus générales de l'activité cognitive ? Il vaut mieux remettre à plus tard l'évaluation des résultats, en se contentant pour le moment d'accumuler des données expérimentales susceptibles, par leur variété, d'éclairer, mieux qu'un raisonnement *a priori*, la nature des conditions les plus générales de la cognition.

Quelque séduisante qu'elle puisse apparaître, et quelque conforme qu'elle soit à l'orientation pragmatique de la discipline comme technologie, cette stratégie de réponse n'en méconnaît pas moins un aspect essentiel de l'intelligence artificielle, « abstraite » ou non. La mise en œuvre d'une modélisation informatique suppose en effet que l'on accepte la pertinence de l'opposition entre des éléments « psychologiques » (au sens de contenus vécus, qualitatifs, dotés d'une signification, etc.) et des composantes « formelles » de la cognition. Newell par exemple revendique explicitement à ce propos l'héritage frégéen : « L'histoire du concept de système symbolique physique commence avec la formalisation de la logique, ou l'accent était précisément mis sur la séparation entre aspects formels et psychologiques ». Or, comme on le sait, la ligne de démarcation frégéenne entre « le contenu de jugement » et « la légitimité de l'acte de juger », présentée en tête des *Fondements de l'arithmétique*, est destinée à radicaliser l'opposition kantienne entre la genèse factuelle et la genèse transcendantale de la connaissance. Ce n'est donc pas la manière dont un sujet parvient à construire ses propres représentations qui intéresse l'I.A., mais ce sont les contraintes formelles qui doivent être satisfaites par un système symbolique physique pour qu'il puisse accomplir des tâches exigeant de l'intelligence. De même donc que Kant déplaçait l'axe de la réflexion de la nature humaine vers les conditions de possibilité d'une science de la nature, d'une mathématique, d'une métaphysique, etc., Newell et Simon proposent de ne pas tant rester fixé sur l'objectif de modéliser tel ou tel comportement humain intelligent que de s'attacher à découvrir les conditions générales, intersystématiques, de l'activité intelligente.

Remarquons à ce propos qu'il y a une relation d'interdépendance entre l'hypothèse selon laquelle il existe une multiplicité de modes d'intelligence et l'idée que le support de l'intelligence soit un système symbolique physique, de même qu'il y a interdépendance entre la thèse kantienne selon laquelle c'est le sujet transcendantal qui est le support des synthèses cognitives et l'idée qu'il puisse y avoir d'autres êtres que l'homme qui aient ses capacités cognitives ou des capacités analogues. Dans cette hypothèse, l'intelligence spontanée des hommes a pour condition de possibilité certaines propriétés logiques ou formelles au sens large (pour Kant : à la fois formelles et transcendantales) matériellement représentables. C'est en explorant ces propriétés que l'on peut comprendre la capacité humaine de produire des comportements adaptés aux situations les plus variées et les plus imprévues et non pas en en restant, comme le feraient des psychologues, au niveau phénoménologique des séquences de comportements effectifs.

Si la perspective de Newell et Simon est bien celle de la genèse fondationnelle, il est vain de mobiliser contre elle des arguments qui appartiennent au registre de la simulation de la cognition humaine. C'est précisément ce que H. et S. Dreyfus tentent de faire dans leur récent ouvrage, *Mind over Machine.* Analysant le processus de l'acquisition de savoir-faire tels que l'apprentissage de la conduite automobile ou celui des échecs, ils discernent cinq phases successives au cours desquelles la transmission et l'application de règles s'effacent progressivement pour faire place à une appréhension globale et intuitive de situations de complexité croissante. De cette régression du rôle des règles, ils pensent pouvoir conclure qu'une discipline telle que l'I.A. classique, fondée sur le caractère réglé du processus cognitif, ne peut, au mieux, que donner une bonne modélisation des phases primitives de la cognition, mais doit nécessairement échouer dès que l'on passe aux niveaux supérieurs de la connaissance.

Si la lecture proposée ici est juste, ce qui rend inefficace l'objection de Dreyfus & Dreyfus tient à ce qu'ils confondent le point de vue psychologique de la genèse concrète des compétences humaines avec le point de vue fondationnel abstrait de ce qui rend possibles la connaissance et l'intelligence en général. Ils mettent en doute le caractère « réaliste » ou authentiquement simulatoire du modèle symbolique de Newell et Simon et lui reprochent de ne pouvoir fournir de modèle adéquat de la compréhension : « Puisque comprendre ne consiste pas en faits et en règles, l'espoir de voir un jour le professeur remplacé par un ordinateur est fondamentalement illusoire », écrivent-ils par exemple (p. 133). Mais le problème de l'I.A. classique, dont Newell et Simon sont les représentants patentés, n'est pas, comme on l'a vu, de respecter les caractéristiques particulières de l'apprentissage humain, mais de fournir une modélisation adéquate des conditions *générales* de possibilité de la cognition, humaine ou non, modélisation qui permette de construire un jour un système capable d'effectuer (d'une manière ou d'une autre) les opérations qui, effectuées par un homme, sont caractéristiques des comportements intelligents.

Le problème de l'universalité.

Pour que l'I.A. puisse atteindre les conditions générales de l'intelligence tout en travaillant sur un système particulier, il lui faut apporter la preuve que ce système est en un sens un « bon représentant » de tous les systèmes possibles, c'est-à-dire qu'il n'est pas essentiellement limité par les caractéristiques de son interface ou par l'étendue de ses capacités de calcul. L'universalité représente de ce fait la contrainte la plus générale qui s'attache à la modélisation de l'intelligence par un système donné. C'est une contrainte analogue que Kant avait placée en tête des conditions d'adéquation des procédures de synthèse. Pour que les conditions *a priori* d'une expérience possible soient aussi les conditions de pos-

sibilité d'objets de cette expérience, il faut d'après Kant que les fonctions de la synthèse soient universelles. Si en effet la synthèse de certaines intuitions par des concepts empiriques n'était qu'un fait contingent, un produit accidentel de l'acte de cognition d'un sujet empirique particulier, on ne pourrait s'expliquer qu'il existe une expérience objective, c'est-à-dire intersubjectivement accessible, dotée de stabilité, et capable d'être répétée. Le problème de l'universalité est ainsi posé comme celui du lien entre l'expérience des objets d'une part et ses conditions *a priori* : les fonctions de la synthèse, aussi bien intuitives que conceptuelles. La solution kantienne du problème passe par l'appel à un principe transcendantal de l'unité, sur lequel s'appuient toutes les synthèses empiriques, c'est-à-dire tous les jugements qui peuvent s'opérer sur la base de données d'expérience. Le concept qui joue un rôle stratégique dans la solution kantienne est celui de *schème* : représentation homogène à la fois à la catégorie, c'est-à-dire à la fonction synthétique *a priori*, et au sensible des phénomènes, le schème garantit la possibilité de l'unité du divers de l'intuition et la reproductibilité de principe des synthèses empiriques.

La notion de schème paraît évidemment annoncer les schémas de l'I.A. Mais il est trompeur de tenter de transposer directement le problème kantien, avec sa solution, dans le domaine de l'I.A., ne serait-ce que parce que la notion de schème est chez Kant étroitement solidaire de sa conviction en vertu de laquelle le temps – détermination transcendantale à la fois pure et sensible – est le seul medium approprié dans lequel les concepts puissent acquérir leur signification, c'est-à-dire leur rapport à des objets.

Il n'en reste pas moins que la manière dont Newell pose, dans ses articles récents, la question de l'universalité de l'I.A., rappelle le problème auquel Kant était confronté. De même en effet que Kant devait résoudre la question de ce qui fonde l'unité de l'expérience, et pensait pouvoir y répondre par une synthèse transcendantale réglée par des schèmes, Newell doit démontrer qu'un système particulier est capable, en dépit des caractéristiques idiosyncrasiques du langage employé, de l'architecture choisie, etc., de produire une fonction entrée-sortie *quelconque*, conformément à la flexibilité requise d'un comportement intelligent. Le critère d'universalité ainsi défini vise à légitimer le caractère authentiquement intelligent du calcul ou de la recherche heuristique effectués par la machine, par opposition à la capacité que l'on pourrait trivialement conférer à un système de répondre à un ensemble limité d'entrées sur le schéma stimulus-réponse.

Pour être menée à son terme, la démonstration de l'universalité comprise en ce sens passe par le développement technologique de l'I.A., en particulier en ce qui concerne la question des limites relatives à l'interface et à la taille des mémoires. Cependant, ces limitations sont des problèmes secondaires dans l'élaboration d'un concept d'intelligence artificielle, parce que ce sont des limitations de fait, dont on sait qu'elles seront probablement bouleversées par les innovations technologiques à venir ; rien ne paraît faire obstacle à l'introduction de capteurs sensoriels diversifiés, ni à une gestion optimale de mémoires sinon illimitées, du moins « ouvertes ». De telles limitations physiques à l'universalité, notons-le en passant, ne sont pas l'apanage exclusif des machines ; elles ne font qu'illustrer la « finitude », reconnue par les philosophes de tous bords, qui caractérise le processus cognitif comme tel (Dieu exclu, mais c'est une question non résolue de la théologie ou de la grammaire philosophique de savoir si l'omniscience divine est une forme de la connaissance ou son point de dissolution).

Si ces limitations physiques paraissent secondaires du point de vue de Newell, c'est qu'elles s'effacent devant des limitations de principe, telles que pourraient l'être des théorèmes formels de limitation. Or l'hypothèse de l'universalité s'appuie solidement sur une thèse formelle assurant la possibilité de principe, pour tout système réalisant une machine de Turing, de simuler toute autre machine ayant la

même propriété. Une machine de Turing est précisément dite « universelle » parce que toute procédure calculable l'est par une machine de Turing. Il n'y a donc pas de calcul effectuable qui puisse, de manière principielle, être ineffectuable par un système « universel ». Il ne s'agit pas là à proprement parler d'un théorème de théorie du calcul ; mais d'une thèse établie séparément par Church et Turing, soit un « fait », mais un fait appelé à jouer un rôle fondationnel précisément dans la mesure où il garantit le caractère essentiel, et non contingent, des propriétés des machines appartenant à la classe des systèmes symboliques universels : « Les machines universelles, remarque Newell, fournissent un point de vue particulier relativement à ce qui est essentiel. Chaque machine universelle manifeste sous une forme ou sous une autre toutes les propriétés de toute machine universelle. » Ce principe d'équivalence formelle entre tous les systèmes de la classe garantit, par exemple, que rien d'extrinsèque ou d'*ad hoc* ne soit introduit dans la modélisation de l'intelligence par le choix de tel langage informatique ou de tel autre. En dépit de la lourdeur de l'utilisation d'un langage comme FORTRAN pour accomplir les tâches accomplies par LISP, la possibilité de simuler le second par le premier est *a priori* garantie. Cette garantie suffit pour ce que l'on attend du langage de calcul, c'est-à-dire qu'il fournisse un instrument fiable pour obtenir, à une isomorphie près, une modélisation des comportements intelligents.

De l'expérience possible à l'intentionnalité.

La condition d'universalité telle qu'elle se trouve définie et satisfaite en I.A., soit en termes de calculabilité universelle, peut cependant paraître manquer la dimension à laquelle Kant liait l'exigence d'universalité, à savoir le rapport à l'expérience possible. Même sans entrer dans les objections relatives à l'indécidabilité qui vient limiter *a priori* la capacité des systèmes formels à démontrer la vérité ou la fausseté de certaines de leurs formules – limitations qui ne peuvent servir d'argument contre l'I.A. dans la mesure où des limitations analogues, sinon plus sévères, affectent le système cognitif naturel[2] –, on peut objecter que Newell ne ménage à ses systèmes symboliques physiques aucun rapport à l'expérience : la pensée symbolique est bien « aveugle » non seulement au sens où, comme dans la caractéristique universelle de Leibniz, il n'est pas nécessaire de s'attacher au sens individuel des symboles pour parvenir à dériver telle ou telle propriété des expressions, mais, plus gravement, au sens où le système n'a pas de véritable ancrage dans le réel. Hilary Putnam, John Searle et Hubert Dreyfus ont avec des arguments différents tenté de démontrer que la dimension proprement sémantique manquait *principiellement* aux langages de programmation : la machine ne connaît pas les états de chose auxquels sont censées renvoyer les formules, elle ne connaît pas les objets que ses symboles sont censés dénoter. Ce que la machine « connaît », à la rigueur, ce sont des séquences de symboles et des opérations à effectuer sur elles, sans qu'aucune référence extra-systématique n'intervienne jamais dans le processus de calcul. Mais est-il encore légitime de parler de « connaissance », quand précisément fait défaut le présupposé de l'activité cognitive, à savoir l'existence de buts sociobiologiques dépendant d'un certain contexte adaptatif ? Il n'y a sens et dénotation que pour l'observateur qui effectue les corrélations entre les symboles et les opérateurs d'une pan et un certain modèle visé de l'autre. C'est l'observateur qui « injecte »

2. Sur la question de la portée des théorèmes de Gödel pour l'intelligence artificielle, cf. Jacques Bouveresse, *La Parole malheureuse*, Paris, Éd. de Minuit, 1971, pp. 406-407.

de l'intentionnalité dans le système, lequel n'est après tout qu'un mécanisme sophistiqué de combinaison d'éléments physiques.

En complète rupture avec cette version réductionniste des machines, Newell pose au contraire le caractère intrinsèquement intentionnel du système symbolique physique. « Le concept le plus fondamental pour un système symbolique, écrit Newell, est ce qui donne aux symboles leur caractère symbolique, c'est-à-dire qui leur fait représenter une certaine entité » (« Physical symbol Systems », p. 156). On dira d'une entité X (un symbole, une expression, un opérateur) qu'elle « désigne » (ou fait référence à) une entité Y relativement au processus P si, « quand P prend X comme entrée, le comportement de X dépend de Y ». On peut évidemment objecter à une telle définition qu'elle ne parvient pas à mettre en évidence de manière univoque la dépendance sémantique, c'est-à-dire l'interprétation souhaitée de X : rien ne vient en effet limiter les entités susceptibles d'influencer pratiquement le comportement de X : un court-circuit peut par exemple être désigné par X relativement au processus P dans la mesure où il influe sur le comportement de X quand il effectue P. Pour supprimer cette indétermination, il faudrait faire valoir une relation proprement sémantique (par exemple, en termes des conventions qui régissent le langage utilisé) venant restreindre la classe des modèles acceptables.

Ce que Newell invoque à l'appui de son interprétation « référentielle », c'est que certains des opérateurs d'un système quelconque accomplissent effectivement une fonction proprement désignative ; on ne peut dire qu'un système ignore tout référent, puisqu'il doit pouvoir faire référence pour pouvoir fonctionner, ne serait-ce qu'à certains des sous-programmes. Tel est par exemple le rôle de l'opérateur ASSIGN du système présenté par Newell à des fins d'illustration. C'est un opérateur qui établit une relation entre deux symboles du langage, ou plus exactement entre un symbole SI et le référent d'un autre symbole, S2. Cet opérateur permet ainsi d'établir la coréférence entre deux symboles, de définir, etc. Une fois l'opérateur ASSIGN activé, on obtient une relation d'ACCES à l'expression, qui rend possible la manipulation des symboles constituant cette expression. L'accès est la capacité de retrouver une information en mémoire, soit selon un mode de recherche séquentielle si l'information est codée sur bande magnétique, soit selon un mode aléatoire pour les mémoires à tores. En laissant de côté cet aspect dynamique de la recherche, l'accès se réduit du point de vue physique à un « mécanisme à deux positions qui ouvre un chemin entre le processus et la chose accédée » (« Physical symbol Systems », p. 158).

On peut être tenté de tirer du mécanisme même de l'adressage et de la recherche en mémoire des arguments contre la portée réellement sémantique de l'accès. Sans doute les relations qui sont spécifiées entre des expressions et entre des formules par ASSIGN peuvent-elles être *destinées à* modéliser un aspect du monde. C'est un fait que la machine parvient à s'orienter, dans l'accomplissement d'un sous-programme quelconque, vers une certaine séquence de signes, un certain opérateur. Il a bien fallu pour cela qu'elle soit capable de les désigner comme objectifs d'une recherche. Mais le fait que l'on puisse activer un opérateur, c'est-à-dire y avoir accès, suffit-il à prouver que le système a fait référence à cet opérateur ? En outre, cette référence « interne » au système fournit-elle un indice adéquat de la capacité générale de faire référence ?

En ce point du débat, partisans et détracteurs de l'intentionnalité machinique doivent bien convenir que leur manière de concevoir « le rapport à l'objet » est à son tour régie par leurs présupposés sur la nature de la cognition. La stratégie intentionnaliste consiste à poser une dualité aporétique du sujet face à un monde d'objets, afin de mieux dissoudre l'aporie en faisant intervenir, véritable *deus ex machina*, une fonction synthétique originaire, l'intentionnalité, qui est constitutive à la fois des significations et

d'un monde d'objets. Remarquons toutefois le contraste entre la richesse descriptive ou phénoménologique du terme d'« intentionnalité » et sa pauvreté explicative. Cette machine de guerre antibehavioriste s'avère, à l'examen, être vide ou circulaire du point de vue d'une théorie de l'intelligence : les théories intentionnalistes font de l'intentionnalité un terme primitif. Elles ne sont donc pas en position d'expliquer à son aide la rationalité ou l'intelligence.

Une façon plus prometteuse d'aborder le problème consiste, selon Dennett[3], à faire de l'intentionnalité une notion purement pragmatique (et non plus théorique ou philosophique). Dennett appelle « système intentionnel » un système dont le comportement peut être prédit en lui attribuant un ensemble de désirs, de croyances, ou, en termes plus neutres, en supposant qu'il a des objectifs précis et qu'il est capable d'utiliser de l'information pour les réaliser. La perspective de Dennett rejoint celle de Newell en ceci qu'il propose une définition susceptible de s'appliquer de manière non discriminative à l'homme et à la machine, le vocabulaire utilisé étant vidé de toute connotation « métaphysique, morale ou divine ». Le point de vue « intentionnel », suggère Dennett, nous est imposé par la nécessité de prédire les comportements d'un ordinateur capable de jouer aux échecs. Les modes d'analyse considérés comme indiscutablement objectifs des performances de l'ordinateur s'arrêtent généralement au niveau de l'architecture ou de l'état physique du système. Or, pour prédire quelle stratégie le joueur d'échecs électronique (par exemple) va utiliser, le point de vue de l'architecture fonctionnelle n'est pas suffisant ; la stratégie est un montage trop complexe pour être analysable en termes de fonctions composantes. Tenter d'aborder le système du point de vue de ses états physiques, comme le font invariablement les détracteurs de l'intentionnalité machinique – souvent dans un esprit néo-vitaliste – consiste à traiter le système comme un ensemble de sous-systèmes physiques – transistors, résistances, circuits, etc. Quoiqu'il soit en principe possible de prédire le comportement sur cette base purement physique, il est clair que la moindre prédiction demanderait un travail colossal de calcul ; l'analyse physique du système n'intervient typiquement que lorsque l'analyse architecturale révèle un dysfonctionnement relevant du niveau inférieur : c'est à ce niveau qu'opèrent les prédictions du réparateur.

C'est donc un fait pragmatique que la prédiction mettant en œuvre des critères de rationalité s'impose à l'analyste quand il s'agit des séquences de comportements complexes qu'un ordinateur adéquatement programmé peut accomplir. Reste alors à « payer les traites », selon l'expression de Dennett, c'est-à-dire à démontrer que les notions intentionnelles sur lesquelles s'appuie la prédiction sont réalisées fonctionnellement. Si l'ordinateur se comporte effectivement comme s'il avait des croyances, si l'ingénieur parvient à faire accomplir par l'ordinateur ses « objectifs », si rien dans sa structure ne s'oppose à cette façon de décrire ses comportements dont on a vu l'intérêt, il faut considérer que l'idiome intentionnel est justifié, sinon à titre de théorie de l'intelligence, du moins à titre de théorie prédictive du comportement. La manière dont Newell met en avant les capacités désignatives d'un système relève d'un affaiblissement similaire des questions intentionnelles. Affaiblissement justifié par le point de vue universaliste adopté.

En effet, si l'on admet avec Newell le caractère universel des systèmes symboliques physiques, on doit admettre que les relations intentionnelles, telles que la référence, la croyance, etc., ne relient pas des termes *hétérogènes* dont il s'agirait de comprendre la liaison. En vertu des caractéristiques du système

3. Daniel Dennett, « Intentional Systems », in *Mind design. Philosophy, Psychology, Artificial Intelligence*, éd. par John Haugeland, Cambridge, The M.I.T. Press, 1981, pp. 220-242.

symbolique physique, toutes les données qui font l'objet d'une référence ou d'une attribution de valeur de vérité (par exemple) doivent *déjà* être accessibles au système. Or, tout système cognitif, *ex hypothesi* est un système symbolique physique. L'idée de donnée radicalement hétérogène au système « connaissant » n'a donc simplement pas de signification. L'hypothèse universaliste étant admise, il est normal de traiter la référence à partir de la structure du système, et de chercher à comprendre l'intentionnalité comme un rapport des commandes et des expressions. Le phénomène crucial est ici celui de la *simulation.* Faire référence, pour un système symbolique physique, c'est nécessairement *simuler* symboliquement la structure et les propriétés d'un objet. La principale présupposition de la simulation est qu'elle préserve la structure de la réalité. Cependant, la notion de « réalité » doit à son tour être comprise dans les termes du système.

Cette circularité entre conception de la référence et théorie du système symbolique est-elle litigieuse ? Elle le serait s'il existait une manière non circulaire de traiter des notions fondamentales de la sémantique. Dans ce cas comme dans d'autres on constate que partisans et adversaires de l'attribution à l'ordinateur de capacités sémantiques ou intentionnelles sont toujours conduits à en appeler à leurs propres présupposés. Quel que soit l'agacement de l'intentionnaliste « irréductibiliste », il n'est pas moins circulaire de considérer (comme le font par exemple John Searle et Hubert Dreyfus) que les ordinateurs sont incapables d'intentionnalité que de leur en conférer une (comme le font Dennett et Newell) en dépouillant ainsi l'intentionnalité de son attrait humaniste.

La réalité des constructions : l'I.A. est-elle idéaliste ?

On peut être tenté de qualifier d'« idéaliste » la position de Newell : rien dans le monde n'« existe » s'il n'est le corrélat d'un état possible du système. Il convient pourtant de modérer cette appréciation. Tout d'abord, Newell maintient, au moins implicitement, que la combinaison des symboles correspond bien à des opérations, des objets, des états de choses concrets ou abstraits dans le monde extérieur. En d'autres termes, il revendique une portée « réelle » pour les assemblages de caractères effectués par la machine. Dira-t-on qu'il s'agit d'un vœu pieux, d'une sorte de foi caractéristique comparable à celle de Leibniz ? L'expression malheureuse d'« action à distance » qu'il emploie pour désigner cette correspondance recherchée entre le symbolisme formel et un modèle – en l'occurrence, le monde ou le micro-monde représenté dans la structure de données – peut laisser penser qu'il s'agit bien d'une assertion typiquement dogmatique, infondée et infondable. Pourquoi « action à distance » ? Voici la réponse : « Le processus se comporte comme si des entrées, éloignées de celles qu'il reçoit en fait, le provoquaient. C'est la propriété symbolique, que le fait d'avoir [le symbole] X revient à avoir [la chose désignée] Y relativement aux objectifs du processus P. »

Le fait que des objets distaux, c'est-à-dire « éloignés » de la périphérie du système, soient représentés par les expressions symboliques, ne semble à première vue guère défendable sur la seule base de l'hypothèse du système symbolique physique, même universel. La seule chose que nous avons pu dériver plus haut de l'hypothèse universaliste, c'est que tout objet du monde est pour le système un objet auquel il a accès, c'est-à-dire qu'il peut traiter comme entrée. Par conséquent, seul l'objet proximal, c'est-à-dire l'objet qui est saisi par les modules sensoriels, semble pouvoir être caractérisé par le système, dans la mesure où le système est structurellement incapable d'extraire de l'information d'un objet qui ne serait pas compatible avec ses capteurs ou avec la structure de sa mémoire. En se plaçant de ce point de vue,

on pourrait dire que l'hypothèse des S.S.P. (systèmes symboliques physiques) conduit à développer une philosophie analogue à l'idéalisme transcendantal de Kant. La position de Kant consiste à affirmer contre le réalisme que les objets sont transformés par le processus cognitif, en sorte que ce sont des phénomènes qui nous sont accessibles, et non les choses-en-soi dont ils sont l'apparaître. Idéalisme transcendantal, toutefois, dans la mesure où non seulement l'existence de choses en soi est maintenue à titre de fondement des phénomènes, mais où les phénomènes sont pourvus de ce que Kant appelle une « réalité empirique » : soit une validité objective qui rend possible l'énoncé de lois nécessaires de la nature.

En invoquant l'existence d'une structure « éloignée » à laquelle correspond la structure symbolique mémorisée, Newell fait une hypothèse comparable au réalisme minimal de Kant. L'objet extérieur, distal, existe, pourrait dire Newell. Mais ce que nous pouvons en connaître est contraint par notre type de réception sensorielle et notre mode de catégorisation des données. L'une des conditions de cette catégorisation consiste dans la loi de transitivité de la désignation, assez voisine de ce que Kant appelait la synthèse de la recognition : « Une loi transitive importante, écrit Newell, consiste dans le fait que si X désigne Y et Y désigne Z, alors X désigne Z. Pour notre cas, il y a d'abord l'acquisition qui, par l'accès à la structure externe réelle *(actual)*, crée une structure dans la mémoire du système qui dépend de cette entité externe ; puis la préservation de cette structure de mémoire dans le temps fournit une structure de mémoire à un temps ultérieur qui dépend encore de l'objet ; en fin de compte, l'accès associé au symbole interne rend la structure disponible pour un processus qui se déroule alors de la même manière, l'attribuant à une nouvelle entité et instanciant la relation de désignation. »

Le moment de l'« acquisition » suppose bien que le contact avec une structure « réelle », caractéristique d'une distribution d'objets dans le monde extérieur, ait été effectué puis qu'une trace de cette structure ait été mémorisée. Ce sont en revanche les caractéristiques formelles du système, (telles que l'articulation d'ASSIGN avec le mécanisme général de l'accès) qui rendent possible la réidentification de l'objet sous des désignations différentes.

Quelle philosophie pour l'I.A. ?

J'ai tenté dans ce qui précède de tirer toutes les conséquences de la suggestion faite indépendamment par Dennett et par Dreyfus du caractère kantien de l'I.A. classique, en proposant une interprétation transcendantale de l'hypothèse de Newell selon laquelle (dans ses propres termes) « une certaine classe de systèmes réalise la nature essentielle des symboles et constitue la condition nécessaire et suffisante d'un agent généralement intelligent » (« The knowledge level », p. 94).

L'analogie entre la démarche de Newell et celle de Kant a pu être développée naturellement et sans accrocs : nous avons repéré ici et là une même façon d'aborder de très haut les conditions de la cognition, le sujet transcendantal se trouvant relayé par le système symbolique physique. Les autres pièces du puzzle se sont emboîtées peu à peu : une distinction analogue entre conditions nécessaires et occasionnelles de la cognition, une recherche menée en termes similaires (quoique non identiques) de ce qui fonde l'universalité de l'agent de la connaissance, la commune quête de conditions valables *a priori*, et enfin la question du « rapport à l'objet » ont apporté une crédibilité nouvelle à ce qui n'était d'abord qu'une sorte de bon mot.

Il ne suffit pourtant pas d'esquisser un parallèle pour être assuré de sa validité. J'ai parlé plus haut de la difficulté particulière que présente la mise en parallèle de textes de l'I.A. théoriques et de philo-

sophie. L'une des difficultés qui se présentent maintenant consiste dans la manière dont Newell et Simon jugent leur propre hypothèse. On observe en effet un décalage entre, d'un côté, la manière dont l'hypothèse est introduite par réflexion sur le concept de système symbolique physique instanciable par une machine donné, et, de l'autre, la façon dont Newell et Simon perçoivent leur activité d'ingénieurs : construisant des machines, les testant, bref se livrant à des expérimentations semblables à celles qui sont effectuées dans les sciences de la nature. « Construire réellement une machine revient à poser une question à la nature ; et nous écoutons la réponse en voyant la machine opérer et en l'analysant par tous les moyens disponibles d'analyse et de mesure » (« Computer science as empirical inquiry », p. 36). Que la construction d'un artefact permette une mise à l'épreuve expérimentale de la théorie, personne ne songera à le nier. Mais cela n'implique pas que l'hypothèse fondationnelle constitutive du champ de recherche, donnant au S.S.P. le statut des conditions nécessaires et suffisantes de tout comportement intelligent, soit à son tour « issue de l'expérience ». Si l'on est prêt à suivre Newell et Simon lorsqu'ils disent que leur hypothèse « a des racines empiriques profondes », dépendante comme elle l'est du développement du traitement de listes, de l'invention du langage LISP, etc., on est plus sceptique sur le caractère authentiquement empirique (par exemple) de leur conception du symbole : « Expliquer la nature des symboles est une proposition scientifique sur la Nature. Cette explication est dérivée empiriquement, selon un développement long et progressif» (*ibid*., p. 37). Doit-on prendre ces déclarations métathéoriques au sérieux ? Leur motivation apparaît clairement : il s'agit de convaincre non seulement les informaticiens, mais les pouvoirs publics et les industriels que l'I.A. a sa place parmi les sciences respectables, comme la biologie et la physique. C'est une réponse dirigée contre ceux qui laissent entendre que l'ingénierie cognitive n'est qu'une discipline bâtarde, à mi-chemin entre la technologie de la programmation et l'illusionnisme.

Ce que je conclurai en ce point, sans espérer avoir l'adhésion des informaticiens soucieux de faire la preuve de la scientificité de leur discipline, c'est qu'il existe au moins une manière philosophique de lire leurs textes. En termes à peine plus hardis, on peut faire l'hypothèse que les chercheurs en I.A. mettent en œuvre une philosophie implicite – dans la conception de la finalité et du fondement de leur recherche, dans l'élaboration de leur théorie de la cognition, de la représentation des connaissances, dans leur conception de l'apprentissage –, qui ne coïncide pas nécessairement avec la manière dont ils se représentent leur propre travail. L'évaluation philosophique de son propre travail par un auteur-concepteur fait intervenir des enjeux de nature polémique ou, plus généralement, sociostratégiques, qui conduisent à privilégier certains mots clefs, certains registres ou certaines alliances. Pour la discipline nouvelle qu'est l'I.A., il est vital de se voir reconnue comme la dernière-née des sciences, et de tracer un arbre généalogique où aucune affiliation avec la philosophie ne vient ternir la pureté de la descendance.

La caractérisation philosophique que j'ai proposée de la manière dont Newell et Simon mettent en place leur hypothèse concernant les systèmes symboliques physiques ne peut évidemment pas être généralisée à l'ensemble de l'I.A. Choisissant le point de vue le plus élevé et le plus éloigné des applications concrètes, ils ont exploré la question des conditions générales de l'intelligence en utilisant une rhétorique et des outils conceptuels qui paraissent bien voisins de la philosophie kantienne. Cette analogie, disons-le rapidement, peut facilement s'expliquer, ne serait-ce que par l'influence directe exercée sur eux par deux courants dérivés du kantisme et apparentés à lui malgré un déplacement des types de réponses, le logicisme de Frege et Carnap et le formalisme de Hilbert. Cela n'implique nullement que d'autres membres de ce qu'on appelle l'« intelligents I.A. » partagent les mêmes présupposés ou posent

les mêmes questions (puisque la philosophie commence avec la sélection des questions dignes d'être posées). Il est au contraire frappant de constater la variété des problématiques philosophiques mises en œuvre dans les divers domaines, qu'elles se trouvent mobilisées à propos de la représentation de l'abstraction, du rapport entre le système et le donné, du choix d'une mémoire épisodique ou sémantique, de l'apprentissage des machines, etc.

Que ceux des informaticiens qui – lassés de voir le philosophe drapé dans son rôle de Cassandre – envisagent avec pessimisme toute collaboration avec les philosophes ne s'inquiètent pas outre mesure : les philosophes ne sont pas là pour redresser les torts ou annoncer d'en haut les impossibilités de principe ou les risques du progrès technique en ingénierie cognitive. Ils peuvent seulement proposer, après coup, des hypothèses de lecture destinées à faire apercevoir tout l'arrière-plan conceptuel du travail d'innovation et à souligner toutes les décisions théoriques impliquées par cet objet technique très particulier qu'est un programme d'I.A. Quant à l'évaluation de l'I.A., elle revient non pas aux seuls philosophes, mais aux intéressés, c'est-à-dire aux agents concernés : les possibilités techniques relèvent de la compétence des ingénieurs, les risques de la compétence des citoyens.

Joëlle Proust.

BIBLIOGRAPHIE

Dennett, Daniel, *Brainstorms. Philosophical Essays on Mind and Psychology*, Montgomery, Bradford Books, 1978.

Dreyfus, Hubert L., *Intelligence artificielle, mythes et limites*, trad. franç. par Rose-Marie Vassallo-Villaneau et Daniel Andler, Paris, Flammarion, 1984.

Dreyfus, Hubert et Stuart, *Mind over Machine*, New York, The Free Press, 1986. Newell, Allen, « Physical symbol Systems », *Cognitive Science*, 1980, 4, pp. 135-183.

Newell, Allen, « The knowledge level », *Artificial Intelligence*, 1982, 18, pp. 87-127.

Newell, Allen et Simon Herbert A., « Computer science as empirical inquiry : Symbols and search », in *Mind Design*, Haugeland J., éd. (voir notre note 4), pp. 35-66.

Putnam, Hilary, *Raison, vérité et histoire*, trad. franç. par A. Gerschenfeld, Paris, Éd. de Minuit, 1984, chap. I.

Searle John R., *Du cerveau au savoir*, trad. franç. par Catherine Chaleyssin, Paris, Hermann, 1985.