Intellectica, 2005/1, **40**, pp. 91-108

L'intentionnalité et le virtuel

Denis BERTHIER*

RESUME. Après avoir rappelé notre définition du virtuel (ce qui, sans être réel, a les qualités du réel, avec force et de manière pleinement actuelle – i.e. pas potentielle), nous expliquons pourquoi les « agents artificiels » qui apparaissent un peu partout dans le monde informatique doivent être conçus comme des agents *virtuels*, pourquoi leur prêter de l'intentionnalité n'est pas une question de choix, contrairement à ce que suggère la « posture intentionnelle » de Dennett, et pourquoi cette intentionnalité doit elle-même être dite virtuelle. Une nouvelle difficulté en résulte pour le projet de naturalisation de l'intentionnalité.

Mots clés : Épistémologie, intentionnalité, agent virtuel, intelligence artificielle.

ABSTRACT. Intentionality and the virtual. After recalling how we define the virtual (which is not real but clearly displays the full qualities of the real, in a plainly actual – i.e. not potential – way), we explain why the "artificial agents" that are appearing in every part of the world of computers should be conceived as *virtual* agents, why attributing intentionality to them is not a matter of choice, contrary to what Dennett's "intentional stance" suggests, and why such intentionality should be called virtual. As a result, the cognitive sciences project of "naturalizing" intentionality faces a new challenge.

Key words: Epistemology, intentionality, virtual agent, artificial intelligence.

I. INTRODUCTION

Le monde est un réseau mouvant de symboles et l'ordinateur le métier sur lequel nous avons cru en produire des simulacres. Mais nous découvrons en réalité, sans encore en mesurer toutes les conséquences, qu'aujourd'hui le retissage permanent du monde ne peut plus guère être isolé de ce qui se tisse en parallèle sur le métier informatique.

De même que, dans l'espace ordinaire, se mêlent indistinctement les reflets aux originaux, les images virtuelles aux images réelles[1], ainsi dans le monde informatiquement augmenté qu'une partie de l'humanité élabore sans répit viennent à s'imbriquer et à interagir de nouvelles catégories d'entités au statut encore passablement incertain, dont l'existence, tout comme celle de nos reflets dans les miroirs, atteint aux sources de notre identité.

Parmi les plus évoluées et les plus intrigantes de ces entités, figurent différents types de programmes informatiques conçus pour nous apparaître naturellement comme des *agents* dotés d'une certaine rationalité et capables de partager avec nous des informations ou des savoirs, c'est-à-dire de communiquer avec nous à leur propos ou de raisonner dessus. Pas plus qu'ils ne se lais-

* Institut National des Télécommunications, Groupe des Écoles des Télécommunications, 9 rue Charles Fourier, 91011 Evry Cedex, France, Denis.Berthier@int-evry.fr, http://www.carva.org/denis.berthier.

[1] Et de même que les premières se sont multipliées avec l'invention des miroirs, lentilles et autres instruments d'optique.

sent réduire à un statut simpliste d'objet ou de système technique[2], ces agents artificiels ne sont totalement compréhensibles à partir du concept d'objet hybride, tel qu'il a été développé par le sociologue Bruno Latour (1991 ; 1997). En fait, avec leur *apparence d'intentionnalité*, ces agents possèdent une propriété qui déborde largement les diverses approches de la sociologie ou de la philosophie des techniques et qui laisse en nous une interrogation lancinante.

De cette apparence d'intentionnalité, nous proposons ici une interprétation basée sur une conception du *virtuel* qui s'appuie sur l'étymologie (de *virtus* – vertu) et sur les usages scientifiques et techniques du terme[3]. Cette conception, ainsi que ce qui en résultera au sujet de l'intentionnalité, peut être comprise à deux niveaux : soit à celui d'une simple analogie optique, le reflet dans un miroir (qui est, techniquement parlant, une image virtuelle) étant considéré comme exemple prototype du virtuel ; soit, selon un cheminement plus long et plus technique, grâce à l'introduction d'un concept dual de celui de virtuel, le concept d'*interopérabilité*, avec lequel nous avons généralisé les divers avatars de ce terme dans le monde des technologies de l'information.

II. POSITION DU PROBLEME : L'INTENTIONNALITE APPARENTE DES AGENTS ARTIFICIELS

Si les agents artificiels nous paraissaient simplement dotés d'une rationalité idéalisée, abstraite, mécanisée, et de capacités communicationnelles de même nature, sans doute atteindrions-nous un degré de compréhension acceptable en adjoignant aux ressources de la logique formelle celles du structuralisme et, en particulier, sa conception des systèmes de signes, avec l'idée que des effets de sens peuvent résulter de la manipulation formelle de symboles a-signifiants : en effet, si une même combinatoire formelle régit nos systèmes de signes et ceux des machines, il est aisément concevable, au moins en principe, que celles-ci puissent participer de-ci de-là, ne serait-ce que comme adjuvants, au tissage du sens.

Mais le fait est que ces agents nous *semblent* non seulement dotés de rationalité et de capacités communicationnelles *formelles*, ils nous *semblent* de surcroît naturellement dotés d'*intentionnalité* – idée qui, cette fois, ne peut pas être assimilée tant elle heurte les dualités établies entre le physique et le mental, entre le corps et l'esprit : comment un programme informatique, implémenté dans un ordinateur et ainsi instancié en un mécanisme, c'est-à-dire *in fine* un objet physique, pourrait-il avoir de l'intentionnalité, caractéristique supposée propre au mental (au moins dans la conception phénoménologique originelle) ?

Le concept d'intentionnalité étant l'un des plus problématiques de la philosophie contemporaine, qu'entendons-nous en disant que des agents artificiels nous en semblent dotés ? Au niveau le plus élémentaire, cela signifie que nous interprétons spontanément leur comportement en termes mentalistes : de désirs, de croyances, de buts, d'intentions, etc.[4] Mais, plus généralement, si on les

[2] Par exemple, par une analyse sommaire, à un niveau conceptuel insuffisant, en termes de mécanisation de la logique formelle.

[3] Cette conception a été développée dans notre livre « Méditations sur le réel et le virtuel » (2004) – MRV par la suite.

[4] Il nous semble capital de souligner dès à présent que nous adoptons en permanence, sans y penser, des formes rudimentaires de cette manière d'interpréter tous les comportements artificiels, par exemple

considère dans toute leur diversité, ces agents nous semblent capables de se référer à tout ce à quoi un être humain peut se référer : tant à des concepts abstraits, des croyances, des buts, des intentions, qu'à des sentiments, des sensations, des objets ou processus physiques du monde réel. Bref, ils semblent dotés d'intentionnalité au plein sens philosophique du terme, c'est-à-dire au sens très particulier qui lui a été donné par la scolastique au Moyen Age[5], puis réactualisé par Brentano et repris par Husserl ; l'intentionnalité signifie la capacité générale de la conscience (de notre conscience à nous, humains) à viser un objet, à *se référer à* « quelque chose », à *porter sur* « quelque chose » (« quelque chose » dont on ne présume en rien du statut d'existence), et à se le représenter :

> « Tout vécu de conscience en général est en lui-même conscience de ceci ou de cela, quoi qu'il en soit de la légitime valeur accordée à la réalité de cet objet […]. Tout cogito, tout vécu de conscience comme nous disons également, vise quelque chose, et porte en soi, sur le mode de ce qui est visé, ce qui est dans chaque cas son cogitatum […]. On nomme aussi intentionnels les vécus de conscience, et le terme d'intentionnalité ne signifie rien d'autre que cette particularité fondamentale et générale de la conscience qui est d'être conscience de quelque chose, de porter, en soi, en tant que cogito, son cogitatum. » (Husserl, Méditations cartésiennes, Méditation 2).

III. AMPLEUR PRATIQUE DU PROBLEME : OMNIPRESENCE DES AGENTS ARTIFICIELS QUI SEMBLENT INTENTIONNELS

Avant de poursuivre, il convient, à travers les exemples de cette section, de souligner deux points : 1°) les agents artificiels semblent pouvoir se référer à toutes les sortes d'« objets » intentionnels recensées par la phénoménologie ; 2°) les questions soulevées par cet état de fait ne peuvent pas être évacuées sous prétexte[6] qu'elles ne se poseraient qu'à propos de techniques informatiques marginales ; au contraire, sous des formes et des degrés de complexité variés, ces agents deviennent omniprésents. En outre, les techniques de réalité augmentée, elles aussi en pleine expansion, tendent à faciliter leur apparition.

III.1 L'intentionnalité des agents artificiels dans l'univers symbolique

Notre vie quotidienne est inscrite dans un tissu de systèmes d'information. Impossible de passer à la caisse du supermarché, de téléphoner, de louer une place de concert, d'acheter un billet de train ou d'avion, de circuler dans certains magasins ou quartiers sous surveillance vidéo assistée par ordinateur, ou de pénétrer sur son lieu de travail, sans éveiller un système d'information ou un autre. Au sens le plus simple, un système d'information est un système permettant de relier les membres d'une communauté humaine à travers des programmes informatiques, des procédures et des données partagées. Mais, de plus en plus fréquemment, certains de ces systèmes ont des fonctionnalités tellement élaborées et/ou sont reliés entre eux de manières tellement complexes

---

quand nous disons qu'une machine s'est trompée : une machine ne peut pas « se tromper » au sens usuel, elle ne peut qu'exécuter son programme (« bugs » y compris).

[5] Ce sens n'a pas de rapport simple ou direct avec le sens actuel de « intention ». Il est dérivé du mot *intentio*, qui désigne l'application de l'esprit à un objet, la tension de l'esprit vers un objet.

[6] Ce qui serait, de toutes manières, du point de vue théorique, une très mauvaise raison.

ou peu transparentes, qu'ils nous apparaissent comme capables non seulement d'accomplir des tâches anodines définies par des procédures formalisées, mais aussi de dialoguer et collaborer entre eux et avec nous, de participer avec nous (voire sans nous) à la prise de décisions – bref qu'ils nous apparaissent comme hébergeant des agents artificiels dotés de certaines capacités « mentales ». Les cas d'interaction avec des agents de ce genre se multiplient, en particulier dans les situations d'interaction anonyme auxquelles nous devons réagir sans savoir par quel type d'agent, humain ou artificiel, celles-ci ont été initiées. Si vous êtres flashé par un radar pour excès de vitesse, comment savoir si votre cas a été examiné par un humain ou si l'amende a été expédiée par un processus totalement automatisé ? De même, dans les systèmes d'information complexes mis en place de plus en plus fréquemment dans les « entreprises intégrées » ou dans les administrations, comment savoir de l'extérieur quelles actions sont lancées, quelles décisions sont prises par le système sans qu'il y ait intervention humaine ? En payant l'amende ou la facture ou vos impôts, vous interagissez avec un agent anonyme, sans savoir s'il est réel ou artificiel – et, généralement, sans même qu'il vous vienne à l'esprit de vous poser la question.

A l'extrémité supérieure de l'échelle de complexité, les « agents intelligents »[7] de l'intelligence artificielle[8] (IA) fournissent des exemples d'agents aux capacités évoluées, qui deviennent capables de partager des savoirs avec nous, c'est-à-dire de communiquer et de raisonner à leur propos. Avant qu'une machine ne démontre un théorème mathématique ou ne joue aux échecs, personne n'aurait cru cela concrètement possible[9] – même si l'on peut faire remonter l'idée abstraite à Descartes et Leibniz. Nous en sommes cependant venus à considérer comme banal le fait qu'une machine soit dotée de ces capacités déductives formelles, au point qu'on recourt aujourd'hui, sans (trop d')états d'âme à l'ordinateur comme assistant pour démontrer des théorèmes (comme la fameuse conjecture des quatre couleurs)[10]. En outre, le partage des savoirs a tendance à se renforcer avec l'émergence des technologies du « Web

[7] D'abord appelés « systèmes experts » dans les années 1970, puis « systèmes à base de connaissances » depuis les années 1980, période où ils ont investi tous les secteurs d'activité, les logiciels développés par l'IA sont aussi considérés comme des « agents intelligents » à la suite de Newell (1982). Nous y reviendrons.

[8] Nous nommons IA tout court la discipline technique, sans préjuger de ses possibles présupposés et/ou interprétations psychologiques et/ou philosophiques. Dans ce qui suit, bien que nos réflexions générales puissent s'étendre à d'autres branches de l'IA, nous considérons principalement l'IA au sens originel strict, c'est-à-dire l'« IA symbolique » (par opposition, par exemple, aux réseaux de neurones formels, modèles numériques de fonctionnement du cerveau, ou à l'IA évolutionnaire, qui cherche des modèles de génération de comportements intelligents et qui est basée sur une conception Darwinienne).

[9] L'étonnement suscité parmi les intellectuels, à la fin des années 1950, par la démonstration des premiers théorèmes des Principia Mathematica s'est rapidement transformé pour certains en engouement pour l'IA. Pour le grand public, les jeux (tic-tac-toe, dames, etc.) ont joué un rôle analogue. Un peu plus tard, dans les années 1970, l'engouement a diffusé vers l'industrie avec le succès des « systèmes experts » sur un certain nombre d'applications pratiques, ayant qui plus est rapidement produit des retours sur investissements. Que l'IA soit passée de mode dans les années 1990 ne doit pas faire oublier que ses techniques ont diffusé dans toute l'informatique et que les applications continuent à se développer à un rythme élevé.

[10] Nous disons bien « démontrer » et non « inventer » : les énoncés à démontrer sont donnés au système, duquel on n'attend pas qu'il évalue leur intérêt. Même dans ce cadre étroit, la démonstration automatique de théorèmes n'est pas un exemple d'école, elle a des applications pratiques en génie logiciel : synthèse et vérification de programmes.

sémantique », qui visent à faciliter pour ces agents l'accès aux masses incalculables de connaissances éparses sur le Web. Si ces « agents intelligents » restent encore pour la plupart restreints à des milieux spécialisés[11], contrairement aux exemples précédents, leur retentissement culturel[12] est énorme car ils touchent à des capacités qui ont longtemps été considérées comme le propre de l'homme. Par ailleurs, leur interfaçage[13] avec la technologie des « services Web » leur ouvre les portes du monde des systèmes d'information qu'elle est elle-même en train d'envahir.

Ainsi l'homme du troisième millénaire débutant se trouve-t-il plongé dans un phlogistique informationnel qu'il partage comme l'air avec ses semblables, mais qu'à la différence de l'air, il partage aussi avec une population croissante d'agents artificiels qui semblent bel et bien capables de se référer aux éléments de notre univers symbolique, c'est-à-dire de faire preuve d'intentionnalité par rapport à certains types d'objets mentaux.

III.2 L'intentionnalité des agents artificiels dans un monde virtuel

Mais il existe aussi de très nombreux cas d'agents artificiels capables de se référer à un environnement externe.

Commençons par un exemple fameux, dont on trouve la trace filmée dans « Le seigneur des anneaux ». Dans ce film, les scènes de combat ont en effet été synthétisées à l'aide du logiciel MASSIVE®, dans lequel chaque combattant est un agent autonome «conscient » de sa situation particulière dans le monde virtuel du combat et régi par des règles de comportement tenant compte de cette situation. Le résultat est saisissant de réalisme et se distingue très fortement des scènes de combat avec clones (comme on en voit dans « La guerre des étoiles ») où tous les clones accomplissent les mêmes gestes au même moment[14].

En ce qui concerne l'intentionnalité d'un agent artificiel dans un monde virtuel (au sens technique de la réalité virtuelle – RV), notons qu'un humain fait preuve d'intentionnalité dans un tel monde tout autant que dans le monde réel (puisque la notion d'intentionnalité ne présuppose rien quant à la réalité de l'« objet » visé). La situation de véritable immersion dans un monde virtuel, au sens strict de la RV, reste encore peu fréquente, essentiellement limitée à certains secteurs professionnels, mais les cas d'agents virtuels capables de se référer aux objets d'un monde virtuel abondent (on trouvera sur le Web plusieurs centaines de pages relatives à de tels systèmes). En outre, cette situation présentera à assez court terme un très large intérêt pratique, ne serait-ce qu'à travers les jeux vidéo, qui en donnent déjà une première approximation. Ce secteur majeur de l'économie des nouvelles technologies est aujourd'hui le champ d'une concurrence des plus féroces et fait l'objet d'investissements considérables. La tendance générale des jeux vidéo les pousse vers un réalisme

---

[11] Diagnostic de pannes, aide à la planification, surveillance de processus, synthèse de cartes électroniques, commerce électronique (FishMarket), synthèse d'informations en provenance de sources hétérogènes (MOMIS), etc. On trouve cependant aussi des agents grand public : réservation (de billets d'avion), etc.

[12] Leur impact épistémique a été analysé dans notre livre « Le savoir et l'ordinateur » – S&O par la suite.

[13] Interfaçage déjà réalisé techniquement, par exemple dans la plateforme JADE.

[14] Sur le site Web du film, on peut créer soi-même des situations de combat et générer les scènes correspondantes.

accru ; à terme, avec la chute prévisible des prix des casques de RV et autres systèmes immersifs, ils permettront une véritable immersion sensori-motrice du joueur dans le monde virtuel du jeu, tandis que les personnages simulés se verront doter de comportements réalistes de plus en plus complexes ; en outre, grâce aux technologies déjà pleinement maîtrisées de mise en réseau, plusieurs joueurs réels pourront se retrouver immergés dans un même monde virtuel. Dans ces jeux, auxquels est exposée toute une génération, il pourra être difficile pour un joueur en situation (comme il l'est déjà avec certains jeux actuels) de savoir quel autre personnage du jeu est réel (c'est-à-dire tenu par un joueur réel) ou artificiel (géré par un agent artificiel) et donc de faire une quelconque distinction entre les uns et les autres quant à cette forme d'intentionnalité qu'est leur capacité à se référer à leur environnement.

III.3 L'intentionnalité des agents artificiels dans le monde réel

On objectera : certes, un agent intelligent peut simuler l'intentionnalité dans un univers virtuel auquel il est adapté par construction. Mais, en interagissant avec nous, une machine peut-elle (donner l'impression de) se référer à la réalité – à la même réalité que nous ? Écartons les quelques cas du début de cette section (impôts, etc.), qui visent pour l'essentiel des « objets » de l'ordre symbolique, pour en venir à la question difficile : un agent artificiel peut-il donner l'impression de se référer à des objets physiques du monde réel ? Le cas des radars sur l'autoroute semble montrer que oui, au moins dans des situations fortement contraintes. Des cas plus élaborés de robots dans des laboratoires, de « rovers » sur les planètes lointaines, ou, au Japon, celui des robots ludiques (Aïbo®) ou des hôtesses d'accueil (Repliee®) offrent aussi des exemples encore simples[15], mais multiples, de comportements autonomes en univers réel.

III.4 L'« augmentation » du monde favorise l'apparition d'agents intentionnels

Il est important de noter par ailleurs que le monde ambiant s'adapte à la venue d'agents intentionnels : avec par exemple les technologies dite de « communication ambiante » (puces Rfid, etc.), une espèce d'« Internet des objets » se met en place, par lequel les objets ordinaires deviennent eux-mêmes aptes à se repérer dans l'espace, à détecter les autres objets de leur environnement et à communiquer spontanément entre eux (et avec des agents artificiels) sans notre intervention, sans même que nous en ayons conscience. On songe aux articles achetés au supermarché qui seront répertoriés automatiquement au passage devant la caisse sans avoir à sortir du chariot, ou au réfrigérateur qui détectera automatiquement son contenu, ou à la voiture qui adaptera sa vitesse en fonction des panneaux rencontrés, ou encore à la puce intégrée à votre veste qui vous signalera que vous venez de passer devant un livre ou un CD qui pourrait vous intéresser, ou de croiser une personne qui partage certains de vos goûts et centres d'intérêt. Ces objets communicants, ces objets « augmentés », sont au monde ambiant ordinaire l'analogue de ce que sont les technologies du Web sémantique dans l'univers symbolique : de même qu'il est très difficile pour une machine de « comprendre » le sens d'une page Web en langage naturel, il est très difficile pour un robot de « reconnaître » les objets de son

[15] En raisons de la très grande difficulté du problème de la reconnaissance des formes, les cadres opératoires de ces robots doivent être fortement contraints.

environnement. Et de même que l'on vise à doter les pages Web ordinaires de compléments sémantiques (en spécifiant les ontologies formelles auxquelles elles se réfèrent), ainsi en vient-on à doter les objets matériels ordinaires de la capacité à s'auto inscrire dans l'espace symbolique en diffusant partout le signal de leur présence, de leurs caractéristiques, de leur position. Ainsi, la communication permanente et tous azimuts est-elle en voie de déborder l'univers symbolique pour investir le monde des objets quotidiens, le monde « réel ». Et c'est l'univers ambiant lui-même qui se transforme pour faire en sorte que des agents artificiels puissent s'y référer et se le représenter plus facilement – puissent y faire preuve d'intentionnalité.

IV. QUEL TYPE D'INTENTIONNALITE ACCORDER A UN AGENT ARTIFICIEL

Or, tel est maintenant le problème auquel nous sommes confrontés, problème dont l'actualité et l'acuité se mesurent à la multiplicité des exemples ci-dessus : si nous devions admettre que des agents artificiels sont pourvus d'intentionnalité, pas seulement en apparence mais en réalité (à supposer qu'une telle distinction ait un sens), cela nous inscrirait dans la lignée des premières interprétations anthropocentriques de l'intelligence artificielle (IA) – interprétations qui concevaient l'IA comme visant à *reproduire* dans une machine les capacités et mécanismes mentaux/cognitifs généraux de l'être humain[16] – et nous nous heurterions immédiatement au leitmotiv de ses critiques philosophiques d'origine phénoménologique : une machine n'a pas et ne peut pas avoir d'intentionnalité, l'IA est donc impossible[17]. S'ouvrirait aussitôt une guerre de tranchées idéologique, dont le seul effet possible serait de redoubler la fracture entre sciences dures et sciences humaines. Notons que, bien que ces débats aient initialement surgi à propos de l'IA et que nous développions un certain nombre de nos considérations sur l'intentionnalité dans ce cadre, la plupart d'entre elles s'appliquent plus largement à des agents artificiels qui ne mettent pas nécessairement en œuvre les techniques spécifiques de l'IA, comme dans les premiers exemples de la section précédente.

Se présente bien entendu la solution de s'en remettre au programme fondateur des sciences cognitives et de miser sur la très hypothétique possibilité de « naturaliser » l'intentionnalité, c'est-à-dire d'en fournir une explication qui soit acceptable par les sciences naturelles[18]. Mais, quels que soient l'intérêt

---

[16] Ces interprétations de l'IA constituent l'« IA forte ». Celle-ci s'oppose à l'« IA faible », ou fonctionnalisme, qui considère que l'IA ne vise qu'à simuler des comportements observables, sans se préoccuper de la nature des structures, mécanismes et/ou capacités internes qui en sont à l'origine. L'IA forte est souvent présentée comme une conception archaïque en voie de disparition, mais c'est sous-estimer la multiplicité des formes qu'elle peut prendre, aussi bien dans certains milieux techniques un peu marginaux et/ou au vocabulaire malheureux que dans l'imaginaire collectif ou que chez certains critiques mal informés. Par ailleurs, étant donnés les résultats incontestables de l'IA (technique), tout discours sur l'impossibilité ou sur l'échec de l'IA (et ils restent très nombreux) ne peut porter en fait que sur les interprétations de l'IA forte.

[17] Cette critique philosophique de l'IA est apparue avec Dreyfus, très peu de temps après la discipline elle-même. Elle est de toute évidence tautologique si l'on adopte la définition husserlienne de l'intentionnalité.

[18] Ce qu'il faut entendre précisément par là varie considérablement d'un auteur à l'autre. On trouvera dans Pacherie (1993) une revue critique détaillée des nombreuses positions possibles et de leurs difficultés : réductionnisme physicaliste, fonctionnalisme (qui présente lui-même de multiples variantes), biosémantique, etc. Proust (1997a ; 1997b) constituent un exemple détaillé des multiples tentatives de

considérable de ce programme et sa fertilité, mesurée par la quantité de questions qu'il soulève, ses difficultés non moins considérables n'autorisent guère à préjuger de son succès. L'intentionnalité apparente des agents artificiels nous conduira d'ailleurs à formuler une difficulté d'un nouveau type.

Reste à revenir aux sources techniques et à concevoir l'IA, d'une manière plus prosaïque et plus conforme à ses développements effectifs[19], comme visant à développer des systèmes logiciels autorisant certaines formes d'*interopérabilité sémiotico-cognitive* entre l'homme et l'ordinateur[20] ; à cette fin, dans le processus de leur développement, les comportements de ces logiciels sont méthodologiquement conçus (au sens de «designed ») en termes mentalistes formels, comme des agents artificiels[21] dotés de «connaissances »[22] sur lesquelles ils peuvent «raisonner », pourvus d'« états mentaux » (buts, intentions, croyances, etc.) et capables d'accomplir certains « actes de parole » ; *in fine*, du côté de l'utilisateur, ces systèmes nous apparaissent effectivement comme des *agents* dotés de ces attributs mentaux – ce qui était purement formel au niveau de la conception acquérant un caractère de naturalité au niveau du perçu. Dans cette approche, d'autres questions, jusqu'alors étouffées par l'énormité philosophique des précédentes, surgissent, qui s'avèrent de fait beaucoup plus abordables. Car nous quittons désormais le paradigme frankensteinien, celui de la création démiurgique, lourde de peurs et de culpabilité souterraines, d'un simulacre d'homme à notre image, pour entrer dans le paradigme sémiotique, celui des interactions sémiotico-cognitives que l'homme est susceptible d'avoir avec certains types de machines conçues à cet effet. En particulier, la question de l'intentionnalité des agents artificiels se présente alors fort différemment.

---

naturalisation. Jacob (2004) se focalise sur la question : « l'intentionnalité est-elle la caractéristique propre du mental ? » et montre qu'aucune réponse satisfaisante n'existe aujourd'hui. Noter que, si le projet de naturalisation est compris en un sens réductionniste, il devient très difficile de ne pas devoir attribuer à un agent artificiel le même type d'intentionnalité (réelle) qu'à un humain. Quant à la notion d'émergence à laquelle il est souvent fait appel dans ce contexte, elle nous semble tout aussi problématique que celle d'intentionnalité.

[19] Il est vrai que cette approche parle moins à l'imagination populaire et qu'elle est plus exigeante, en termes de connaissance de la réalité des sciences et techniques ; elle est sans doute aussi globalement plus proche de la vison des ingénieurs en IA que de celle des chercheurs en sciences cognitives. Cette démarche a été introduite dans Berthier (1994) et suivie dans nos livres S&O et MRV.

[20] L'interopérabilité sémiotico-cognitive a été définie dans S&O (chapitre 9 et partie 4) et dans MRV (chapitre 6) comme prolongeant naturellement la notion purement technique d'interopérabilité, désormais classique et omniprésente dans l'univers de l'informatique et des réseaux, où elle exprime un problème majeur du génie logiciel.

[21] Désormais, le « paradigme agent » dépasse largement l'IA et les « agents intelligents » de l'IA sont des cas particuliers d'« agents artificiels ». Tandis que le qualificatif d'« intelligent », si l'on veut (à tort) décortiquer la première expression, apparaît comme une pétition de principe, et que celui d'« artificiel » ne fait que décrire leur mode de production, la vision du système comme *agent* correspond à toute une conception de ce qu'est l'objet de l'IA et elle s'appuie sur toute une méthodologie de développement (Newell, 1982 ; Schreiber & al., 1993 ; Berthier, 1994 ; 2002 ; 2004). Elle est aussi liée au fait que ces agents sont désormais capables de communiquer entre eux par des primitives standardisées de haut niveau conceptuel, inspirées de la théorie des actes de parole d'Austin et Searle, et indépendantes des protocoles réseaux sous-jacents.

[22] Exprimées dans des systèmes formels de représentation, sous une forme de nature fondamentalement logique, à la fois compréhensible par un humain ne sachant rien des subtilités informatiques d'une machine capable de les mettre en œuvre par inférence formelle et exploitable par une machine ne « sachant » rien des subtilités psychiques d'un humain capable de les comprendre.

V. L'INTEROPERABILITE ET LA CATEGORIE DU VIRTUEL

V.1 Les deux formes supérieures d'interopérabilité

Avant d'interpréter l'intentionnalité apparente des agents artificiels, nous devons encore rappeler le parallèle naturel que nous avons établi ailleurs[23] entre l'IA et la réalité virtuelle (RV) : si l'IA vise à développer l'interopérabilité sémiotico-cognitive entre l'homme et l'ordinateur, la RV vise à développer l'*interopérabilité sensorimotrice*. Nous avons introduit cette deuxième forme d'interopérabilité pour regrouper les critères classiques (purement techniques) de la RV[24] : perception spatiale en trois dimensions, immersion sensorimotrice, interaction en temps réel, caractère naturel de l'immersion et de l'interaction. Conjointement donc, l'IA et la RV, ces deux disciplines phares de l'informatique, visent à établir l'interopérabilité entre l'homme et l'ordinateur dans les deux ordres majeurs de l'expérience humaine ordinaire.

Il convient de noter dès à présent que le concept d'interopérabilité, que ce soit dans le sens technique qu'on lui attribue classiquement dans le monde de l'informatique et des réseaux, ou que ce soit dans les sens élargis que nous lui avons donnés, sous-entend toujours un cadre d'opération bien déterminé. Cette restriction essentielle, qui s'applique en particulier à chaque agent artificiel de l'IA et à chaque monde virtuel de la RV, est sans doute drastique, mais c'est elle qui sépare les réalités technologiques de la science-fiction[25] ; elle se trouve explicitement exprimée et formalisée dans toutes les méthodologies de développement d'agents artificiels ou de mondes virtuels[26]. Elle n'implique cependant aucune restriction *a priori* quant à la diversité des cadres d'opération ou à la multiplicité des types d'agents artificiels et de mondes virtuels qu'il est possible de développer. En outre, les contraintes sémiotico-cognitives et/ou sensorimotrices qui conditionnent l'interopérabilité du côté de l'humain sont en général beaucoup moins difficiles à satisfaire qu'on ne pourrait le craindre, car de nombreux facteurs psychologiques (comme la concentration sur la tâche à accomplir) ou sociologiques (de même nature que celles qu'il rencontre dans la vie ordinaire[27]) peuvent contribuer à le maintenir naturellement dans le cadre opératoire fixé.

V.2 La modalité ontologique du virtuel

Outre le parallèle entre l'IA et la RV, établi sur la base de cette notion généralisée d'interopérabilité, nous avons redéfini le virtuel en tant que *modalité ontologique*[28] *générale* devant être clairement distinguée des autres modalités

---

[23] Dans MRV, d'où sont aussi issues d'autres idées de cette section.

[24] Nous parlons ici de la RV au sens originel strict, c'est-à-dire immersive.

[25] On peut considérer cette caractéristique des agents artificiels comme ce qui les sépare fondamentalement de l'humain. Dans un cadre un peu différent, sur l'opposition entre le caractère fini de l'ordinateur et infini de l'esprit humain, voir par exemple (Chazal, 1995).

[26] Il existe bien entendu de nombreuses tentatives pour doter les agents artificiels de capacités sémiotico-cognitives générales (connaissances de sens commun, apprentissage « automatique », etc.) qui leur permettraient d'opérer dans des cadres moins rigidement délimités, mais les succès en la matière restent plus que modérés en termes d'applications opérationnelles.

[27] Par exemple, au cours d'une réunion professionnelle d'une certaine tenue, il y a un ordre du jour et l'on ne se met pas à parler de n'importe quoi.

[28] Il ne faut pas comprendre ce terme au sens d'essences éternelles, indépendantes de nos capacités perceptives et cognitives. Pour nous, rien n'ayant d'essence, l'ontologie est une partie de

de l'être, celles du potentiel, du latent, du possible, et celles de l'imaginaire ou de l'illusoire, toutes modalités avec lesquelles elle est généralement confondue[29]. Pour nous, *est virtuel ce qui, sans être réel, possède avec force et de manière pleinement actuelle (i.e. pas potentielle), les qualités (propriétés ou qualia) du réel*. Nous avons commencé par construire cette définition en nous basant sur l'étymologie (à partir de *virtus* – vertu – et non de *virtualis*, inventé au Moyen Age) et nous avons montré qu'elle est conforme aux usages scientifiques et techniques du terme (en optique géométrique et en RV)[30], donc qu'il existe bel et bien du virtuel en ce sens.

Le reflet dans un miroir (qui est, techniquement parlant, une image virtuelle) constitue, dans la modalité visuelle, un cas prototype de virtuel en ce sens. Pour prendre un autre exemple bien connu hors de la modalité visuelle, chacun d'entre nous peut constater (avec une bonne chaîne audio ou dans une bonne salle de cinéma) combien il est facile de fabriquer des sons virtuels, c'est-à-dire des sons qui nous semblent provenir d'un point de la salle d'où aucun son n'est réellement émis ; il est également très facile de produire pour ces sons l'impression du mouvement de leur point d'émission[31].

Il s'ensuit tautologiquement que, dans la (ou les) modalité(s) concernée(s), *le virtuel ne peut pas être distingué du réel par des critères généraux* (ce qui n'implique pas que la distinction soit inopérante, car il peut en général être repéré sans difficulté, dans chaque situation concrète, par des critères spécifiques – comme de sortir du cône de visibilité du reflet dans le miroir ou de recourir à une autre modalité). Il s'ensuit aussi que *la capacité du virtuel à s'imposer à nous dans la (ou les) modalité(s) concernée(s), (comme un reflet ou un son), voire à se saisir de nous (comme un monde virtuel), devient compréhensible en principe*, alors qu'elle reste totalement inexplicable (et fortement contradictoire) dans les conceptions, héritées de la scolastique du Moyen Age et aujourd'hui largement répandues, pour lesquelles le virtuel serait en attente d'actualisation[32].

En outre, l'exemple du reflet dans un miroir illustre parfaitement le fait général que *d'un objet virtuel peuvent être issus des effets réels* (comme les rayons réfléchis), de sorte que la perception qu'on en a et toute notre relation à lui sont bien réelles, comme le sont (dans le champ visuel) celles du reflet ou (dans le champ auditif) celle du son virtuel. Le monde dans lequel nous nous

---

l'épistémologie – la partie qui concerne le recensement des diverses catégories d'« objets » et les modes d'être de ces « objets ». Nous adoptons ici un point de vue élémentaire, consistant à admettre que notre expérience nous fait distinguer divers modes d'être (pour le formuler en termes husserliens), dont les exemples cités.

[29] Ainsi, par exemple, aucune de ces deux références majeures que sont le Trésor de la Langue Française et l'Encyclopædia Universalis ne permet d'établir une distinction claire entre le virtuel et le potentiel.

[30] Cela a été établi dans MRV, respectivement aux chapitres 3 et 4.

[31] Ces effets sont obtenus en jouant sur différents paramètres physiques : intensités relatives dans les haut-parleurs, rapport d'énergie entre son direct et son artificiellement réverbéré, effet Doppler (changement de fréquence lié au mouvement). Ces effets peuvent être transposés dans un casque, en jouant sur les décalages temporels entre ce qui parvient aux deux oreilles.

[32] Comme celle de Deleuze (1968a ; 1968b) – et, à sa suite, celle de Pierre Lévy – pour qui le réel s'oppose au possible, et, au sein du réel, le virtuel s'oppose à l'actuel. Pour nous, l'actuel s'oppose au potentiel, et, au sein de l'actuel, le virtuel s'oppose au réel. Quant à la conception de Bergson (1939), dont Deleuze se veut le continuateur sur ce point, nous n'en parlerons pas ici, le terme « virtuel » y étant utilisé sans définition explicite, mais toujours avec le sens implicite de latent ou de potentiel.

trouvons immergés à un moment donné peut être virtuel, il n'en reste pas moins que les expériences mentales que nous y vivons et les émotions que nous y ressentons sont bien réelles et ont sur nous des effets bien réels, y compris d'ordre physique. Ce fait évident a des applications pratiques : on l'utilise, par exemple, avec succès pour traiter des phobies en exploitant les techniques de la réalité virtuelle ; ainsi, dans le cas de l'agoraphobie, amène-t-on les patients à sortir dans la rue, dans un monde virtuel. La conception courante du virtuel comme étant en attente d'actualisation ne peut que rendre ces effets totalement incompréhensibles. Ces remarques illustrent aussi à quel point la notion courante d'illusion est impropre à saisir les subtilités du virtuel.

V.3 Lien entre virtuel et interopérabilité

Notons que *les notions d'interopérabilité et de virtuel sont indissociables : le virtuel est la modalité ontologique fondamentale pour la description phénoménologique de situations qui peuvent aussi être expliquées de manière analytique avec le concept plus technique d'interopérabilité* – de la même manière que le reflet est le concept permettant la description phénoménologique d'une situation qui peut être analysée de manière plus savante en termes de rayons incidents et réfléchis.

V.4 Agents virtuels

Dès lors, la boucle se referme : un agent artificiel est un *agent virtuel*. Les deux qualificatifs ne sont pas homologues : alors que celui d'artificiel est purement technique et factuel, comme il a déjà été indiqué, celui de virtuel véhicule ici une affirmation précise sur le mode d'être de cet agent. En effet, grâce à ses capacités techniques d'interopérabilité sémiotico-cognitive, il *semble naturellement* faire preuve, dans le contexte d'opération bien délimité qui est le sien et pour des humains en situation d'interaction avec lui, des facultés sémiotico-cognitives elles aussi bien délimitées qui seraient celles d'un agent humain assigné à l'accomplissement des mêmes tâches. Dans semblables contextes, de tels agents ne peuvent pas être distingués d'un agent humain, du point de vue de leur interlocuteur en situation[33].

Insistons : c'est en raison de ses capacités techniques d'interopérabilité sémiotico-cognitive qu'un système d'IA apparaît comme un agent virtuel ; de même, c'est en raison des possibilités d'interopérabilité sensorimotrice avec un système de RV que nous avons l'impression d'immersion dans un monde virtuel.

VI. PLUS FAIBLE QUE L'IA FAIBLE

Selon notre interprétation de l'IA, les questions soulevées par l'IA forte ne sont pas complètement éliminées, mais elles se trouvent reformulées. *La notion d'agent virtuel désamorce automatiquement les objections philosophiques habituellement soulevées à l'encontre de l'IA, puisque, en même temps que l'agent lui-même, tous ses attributs cognitifs ou mentaux, et en particulier une*

[33] Ces affirmations ne présupposent pas que le test de Turing (1950) puisse être passé avec succès. Pour nous, la formulation habituelle de ce test est liée à une conception ancienne et très ambitieuse de l'IA, en termes de capacités cognitives générales. Il est beaucoup trop vague et ne tient pas compte des conceptions plus récentes et plus terre à terre, qui présupposent la notion de cadre opératoire limité.

éventuelle intentionnalité, devront être systématiquement qualifiés de virtuels. Et il ne s'agit pas là d'une simple pirouette linguistique, puisqu'un sens précis a été attribué à ce terme – un sens qui nous a contraint à remonter jusqu'au niveau des modalités majeures de l'être : du réel, de l'actuel, du potentiel, du possible, etc. Cela s'applique également aux connaissances de l'agent : bien que nous puissions interopérer avec, nous n'avons pas besoin d'admettre qu'un ensemble de règles d'expertise en Prolog, par exemple, serait de la connaissance au même sens réel qu'une connaissance humaine[34].

Qualifier de virtuelle l'intelligence, l'intentionnalité ou la conscience d'un agent artificiel permet de poursuivre les recherches en ces domaines sans pour autant prétendre reproduire l'intelligence, l'intentionnalité ou la conscience humaines (même si l'on s'inspire de certains de leurs aspects) ; cela conduit à s'interroger sur les facteurs qui nous conduisent à accorder la réalité de ces attributs plutôt que seulement l'apparence ; et cela amène finalement à distinguer interopérabilité et intersubjectivité, tout en reconnaissant que les circonstances produites par la première peuvent nous amener à un point où apparaît la tentation intellectuelle d'admettre l'autre, comme le font les tenants de l'IA forte ; à notre avis, seule la catégorie du virtuel peut actuellement justifier, autrement que pour des motifs religieux ou philosophiques, ou par évocation à l'« intuition », qu'on repousse cette tentation.

Or, tout en désamorçant l'IA forte, notre interprétation ne s'identifie pas non plus à l'IA faible. S'il est clair que l'interopérabilité ne présuppose pas l'identité structurelle, elle ne présuppose pas non plus l'identité fonctionnelle. En parlant avec un autre humain, j'interopère avec lui dans l'ordre sémiotico-cognitif ; rien cependant ne permet d'affirmer que nous soyons fonctionnellement identiques en toutes circonstances dans cet ordre ; rien même ne permet d'assigner un sens précis à cette expression[35]. De même, pour tout observateur humain en situation, le reflet d'un objet dans un miroir opère dans la modalité visuelle comme un objet réel – et cela, de manière objective, totalement indépendante de cet observateur[36] ; néanmoins, pour un observateur hors situation, il n'en va pas de même[37]. L'interopérabilité, quoique objective pour tout observateur humain en situation (au sens où elle ne dépend pas de sa subjectivité individuelle ou de son imagination), ne l'est *que* pour un observateur humain en situation.

Notre interprétation de l'IA comme visant à développer l'interopérabilité sémiotico-cognitive entre l'homme et l'ordinateur *est donc plus « faible » que les formulations usuelles du fonctionnalisme ou IA faible* : au lieu de nous intéresser aux « comportements observables » du système en général (c'est-à-dire objectivement, par tout moyen externe d'observation imaginable), nous

---

[34] Voir plus haut – note 22 – comment nous avons formulé la différence avant de disposer du concept de virtuel.

[35] On connaît d'ailleurs les difficultés considérables auxquelles se heurte toute tentative de formaliser le fonctionnalisme (voir Pacherie, 1993). Et si toutefois un sens clair venait à être défini, il est plausible qu'on doive conclure que deux humains quelconques ne sont pas fonctionnellement identiques dans cet ordre.

[36] L'observateur peut même être une machine basée sur certains mécanismes perceptifs compatibles avec la perception humaine, comme un appareil photo autofocus, dont on sait qu'il fait la mise au point par analyse des micro-contrastes.

[37] Et il n'en va pas de même non plus pour un appareil basé sur d'autres mécanismes perceptifs, comme un système qui ferait le point par télémétrie laser.

nous intéressons seulement aux comportements du système observables par un sujet humain en situation d'interopérabilité avec lui, dans un cadre opératoire clairement prédéfini. Soulignons qu'il y a là deux restrictions complémentaires, de natures différentes : 1°) la première concerne l'agent artificiel et porte sur son cadre opératoire délimité[38] ; 2°) la deuxième restriction concerne l'observateur, dont on fait explicitement un agent humain en situation ; par exemple, bien qu'objective et pleinement actuelle, l'intentionnalité de l'agent artificiel ne sera perçue en tant qu'intentionnalité qu'en raison de nos capacités sémiotico-cognitives humaines à la percevoir comme telle dans les conditions adéquates, tout comme un reflet dans un miroir n'est perçu qu'en fonction des capacités visuelles spécifiques à notre espèce (une mouche ne voit pas le reflet) et lorsque nous sommes dans le cône de visibilité.

Ainsi, notre conception explicite-t-elle le fait que l'IA doit bel et bien exploiter au maximum la spécificité de nos capacités sémiotico-cognitives humaines, quoiqu'elle ne doive pas le faire pour tenter de les *reproduire* de manière générale, selon un mythe initial qui persiste dans l'imaginaire collectif, mais pour tenter de les *leurrer* dans des situations particulières prédéfinies[39]. Cela constitue une différence notable, tant sur le plan théorique de l'épistémologie que sur le plan pratique des méthodes de développement de systèmes d'IA.

VII. PLUS FORT QUE LA « POSTURE INTENTIONNELLE » DE DENNETT

Les concepts d'agent virtuel et d'intentionnalité virtuelle permettent non seulement de préciser en quel sens nous «prêtons » de l'intentionnalité à un agent artificiel – pour reprendre l'expression de Dennett (1987) ; ils permettent aussi de comprendre pourquoi nous n'avons pas le choix, pourquoi son intentionnalité s'impose à nous (comme le fait un reflet dans un miroir), y compris dans les cas où nous admettrions par ailleurs, intellectuellement, qu'il n'en a pas vraiment.

Pour Dennett, la *posture intentionnelle* (« intentional stance ») est la *stratégie* consistant à *interpréter* le comportement d'un objet (vivant ou pas) *comme si* c'était un agent rationnel dont les actions sont déterminées par ses croyances et désirs. Mais cette manière de présenter les choses, qui s'apparente à la démarche de modélisation, laisse entendre, jusque dans le vocabulaire utilisé (mots en italiques ci-dessus), que nous pourrions avoir le choix de résister à l'adoption spontanée de cette posture : il y a dans le mot « stratégie », ainsi que dans l'anglais « stance », ou dans sa traduction française « posture » (qui nous semble plus juste qu'« attitude »), l'idée d'une attitude non naturelle. Or, la « posture intentionnelle » est justement celle que chacun d'entre nous adopte spontanément quand il parle de lui-même ou de ses semblables. De ce point de vue, la formulation de Pierre Jacob est plus explicite, pour qui, pas plus qu'à la physique naïve, nous ne pouvons échapper à la psychologie naïve :

---

[38] On peut admettre que le fonctionnalisme partage cette première restriction, la notion de « cadre opératoire » étant implicitement incluse dans celle de « fonction » du système.

[39] En l'état actuel des techniques, l'effet de leurre pourrait ne pas durer très longtemps si la communication avec l'agent artificiel n'était pas fortement contrainte. Le cadre opératoire doit, entre autres, contribuer à maintenir cet effet. Noter qu'il suffit parfois que l'effet se produise initialement pendant un temps assez court pour qu'une conséquence ou un acte irréversible s'ensuive, d'où peut résulter un certain « verrouillage » de l'acteur humain dans le système.

> « Chaque être humain est tellement prédisposé par la psychologie naïve à concevoir ses actions et celles d'autrui comme le résultat de ses buts, intentions, désirs et croyances que *le moindre comportement* non humain est *irrésistiblement* interprété comme celui d'un agent doué d'une intention ou d'un but. » (Jacob, 2004, p. 13 – c'est nous qui soulignons)

Cependant, cela ne nous satisfait pas encore totalement. Certes, puisque la physique naïve nous pousse irrésistiblement à percevoir le monde réel en termes d'objets, nous pouvons considérer, par extension, qu'elle nous pousse aussi irrésistiblement à réifier le reflet. Mais seule l'analyse du reflet comme image virtuelle nous permet de comprendre pourquoi il doit en être ainsi, pourquoi notre perception de lui comme objet n'est en aucune manière altérée quand nous savons que ce n'est pas un objet réel, pourquoi, en somme, l'ontologie de la physique naïve surgit de nos facultés sensorimotrices bien avant de se transcrire dans l'ordre sémiotico-cognitif. De la même manière, la psychologie naïve nous pousse irrésistiblement à percevoir de l'intentionnalité chez un agent artificiel, mais seule la notion d'intentionnalité virtuelle nous permet de comprendre pourquoi il doit en être ainsi, pourquoi nous ne pouvons pas choisir de ne pas lui en « prêter ». Dans chaque modalité, sensorimotrice ou sémiotico-cognitive, il se trouve que les critères de l'interopérabilité sont aussi ceux qui constituent la trame du domaine de la physique naïve ou de la psychologie naïve. Que le retissage permanent du monde ne puisse plus être isolé de celui de ses simulacres informatiques prend ici un sens bien précis – un sens qui, en plaçant au centre les capacités sensorimotrices et sémiotico-cognitives humaines (avec toutes les possibilités de leurre qu'elles recèlent), respecte l'esprit de la « révolution copernicienne » opérée par Kant, mais en l'épurant de ses *a priori* transcendantaux, comme le font de leur côté les sciences cognitives – un sens aussi qui tient à l'écart la tentation nihiliste (par exemple de Baudrillard, 1981).

VIII. DEFI AU PROJET DE NATURALISATION DE L'INTENTIONNALITE

La conception du virtuel et de l'intentionnalité virtuelle qui vient d'être présentée semble poser un défi nouveau aux sciences cognitives, auquel nous n'avons pas de solution à proposer : *comment interpréter l'objectif de « naturalisation » de l'intentionnalité si rien ne permet de distinguer formellement, par des critères généraux, le réel du virtuel – en particulier, une intentionnalité réelle d'une intentionnalité virtuelle ?* Il semble en effet qu'on doive adopter l'une ou l'autre des deux positions suivantes :

1°) ou bien l'on admet qu'il n'y a définitivement pas lieu de distinguer entre intentionnalité réelle et intentionnalité virtuelle[40] et qu'en conséquence la même explication « naturalisée » s'appliquera aux deux, ce qui laisse à nouveau deux possibilités opposées (qui s'appliquent également en dehors de tout projet de naturalisation) :

1a) les agents artificiels ont réellement de l'intentionnalité ;

1b) notre intentionnalité à nous humains est virtuelle ;

[40] Cela ne s'applique qu'à l'intentionnalité (et aux concepts mentalistes) et n'implique pas qu'on doive renoncer en général à la distinction que nous avons établie entre le réel et le virtuel, suffisamment étayée par ailleurs (voir MRV).

2°) ou bien l'on admet que, bien qu'elles ne puissent pas être distinguées par des critères généraux, intentionnalité réelle et intentionnalité virtuelle peuvent l'être dans chaque situation spécifique au niveau des explications produites par la « naturalisation ».

En dépit de la forme logique disjonctive anodine de cette présentation, chaque possibilité est lourde d'associations et/ou d'implications. Passons sur l'hypothèse (1a), qu'on peut assimiler à l'IA forte.

L'hypothèse (1b) évoquera sans doute, pour les chercheurs en sciences cognitives, les considérations de Varela au sujet de notre Moi, qu'il qualifie de virtuel ; cette idée du Moi virtuel traduirait certaines doctrines asiatiques, comme la théorie de l'*anâtman* (non-Moi ou non-Essence ou encore « vacuité » – centrale dans le bouddhisme Mahâyanâ, et développée en particulier dans l'école Mâdhyamika de Nagarjuna)[41,42]. Dans le contexte occidental des «Méditations cartésiennes » qui, pour beaucoup, sera plus familier, elle évoque une remarque[43] au sujet du mot « aussi », dans la phrase suivante de Husserl (fin du §15) : « moi qui suis dans l'attitude naturelle, je suis aussi et toujours le je transcendantal, mais je ne le sais qu'en opérant la réduction phénoménologique ». Ingarden signale :

> « … la grande difficulté que personne, à ma connaissance, n'a encore indiquée : comment le je pur constituant [i.e. l'*ego* transcendantal de la phénoménologie] et le je ontique constitué [i.e. le Moi ordinaire] peuvent être en même temps un et le même si les propriétés qui leur sont attribuées s'excluent mutuellement […]. Ce n'est que si on tenait d'emblée le je constitué pour une *illusion* – tout comme la totalité du monde ontique (*real*) constitué – que cette difficulté pourrait être résolue au sens où n'existerait que le je pur, et où le je ontique ne serait qu'une fiction transcendant le je pur, bien que prescrite par le cours de ses expériences. »

L'hypothèse du Moi-illusion envisagée par Ingarden semble très proche des théories bouddhiques ou de la position de Varela ; Ingarden montre aussitôt en quoi Husserl s'en distingue fondamentalement, lorsqu'il précise : « Mais Husserl protesterait vivement contre une telle interprétation de l'idéalisme [transcendantal] selon laquelle le constitué serait identifié à une fiction ».

---

[41] Ces doctrines sont indissociables de traditions spirituelles et de pratiques dont les premières étapes (consistant à adopter une attitude de « présence attentive » ou une position de « spectateur impartial » ou de « spectateur désintéressé » – *turiya* dans les traditions hindoues) constituent une sorte d'*époché* phénoménologique. A cette nuance capitale près qu'elle sont expérimentées dans la méditation plutôt que pensées intellectuellement. Semblables rapprochements doivent donc être considérés avec la plus grande prudence. D'autant que l'objectif Husserlien de « fonder » la philosophie, et toute sa problématique de la constitution, restent fondamentalement étrangers à ces doctrines.

[42] Varela (1992) ne parle que de la virtualité du Moi, pas de l'intentionnalité. Il emploie ce terme de « virtuel » en un sens informel, mais nous considérons que notre définition permet de donner sa pleine signification à sa conception. Par ailleurs, si l'hypothèse de virtualité de l'intentionnalité *évoque* l'idée de virtualité du Moi et justifie les remarques de ce paragraphe, nous ne suggérons pas qu'inversement la virtualité du Moi doive entraîner celle de l'intentionnalité (ni, *a fortiori*, comme Ingarden semble le déduire un peu vite, celle du « monde ontique constitué » – voir suite du texte) ; l'intentionnalité pourrait bien être comme un rayon lumineux issu d'une image virtuelle : virtuel sur une partie de son trajet, réel sur l'autre. Cependant, si l'on revient à la philosophie Mâdhyamika, ce sont bel et bien trois « vacuités » qui y sont affirmées : du sujet, de l'objet ET de la relation (aspect qui est souvent oublié). Toute assimilation hâtive mise à part, la relation en question ici ressemble à s'y méprendre à l'intentionnalité, auquel cas Varela aurait dû étendre la virtualité du Moi à celle de l'intentionnalité.

[43] Roman Ingarden : « Remarque à propos de la page 75 » (reproduite dans Husserl, 1994).

Laissons ouvert le débat[44]. Le cours de ces réflexions n'aurait-il pas été différent si, au lieu des termes fortement connotés d'« illusion » et de « fiction », on avait employé celui de «virtuel» – en rappelant qu'il n'y a rien, dans notre conception du virtuel, qui relève de l'illusion ou de la fiction au sens courant de ces termes.

L'hypothèse (2), quant à elle, pose de très fortes contraintes sur ce qu'on peut accepter comme explication « naturelle » de l'intentionnalité ; non seulement devra-t-on expliquer l'intentionnalité sur la base des spécificités (par exemple biologiques, phylogénétiques, etc.,) de notre espèce (ou de celles auxquelles on attribue de l'intentionnalité), encore devra-t-on le faire d'une manière qui ne puisse pas s'appliquer à des artefacts simulant les spécificités invoquées. Il en résulte qu'on voit mal comment des approches très générales, comme par exemple la Morphodynamique de Thom et Petitot, pourraient permettre d'établir ces distinctions. Jusqu'à quelles profondeurs d'organisation de la matière faudra-t-il alors pousser l'idée d'une « inscription corporelle de l'esprit »[45], quelles spécificités de notre (ou nos) mode(s) de présence au monde faudra-t-il invoquer, pour atteindre un tel objectif ?

Cette hypothèse (2) suggère une remarque complémentaire sur le projet de « naturalisation », remarque qu'on comprendra mieux à partir d'une analogie optique. Nous savons tous que, pour les images de l'optique géométrique (disons les images produites par des jeux de lentilles), la nature réelle ou virtuelle de l'image A' d'un « objet » A peut être déterminée (hors situation) en fonction de la position de A par rapport au foyer. Or, en raison de la loi de propagation inverse, cette affirmation banale suppose une hypothèse implicite : nous ne pouvons en effet faire la distinction pour A' que si nous avons un point de référence pour le réel, soit l'information sur la nature réelle ou virtuelle de A, soit la place de l'observateur – ce fameux œil qui apparaît dans les livres de physique du collège et dans lequel les philosophes se plaisent à reconnaître le symbole du Sujet cartésien. Transposé du côté de l'intentionnalité, cela signifie métaphoriquement que la différence entre intentionnalité réelle et virtuelle ne peut avoir cours que du point de vue d'un *ego* qui s'affirme *a priori* comme référence ultime. Autrement dit, le projet de « naturalisation » ne présuppose-t-il pas la conception phénoménologique de l'*ego* transcendantal qui a placé l'intentionnalité sur le devant de la scène philosophique ?

IX. CONCLUSION

Bien qu'à l'origine purement philosophique et relative aux seuls humains, la question de l'intentionnalité s'impose à notre attention lorsque nous sommes confrontés à des agents artificiels qui en semblent pourvus, et elle s'impose d'autant plus vigoureusement que ceux-ci deviennent omniprésents dans un monde qui est lui-même en permanence « augmenté » pour mieux les accueillir. Notre définition précise du virtuel permet de considérer ces agents artificiels comme des agents virtuels et d'affecter le même qualificatif à tous les concepts mentalistes sous lesquels nous les percevons spontanément (connaissances, buts, intentions, actes de parole, etc.), dissolvant ainsi automatiquement les objections usuelles à l'IA «forte». Notre interprétation de l'IA (et plus

[44] Aborder sérieusement cette question obligerait à analyser les multiples critiques, variantes et reconstructions de la phénoménologie et de la notion d'intentionnalité.

[45] Pour reprendre l'expression de Varela & al. (1993).

généralement de tous les agents artificiels, qu'ils recourent ou non aux techniques d'IA) à partir du virtuel, plus faible que l'IA « faible », est cependant plus « forte » que la posture intentionnelle de Dennett, le fait que le virtuel soit pleinement actuel permettant de comprendre en particulier pourquoi l'apparence d'intentionnalité de ces agents s'impose à nous avec autant de force qu'un reflet dans un miroir. Enfin, notre approche pose un nouveau défi au projet de naturalisation de l'intentionnalité (et des concepts mentalistes associés).

Il n'est pas anodin que l'informatique, à travers l'interopérabilité qu'elle développe progressivement entre l'homme et l'ordinateur dans les deux ordres majeurs de l'expérience ordinaire, sensorimoteur et sémiotico-cognitif, nous amène à pousser toujours plus loin la quête de notre identité[46].

Il est clair, par exemple, grâce à l'IA, que le raisonnement déductif formel, dont on a pu faire un temps la gloire de notre espèce, ce qui nous distinguait des animaux, et qui passe malheureusement encore aux yeux de beaucoup pour l'essence des mathématiques, avec toutes les conséquences pédagogiques catastrophiques que l'on peut imaginer, n'est plus l'apanage de l'homme puisqu'il est accessible à une machine ; cela conduit immanquablement à réviser l'idée qu'on s'est longtemps faite de l'intelligence et à accorder plus de crédit aux psychologues, eux qui savent depuis longtemps qu'il en existe plusieurs formes. Pourtant, on peut se demander, inversement, si l'expansion prévisible des agents artificiels ne conduira pas de fait à un certain « formatage » de la pensée[47], à la suprématie d'une forme hautement formalisée et appauvrie de « grammaire de l'intelligence », au sens de Ferry (2004).

De la même manière, en paraissant dotés d'intentionnalité, les agents artificiels nous conduisent à nous interroger un peu différemment sur notre (ou nos) mode(s) de présence au monde. En particulier, nous avons montré comment les différentes attitudes possibles vis-à-vis de la prise en compte du virtuel dans le projet de naturalisation de l'intentionnalité distribuent certaines positions philosophiques – comme l'IA forte, la conception du Moi de Varela ou la phénoménologie – et comment elles lui imposent de fortes contraintes.

BIBLIOGRAPHIE

Baudrillard, J. (1981). *Simulacres et Simulation*. Galilée, Paris.

Bergson, H. (1896/1970). *Matière et mémoire*. in Œuvres, PUF, Paris (aussi disponible sur le Web).

Berthier, D. (1994). L'agent rationnel abstrait, objet de l'IA. *Revue d'intelligence artificielle*, Vol 8, n° 4, pp 327-359.

Berthier, D. (2002). *Le savoir et l'ordinateur*. L'Harmattan, Paris.

[46] Sur ces points, notre approche ne s'oppose pas radicalement à la « philosophie de l'analogie » prônée par Chazal (1995), consistant à rechercher en quoi nous ressemblons à l'ordinateur, et, au-delà, en quoi nous en différons. Ces questions peuvent en effet constituer la première étape légitime d'une interrogation philosophique. Nous pensons cependant avoir défini une épistémologie plus complexe du virtuel et de l'interopérabilité, qui n'est pas soumise à cette idée préalable de ressemblance pour analyser nos modes d'interaction avec l'ordinateur.

[47] C'est un risque que nous évoquions déjà dans S&O et dont les manifestations effectives restent en grande partie à explorer. Ce risque illustre le point soulevé dans la note précédente : c'est l'interaction répétée qui provoquerait le formatage, pas une ressemblance définie au préalable. La mise en œuvre concrète de nos concepts suppose donc la prise en compte de la dimension socio-historique de l'informatique.

Berthier, D. (2003). La culture et la rationalité modernes, du structuralisme à l'IA symbolique, *Automates Intelligents*, n° 42, Paris, Juin.
Berthier, D. (2004). *Méditations sur le réel et le virtuel*. L'Harmattan, coll. « Impacts des Nouvelles Technologies », Paris.
Berthier, D. (2004-2005). Penser notre relation à la Machine, plutôt que nous penser comme des machines, *Terminal*, n° 92, Hiver.
Breton, Ph. (1996). *A l'image de l'homme*. Seuil, Paris.
Chazal, G. (1995). *Le miroir automate, introduction à une philosophie de l'informatique*. Éditions Champ Vallon, Seyssel, France.
Deleuze, G. (1968). *Le bergsonnisme*. PUF, Paris.
Deleuze, G. (1968). *Différence et répétition*. PUF, Paris.
Dennett, D. (1987). *The intentional stance*. MIT Press, A Bradford Book, Cambridge, MA.
Ferry, J.-M. (2004). *Les grammaires de l'intelligence*. Editions du Cerf, Paris.
Husserl, E. (1931/1994). *Méditations cartésiennes*. Armand Colin, PUF, Paris.
Jacob, P. (2004). *L'intentionnalité*, Odile Jacob, Paris.
Nagarjuna, (1995). *Traité du Milieu*. trad. fr., Seuil, Paris.
Newell, A. (1980). Physical Symbol Systems, *Cognitive Science*, Vol 4, pp 135-183.
Newell, A. (1982). The Knowledge Level, *Artificial Intelligence*, Vol 59, pp 87-127.
Pacherie, E. (1993). *Naturaliser l'intentionnalité*. PUF, Paris.
Proust, J. (1997a). *Comment la pensée vient aux bêtes*. Gallimard, Paris.
Proust, J. (1997b). *Perception et intermodalité*. PUF, Paris.
Turing, A. (1950). Computing machinery and intelligence, *Mind*, LIX (236).
Varela, F., Thomson, E. & Rosch, E. (1993). *L'inscription corporelle de l'esprit*. Seuil, Paris.
Varela, F. (1996). *Un know-how per l'ética*. Editori Laterza, Bari, 1992 ; trad. fr. : *Quel savoir pour l'éthique ?* Éditions La Découverte, Paris, 2004.