
CE QUI FAIT ENCORE CRUELLEMENT DEFAUT A L'INTELLIGENCE ARTIFICIELLE

Paul Jorion

MSH Paris

Bien entendu nous aimerions être plus avancés que nous ne le sommes aujourd'hui . Des progrès ont été réalisés et très loin d'être négligeables. Mais par rapport aux ambitions affichées à l'époque du Handbook of Artificial Intelligence [Barr & Feigenbaum 1981, 1982 ; Cohen & Feigenbaum 1982], ou de PDP [Rumelhart & McLelland 1986 ; McLelland & Rumelhart 1986], il s'agit bien seulement de pas de souris.

Nous mimons l'intelligence sous quelques-uns de ses aspects et prétendons quelquefois n'avoir jamais rien visé d'autre. Mais lorsque nous nous contemplons le matin dans le miroir, il nous faut bien admettre que nous avons en tête l'émergence d'un sujet au sein de la machine. Ce que nous imaginions, c'était une machine dont l'expression manifesterait au-delà de tout doute possible l'existence d'une individualité - sinon d'une conscience.

Ce n'est donc pas par hasard si une partie du débat — par opposition aux réalisations effectives — s'est déplacé récemment vers cette question - même de la conscience. Et que l'on nous annonce la conscience pour demain — de la même manière que l'on nous annonçait, il y a quarante ans, l'intelligence pour demain.

S'il me fallait mettre une étiquette sur ce qui fait encore cruellement défaut à l'Intelligence Artificielle, je dirais une "théorie de la signification". Par, là, je n'entends pas une Linguistique, ni non plus une Logique, ni même une Psycholinguistique, mais une théorie qui nous explique ce qui se passe quand je dis "Est-ce que tu pourrais ouvrir la fenêtre ?" et que, sans même m'écouter, machinalement comme on dit, vous tendez le bras et ouvrez la fenêtre — sans même m'avoir réellement entendu.

Mais cela, me répondez-vous, cela nous sommes arrivés à le réaliser au cours de ces quarante années qui nous séparent de la Dartmouth Conference, nous possédons des robots qui ouvrent la fenêtre quand on leur demande de le faire. C'est vrai, et c'est pour cela qu'il n'y a pas lieu du tout de se décourager : nous en savons en effet suffisamment sur la signification pour réaliser cela. Rien n'aura été aussi inutile au fil des années que les distractions causées par l'un ou par l'autre, qui se fondant sur l'herméneutique, qui sur le théorème de Gödel, ou sur que sais-je encore, déclarèrent solennellement que, la preuve était désormais faite — par eux — que l'Intelligence Artificielle n'aurait pas lieu.

S'il avait sans doute été imprudent d'annoncer pour l'IA une date précise, combien plus imprudent n'est-il pas de se porter garant aujourd'hui de son impossibilité irrémédiable ? Le risque encouru par John McCarthy est de laisser son nom dans l'histoire comme celui d'un enthousiaste qui se laissa aller à des affirmations

prématurées. C'est là un risque sympathique et que l'on trouve mentionné dans les livres d'histoire avec mansuétude, mais il n'existe aucun livre qui établisse la liste de tous ceux qui affirmèrent un jour — preuves à l'appui — que l'homme n'irait jamais dans la lune. Un tel livre serait un très lourd bottin que nul ne songerait à consulter.

L'algorithme qui définit les procédures permettant à un robot d'ouvrir une fenêtre n'est en rien mystérieux. Ils n'a nécessité aucune révolution majeure dans le savoir, dans la manière dont nous concevons l'humain — et que constituerait a contrario une explication convaincante de ce que c'est que la conscience. Et j'entends par là une explication de la conscience qui en offre un modèle formel, et non ce que l'on nous propose ces jours derniers, que la conscience est bien tombée du ciel comme nous l'avions toujours soupçonné, ou qu'il ne s'agit que d'une illusion. Car s'il s'agit d'une illusion il faut nous dire alors pourquoi nous en sommes si unanimement les victimes, et l'explication requise n'a fait que se déplacer d'un cran.

Alors prenons un exemple un peu plus complexe dans la signification que la fenêtre qu'il s'agit d'ouvrir. Imaginez que vous émergiez de l'évanouissement après un accident affreux dont vous sortez miraculeusement indemne et qu'un visage penché sur vous prononce alors d'une manière très grave les mots suivants : " *Vous rendez vous compte que si vous êtes vivant, c'est parce que Jésus vous aime ?*". Imaginez aussi qu'ébranlé par ces paroles tout empreintes d'une puissante signification, vous dont la religion se limitait jusque-là à vous souvenir le premier décembre que Noël est pour bientôt, vous subissiez aussitôt une étonnante métamorphose dans la manière dont vous posez votre regard sur le monde et que chacun de vos gestes acquière désormais une nouvelle signification.

D'un tel phénomène, d'une authentique conversion résultant du pouvoir des mots, nous sommes, cela va sans dire, incapables aujourd'hui de définir l'algorithme. Nous ne sommes pas à même de déterminer la fonction mathématique dont les valeurs en entrée est la phrase prononcée par votre mystérieux interlocuteur et dont les valeurs en sortie sont les modifications de vos prédispositions, dont votre comportement ultérieur portera désormais la marque indélébile.

Mais s'agit-il dans l'exemple de la fenêtre à ouvrir et de la conversion religieuse, de deux phénomènes entièrement distincts ? Personnellement je ne le pense pas et c'est en grande partie cette conviction qui m'avait conduit à écrire dans *Principes des systèmes intelligents* [Jorion 1990] que la conversion est un réaménagement de la dynamique d'affect du réseau mnésique tel que ce sont d'autres traces mnésiques qui accèdent désormais au statut de noyaux de croyance,

passages contraints pour la relaxation provisoire du système (atteinte d'un puits de potentiel).

Il me semble qu'il y a à l'intérieur de nous tous quelque chose de l'ordre du préjugé, de ce qui ne se modifie qu'en tout dernier recours dans l'organisation conceptuelle (ce que j'appelle noyau de croyance), qui nous fait supposer qu'il s'agit avec la fenêtre à ouvrir et avec la conversion, de deux phénomènes d'ordre radicalement différent. Il me semble aussi que c'est cette croyance personnelle — mais culturellement partagée — qui d'une certaine manière nous interdit d'aller aussi loin que nous aimerions le faire dans la façon dont nous concevons les prochaines étapes de l'IA. Quelque chose de l'ordre d'un tabou : quelque chose qui provoque la crainte ou la colère si l'on y touche.

Lorsque l'on a terminé de lire la maigre littérature qui porte spécifiquement sur l'objet que j'appelle une "théorie de la signification", on reste sur sa faim. Qu'il s'agisse de la Philosophie, de la Psychologie ou de la Linguistique, on découvre sans doute ici ou là des voies très prometteuses, telle ce que j'ai appelé la "linguistique" d'Aristote [Jorion 1996] ou la théorie scolastique du complexe significable (voir [Jorion 1997] mais au bout du chemin, le mystère demeure égal : qu'est-ce bien que la signification ?

L'histoire des sciences nous est peut-être ici d'un certain secours. Lorsque Max Planck pose les jalons de la mécanique quantique, il nous est aisément loisible d'énumérer ses prédécesseurs : il bâtit sur les fondations posées par Clausius, Maxwell et Boltzmann. Lorsque Darwin met au point sa théorie de l'évolution des espèces ou lorsque Freud développe la méta-psychologie freudienne, on aurait au contraire bien du mal à déterminer leurs prédécesseurs (l'oeuvre parallèle de Wallace est contemporaine de celle de Darwin). On peut toutefois leur trouver ici et là dans l'histoire (et non quelques années auparavant), des précurseurs (la tâche est plus aisée pour Darwin que pour Freud), des penseurs qui exprimèrent des vues où l'on retrouve en germe, sous forme ébauchée et souvent d'idée isolée, ce qui ne prendra tout son sens que dans la théorie complète que Darwin ou Freud développèrent ensuite. Lorsque des prédécesseurs existent, comme c'est le cas pour Planck, la quête de précurseurs apparaîtrait bien vaine puisqu'une ligne continue de prédécesseurs conduirait jusqu'à eux.

Qu'est-ce qui distingue alors les découvertes de Darwin ou de Freud, sinon leur réelle nouveauté ? "Qu'elles ne constituent pas des théories à proprement parler" disent aujourd'hui certains, "du fait qu'elles ne sont pas falsifiables, qu'elles ne se prêtent pas à la contre épreuve". L'argument est sans mérite : leurs théories sont falsifiables, au même titre que le "Big bang" par exemple, même si ce cela exige davantage d'argumentation discursive que de recours à la vérification expérimentale pure. Ce qui distingue leurs théories, c'est qu'il est difficile, au sens de "dur" psychologiquement, pour un auteur de les formuler. Il existe pour ces théories un tabou à surmonter, une conversion à réaliser, d'une part pour son auteur, au moment où il la formule, d'autre part pour son lecteur au moment où il doit se laisser convaincre, au moment où certains remparts dressés par son affect doivent s'effondrer pour faire place à la nouvelle conception.

Ce qui caractérise le darwinisme ou le freudisme, c'est que s'ils sont vrais, d'une manière automatique le mérite de Darwin et de Freud, en tant qu'ils en sont les auteurs en est diminué. Si nous ne sommes que les descendants de grands singes, alors le darwinisme lui-même a pour auteur le descendant d'un grand singe (les caricaturistes de l'époque s'en sont d'ailleurs donné à coeur joie), de même, si toute oeuvre humaine est un moyen détourné de satisfaire une pulsion d'ordre sexuel, alors la méta-psychologie freudienne elle-même est un moyen détourné pour son auteur de satisfaire une pulsion d'ordre sexuel.

La théorie de l'évolution de Darwin ainsi que la psychanalyse — on l'a écrit — impliquent une dévaluation, un rabaissement de l'image que se fait la race humaine d'elle-même. La vanité de l'espèce en prend un mauvais coup, car il s'agit de bien plus que d'une théorie nouvelle, il s'agit en même temps d'une leçon d'humilité. Copernic en avait fait autant lorsqu'il déplaça la terre du centre vers la périphérie (la claque de l'héliocentrisme est cependant plus sérieuse au sein de la cosmologie chrétienne qu'au sein de l'univers païen polythéiste qu'Aristarque de Samos avait à peine ému avec une hypothèse identique) ou lorsque Linné le premier classa l'homme au rang des mammifères.

Alors qu'est-ce qui nous empêche de dire qu'ouvrir la fenêtre machinalement et subir dans son être une conversion religieuse ce sont des effets de signification du même ordre ? Probablement un mécanisme psychologique du même ordre que celui que je viens d'évoquer à propos de Darwin et de Freud : si c'est le cas en effet, alors composer la Neuvième Symphonie ou peindre La ronde de nuit, sont des réalisations, individuelles sans doute, mais ni plus ni moins liées à un sujet humain authentique qu'ouvrir la fenêtre machinalement. Quant à celui qui attacherait son nom à la découverte que de tels effets de signification diffèrent seulement par leur degré de complexité, il rabaisserait d'autant sa propre découverte : elle aurait été tout aussi machinale, selon l'automatisme qu'il aurait mis en évidence. Il serait l'auteur de sa découverte par un mécanisme dont — il l'aurait prouvé — sa personne n'est le support que pour des raisons parfaitement fortuites au regard de l'histoire. Tout ce qu'il pourrait affirmer quant au fait que la découverte ne pouvait avoir lieu que par lui se trouverait automatiquement disqualifié : ce ne pouvait être que lui sans doute mais sans pour autant que la paternité en revienne à ce "moi, je" dont il aime ponctuer son discours.

Voilà en quelques mots ce qui expliquerait pourquoi les penseurs qui se sont penchés sur la signification seraient arrêtés sur son bord — puisque la théorie à découvrir les priverait de la satisfaction de mettre en avant leur propre personne — satisfaction qui guide en temps normal le processus de la découverte.

Le raisonnement que je viens de tenir, n'est bien entendu pas probant sous la forme où je l'ai présenté : il a été produit de manière indirecte, à la lumière de l'éclairage que procure l'histoire des sciences, il conviendrait encore de le prouver ou, sinon le prouver, du moins de le soutenir par des arguments suffisamment probants tels ceux qui supportent la théorie de l'évolution darwinienne ou la méta-psychologie freudienne.

Je ne pensais à rien de tout cela quand j'ai proposé dans *Principes des systèmes intelligents* [Jorion 1990] un modèle de la pensée comme un gradient à l'intérieur d'un réseau mnésique sous-tendu par une dynamique d'affect. Le modèle permettait toutefois déjà d'éliminer l'intentionnalité, cette projection en avant, cette tension qui nous guiderait vers un but. S'il s'agit d'un simple glissement selon la plus grande pente débouchant sur la relaxation qui s'opère lorsqu'est atteint un puits de potentiel, alors, la conscience se trouve déjà privée de la fonction la plus ambitieuse que nous lui attribuons ordinairement, de se diriger de manière délibérée vers des buts qu'elle s'est préalablement assignés. Cette logique de gradient où nous, sujets humains, nous laissons capturer (comme le badaud est capturé machinalement par les vitrines de la rue où il déambule) par la pente la plus raide de l'espace de configuration que constitue le monde où nous vivons, je l'ai décrite — sous une forme épurée — dans un texte publié quelques années plus tard [Jorion 1994].

Même privée de l'intentionnalité comme l'un de ses attributs, la conscience demeure toutefois comme un donné irréductible. J'ai dit plus haut qu'il est sans conséquence d'affirmer que la conscience est illusoire : son expérience quotidienne nous la rend incontournable, nous la vivons de manière trop certaine pour nous convaincre qu'elle ne serait pas là. Mais quel pourrait être un autre type de rabaissement qui ferait que nous, humains, n'avons jamais pu saisir ce qu'il en est vraiment de la signification ? Eh bien, par exemple que si la conscience joue bien un rôle, que si son existence — aussi certaine que nous puissions la vivre — est avérée, que ce rôle soit en réalité trivial, "futile" (*trifling*) comme s'exprimait Locke.

Voici alors ce que je vais faire pour conclure : je vais prolonger la voie ouverte par le modèle en gradient, en y situant la conscience — dont je tiens l'existence pour indubitable — en tant qu'épiphénomène ne jouant qu'un rôle "futile" dans le processus global de la pensée.

Je rappelle la problématique telle que je l'ai caractérisée dans des textes déjà publiés :

"*Pourquoi la rivière va-t-elle à la mer ?*" Nullement parce que A provoque B, et que B entraîne C, jusqu'à ce que soient atteints les rivages de l'océan, mais parce qu'il existe un gradient, on dirait dans ce cas-ci, une "pente" : l'eau suit la courbe de la pente la plus raide, jusqu'à ce qu'elle atteigne le niveau de la mer. C'est un effet de la gravité : le centre de la terre joue le rôle d'un attracteur, et c'est en se dirigeant vers celui-ci que l'eau, dans un mouvement de spirale centripète (le méandre de son cours n'est rien d'autre), s'arrête un beau jour à la mer, dans l'impossibilité d'aller plus loin, de se rapprocher davantage de son attracteur.

C'est Freud qui observa le premier que sans une dynamique d'affect pour provoquer la remémoration, la mémoire ne serait qu'un réseau statique qui demeurerait sans expression extérieure. Aux mots sont attachés une valeur d'affect qui rayonne sur ceux qui leur sont connectés. Ces valeurs sont en perpétuelle variation : il faut imaginer la mémoire associée à l'affect comme la surface de l'océan par grand vent.

Dans la veille, le réseau des traces mnésiques est parcouru selon la dynamique de l'affect : l'affect

canalise, décide à certaines bifurcations de diriger le train d'impulsions vers tel branchement plutôt que vers tel autre. En fait, l'affect joue exactement le rôle d'un gradient : il décide entre deux ramifications dendritiques laquelle présente "la pente la plus raide".

Le petit ru fera la grande rivière et au moment où la petite goutte de pluie qui d'abord ruissela sur le sol rejoint la mer, tout se sera passé comme si cette dernière l'avait appelée, à partir de ce moment situé dans l'avenir où la jonction a enfin lieu. Or l'océan s'est contenté d'être là : il se fait simplement que c'est lui que l'on trouve au bout du gradient, il est ce que l'on appelle en physique, un puits de potentiel. Ces soucis qui encouragent, qui guident l'association libre vers eux de manière insistante, toujours renouvelée, j'en ai donné des exemples dans *Principes des systèmes intelligents* [Jorion 1990], "*se souvenir d'acheter du beurre*", "*le rappel d'impôt aussitôt démenti*", etc. Le souci est un puits de potentiel auquel nous ne pouvons rien faire, il appelle l'association libre vers lui.

L'intention elle-même peut être envisagée comme un "souci" qui nous advient par devers nous-mêmes. Aussitôt que l'intention est présente, dès que le projet existe, tel la réalisation d'un objet — et quelque soit la manière dont ce souci s'est imposé — la vision future du projet accompli, du but atteint, agit comme un puits de potentiel. Celui-ci ne sera sans doute effectivement atteint que dans un moment à venir, mais il guide vers lui parce qu'il existe dès l'origine, c'est-à-dire qu'il existe de manière contemporaine à l'ensemble des opérations, à la réalisation de chacun des sous-buts, qui mènent vers lui. Le but à atteindre est un puits de potentiel qui pré-existe aux étapes vers sa réalisation, que l'on peut envisager comme ses effets. L'intention (le puits) est présente avant la réalisation de l'action envisagée, seul l'accomplissement aura effectivement lieu plus tard : au moment où le puits de potentiel aura été atteint [Jorion 1994 : 94-98].

Reprenant des arguments avancés autrefois par Wittgenstein et par Merleau-Ponty, j'avais, selon la même logique, avancé dans *Principes des systèmes intelligents* que l'on n'a pas le temps matériel d'"avoir l'intention de dire" tout ce que l'on dit effectivement dans le feu de l'action d'une conversation ou d'un discours quelconque, c'est-à-dire qu'une fois lancée sur sa pente, la parole se poursuit jusqu'à extinction, relaxation, à savoir jusqu'à ce que le gradient d'affect vienne mourir dans un puits de potentiel. Dans la conversation, c'est le discours de l'autre qui, mettant mon affect en émoi, relance le processus, à savoir crée un nouveau gradient.

Or, et la chose va de soi, on ne s'arrête pas de penser quand on s'arrête de parler.

Socrate — *Appelles-tu penser ce que j'appelle penser ?*

Thééthète — *Dis moi ce que c'est.*

Socrate — *Le discours (logos) que l'âme se tient à elle-même sur ce quelle voit. Il me semble que, lorsqu'elle pense, elle ne fait rien d'autre que converser, poser des questions et proposer des réponses, affirmant et niant. Lorsqu'elle se fixe et ne met plus en doute, nous appelons cela "opinion" (doxa). Donc ce que j'appelle se*

former une opinion c'est discourir (légein) et l'opinion, c'est la parole que l'on énonce non pour quelqu'un d'autre ou tout haut (phonè) mais en silence pour soi-même.

(Platon, *Thééthète*, 189E-190A, trad. de J. Burnet, dans [Kneale & Kneale 1986 : 17-18]).

On s'entend parler quand on parle, mais on s'entend parler tout aussi bien quand on se contente de penser. Je vais plus loin : si ce que l'on dit, on n'a jamais eu " *l'intention de le dire* ", alors ce que l'on dit, on l'apprend seulement — comme quiconque — au moment où on se l'entend dire. Et si l'on s'entend dire ce que l'on dit soi-même, alors ce qui est entendu nous met en émoi au même titre que ce que l'on entend dire par autrui.

Qu'est-ce que ceci implique ? Ceci implique la chose suivante : notre discours (aussi bien intérieur qu'extérieur) au moment où il est entendu, modifie notre affect, c'est-à-dire modifie le profil du gradient d'affect qui sous-tend notre discours alors même que celui-ci est en train de se dérouler. Il y a rétroaction (feedback), effet en boucle, et comme pour tout effet en boucle — effet cybernétique — la dynamique se nourrit du léger retard qui existe entre le " *me l'entendre dire* " et " *me mettre en émoi* ".

La conscience à mon sens se situe là : dans le temps qu'il me faut pour " *m'entendre penser* ". La durée psychologique du sentiment vécu du " présent " a probablement sa source là également : dans ce léger temps, ce court délai, qu'il faut à la boucle pour se boucler.

Quelle fonction attribuer alors à la conscience ? aucune : il s'agit d'un épiphénomène résultant du temps nécessaire à la dynamique d'affect pour se remettre à jour — que l'input ait une source extérieure (les autres), ou tout aussi bien intérieure (moi-même). Que la conscience se confonde avec le sentiment vécu d'un " *libre-arbitre* ", d'un pouvoir décisionnel entre des choix multiples, c'est là que réside l'illusion. Le phénomène de la conscience, lui, est bien réel, seul son pouvoir est illusoire.

Le français n'a qu'un seul mot pour " *conscience* " là où l'anglais en a deux, *awareness*, pour " *conscience-de-soi* ", et *consciousness*, pour " *être-conscient* ". La " *conscience-de-soi* " accompagne sans doute de manière automatique la perception, la dynamique d'affect existe là d'ores et déjà mais en prise seulement sur le système moteur ; l'" *être-conscient* " est indissolublement lié au fait que chez l'être humain la dynamique d'affect anime un réseau dont certaines des traces mnésiques sont des signifiants. Il n'y a sans doute jamais que des actes réflexes, mais chez les hommes, certains de ces actes réflexes sont des phrases. Au tout début il y a l'affect, mais au commencement de ce qui distingue l'être humain des autres animaux, il y a effectivement le Verbe.

Références bibliographiques

[Barr & Feigenbaum 1981] Barr, A. et Feigenbaum, E.A., *The Handbook of Artificial Intelligence*, Vol. I, Los Altos (Cal.) : William Kaufmann, 1981

[Barr & Feigenbaum 1982] Barr, A. et Feigenbaum, E.A., *The Handbook of Artificial Intelligence*, Vol. II, Los Altos (Cal.) : William Kaufmann, 1982

[Cohen & Feigenbaum 1982] Cohen, P.R et Feigenbaum, E.A., *The Handbook of Artificial Intelligence*, Vol. III, Los Altos (Cal.) : William Kaufmann, 1982

[Jorion 1990] Jorion, P., *Principes des systèmes intelligents*, Paris : Masson, 1990

[Jorion 1994] Jorion, P., *L'intelligence artificielle : au confluent des neurosciences et de l'informatique*, *Lekton*, vol IV, N°2: 85-114, 1994

[Jorion 1996] Jorion, P., *La linguistique d'Aristote*, in V. Rialle & D. Fiset (eds.), *Penser l'esprit : Des sciences de la cognition à une philosophie cognitive*, Grenoble : Presses Universitaires de Grenoble : 261-287, 1996

[Jorion 1997] Jorion, P., *Jean Pouillon et le mystère de la chambre chinoise*, *L'Homme* 143, juil.-sept. 1997 (à paraître)

Kneale, William & Martha Kneale, *The Development of Logic*, Oxford : Clarendon Press 1986 (1962)

[McLelland & Rumelhart 1986a] McLelland, J.L. et Rumelhart, D.E., *Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Volume 2, Psychological and Biological Models*, Cambridge (Mass.) : MIT Press, 1986

[McLelland & Rumelhart 1986b] Rumelhart, D.E. & McLelland, J.L., *Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Volume 1, Foundations*, Cambridge (Mass.) : MIT Press, 1986