ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 553

SAINT THOMAS D'AQUIN CONTRE LES ROBOTS?

PISTES POUR UNE APPROCHE PHILOSOPHIQUE DE L'INTELLIGENCE ARTIFICIELLE

MATTHIEU RAFFRAY*
*Pontificia Università San Tommaso d'Aquino, Roma***

ABSTRACT

In light of the pervasive developments of new technologies, such as NBIC (Nanotechnology, biotechnology, information technology, and cognitive science), it is imperative to produce a coherent and deep reflexion on the human nature, on human intelligence and on the limit of both of them, in order to successfully respond to some technical argumentations that strive to depict humanity as a purely mechanical system. For this purpose, it is interesting to refer to the epistemology and metaphysics of Thomas Aquinas as a stable philosophical reference on Human Nature. Indeed, we find in the works of Aquinas some of the most productive elements that could form a base to our deeper understanding of, and possibly even solutions to some of the most perplexing questions raised in our times by the existence of AI.

IL est évidemment paradoxal de mettre en opposition un philosophe médiéval et une question aussi contemporaine que celle de l'Intelligence Artificielle. Néanmoins, pour qui veut se pencher sur les défis posés à l'intelligence par les avancées technologiques les plus récentes — et avant même de pouvoir en aborder les conséquences éthiques, anthropologiques ou sociales — il est indispensable d'en examiner les

* Contact: raffray@pust.it

** Facoltà di Filosofia.

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 554

fondements épistémologiques et même métaphysiques. Les progrès récents des NBIC («Nanotechnology, biotechnology, information technology, and cognitive science»), qui visent explicitement à transformer l'humain pour augmenter ses capacités, rendent en effet urgente une réflexion solide et profonde sur la nature humaine, ses limites et ses présupposés, réflexion qui soit capable d'argumenter face à la vacuité philosophique qui se manifeste le plus souvent dans les discours et les projets des technologues les plus avancés.

En ce sens, la permanence de la vérité comme l'atemporalité de la nature humaine justifient donc le recours à un penseur comme saint Thomas d'Aquin: la nature de l'intelligence ou le fonctionnement de l'esprit humain n'ont pas varié depuis les temps reculés du Moyen Âge. Ce qui était vrai au 13ème siècle est toujours vrai aujourd'hui; il faut néanmoins exploiter ces données philosophiques non pas comme de beaux reliquats du passé, mais au contraire pour les faire tourner au profit d'une compréhension inédite de problèmes résolument nouveaux, qui eux n'existaient pas il y a 800 ans. Se contenter de répéter une doctrine du passé — même aussi cohérente et aussi brillante que celle d'un Thomas d'Aquin — n'aurait donc ici, sur la question de l'IA, aucun intérêt. Notre ambition est plutôt de montrer en quelle mesure certains éléments de la pensée thomasienne ou thomiste peuvent venir éclairer de façon très perspicace les problèmes posés aujourd'hui par les «machines intelligentes».

Nous nous contenterons, ici, de suggérer quelques pistes de recherche, en concentrant notre réflexion sur un point fondamental, à la racine même de tous les problèmes liés à l'Intelligence Artificielle: comment ces deux termes (« intelligence» et «artifice») peuvent-ils être pensés ensemble? N'y a-t-il pas une contradiction, une incompatibilité essentielle, entre le fait d'être une machine, construite à l'aide de matériaux inanimés, et le fait de posséder une capacité intellectuelle au sens propre? Faut-il, en d'autres termes, n'entendre la notion d'intelligence artificielle qu'en un sens métaphorique, et reléguer le projet de robots intelligents au rang d'une utopie?

1.—Une machine peut-elle penser?

Dans son célèbre article paru dans la revue *Mind* en octobre 1950, et in-

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 555

titulé «Computing Machinery and Intelligence»,[1] Alan Turing, le père de l'informatique moderne, se propose de précisément «considérer la question 'Une machine peut-elle penser?'». Mais il faut — pour éviter, dit-il, une interprétation triviale de cette question, et donc une réponse absurde — la transformer en une autre question, exprimée en des termes «relativement non-ambigus», en précisant son sens.

Il propose alors d'abord de redéfinir l'idée de «pensée» en se représentant un «jeu d'imitation» (*Imitation Game*), classique dans les salons anglais du début du 20ème siècle. Dans ce jeu, un interrogateur (C), situé dans une pièce séparée, doit chercher à découvrir à l'aide de questions libres, posées par écrit, lequel de ses deux interlocuteurs (A) ou (B) est une femme — l'un et l'autre essayant de le convaincre ou de le tromper. Turing propose alors de transformer ce test en remplaçant l'un des interlocuteurs (A) par une machine, qui doit faire croire à l'interrogateur qu'elle est un homme: si celle-ci peut tromper l'interrogateur en moyenne aussi souvent qu'un homme, on dit alors que la machine a passé l'épreuve. Ce célèbre «test» (désormais connu sous le nom de «test de Turing»), présente l'avantage de discriminer la pensée ou l'intelligence non point sur des critères subjectifs ou théoriques discutables, mais sur des performances concrètes, calculables et comparables.

Ensuite, afin de donner un sens plus précis à la notion de «machine», Turing propose de s'en tenir à ce qu'il appelle une «machine digitale» (*digital computer*), c'est-à-dire un système de traitement binaire de l'information (à l'aide de 0 et de 1) qui se contente d'exécuter une liste d'instructions précises (lire, effacer, écrire sur un ruban théoriquement infini), lesquelles sont déterminées par des états différents de la machine.[2] Si cette réduction semble drastique, elle présente néanmoins un double avantage: d'une part on ne spécule pas sur la possibilité de l'existence d'une machine intelligente — puisque la machine de Turing ainsi présentée est trivialement réalisable; et d'autre part cette machine, mal-

1. Alan TURING, *Computing Machinery and Intelligence*, in «Mind», 59 (1950), pp. 433–460.

2. On trouve la première description de cette «Machine de Turing» dans un article de 1936 (A. TURING, *On Computable Numbers, with an Application to the Entscheidungsproblem*, in *Proceedings of the London Mathematical Society*, série 2, vol. 45, 1936, pp. 230–265), qui reprend les intuitions de la «machine analytique» de Charles Babbage (1791–1871).

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 556

gré sa simplicité, possède néanmoins un potentiel universel: selon la «Thèse de Church-Turing», une telle machine de Turing est en effet capable de simuler le comportement de n'importe quelle autre machine digitale, donc de n'importe quel ordinateur — si on lui donne une liste d'instruction convenable et un ruban de lecture/écriture suffisant[3] — et donc finalement d'effectuer n'importe quelle fonction «calculable».

A l'aide de ces deux précisions (le «test de Turing» et la «machine de Turing»), notre question initiale «Une machine peut-elle penser?» peut donc être traduite par la question suivante: «Peut-on imaginer qu'une machine digitale puisse passer avec succès le test du jeu d'imitation?». À en croire Turing, une telle machine devait voir le jour avant l'an 2000, c'est-à-dire passer avec succès le test dans au moins 70% des cas.[4] Il en concluait alors que «à la fin du siècle, l'usage des mots et l'opinion générale moyenne auront tellement changés qu'on pourra parler de machines pensantes sans s'attendre à être contredit».[5]

Indépendamment des résultats effectifs du test — chaque année depuis 1991, le Prix Loebner met des intelligences artificielles en compétition pour passer le Test de Turing, et en 2014, une IA russe dénommée «Eugene Goostman» est parvenue à convaincre 33% des juges qu'elle était humaine[6] — il est tout à fait intéressant, au point de vue philosophique,

---

3. Alan Turing, *Computing Machinery and Intelligence*, in *op. cit.*, pp. 441–442 : «This special property of digital computers, that they can mimic any discrete state machine, is described by saying that they are *universal* machines. The existence of machines with this property has the important consequence that, considerations of speed apart, it is unnecessary to design various new machines to do various computing processes. They can all be done with one digital computer, suitably programmed for each case. It will be seen that as a consequence of this all digital computers are in a sense equivalent».

4. Ivi, p. 442: «I believe that in about fifty years'time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning».

5. Ivi, p. 442: «The original question 'Can machines think ?' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted».

6. Le succès d'Eugene Goostman a été mis en cause par la communauté scientifique, en raison de l'aspect partiel du test effectué : dans le système qui le 7 juin 2014 a pré-

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 557

de noter que Turing s'en tient ici à une description externe et purement comportementale, ou «imitationnelle» de l'intelligence: est dit intelligent celui qui peut passer pour tel, ou qui est considéré comme tel par une majorité d'êtres humains. Les chatbots qui tentent de passer le test de Turing sont d'ailleurs aujourd'hui plus entraînés à tromper les juges qu'à développer de véritables performances dans le domaine de l'intelligence.

Au-delà de la validité du test, qui a été largement mise en cause, nous trouvons là une critique fondamentale qui peut être adressée à la conception de l'intelligence présupposée par Turing. Le chercheur Robert French,[7] par exemple, l'illustre de la façon suivante: imaginons un peuple qui ne connaîtrait qu'une seule espèce d'oiseaux, par exemple les mouettes. Ce peuple se poserait le problème de réaliser une machine volante et, pour savoir s'il a réussi, utiliserait le *test de la mouette*: une machine sera dite volante s'il est impossible de la distinguer d'une mouette dont le comportement est observé à l'aide d'un radar. Les avions, les hélicoptères, les montgolfières et même les autres oiseaux ne réussiront pas le *test de la mouette* et ne seront donc pas considérés comme «capables de voler» — parce que la définition de voler est réduite dans cet exemple à l'unique cas particulier que ce peuple connaît. En d'autres termes, pour French, le test de Turing est une condition suffisante d'intelligence, mais seulement d'intelligence humaine, et il reproche donc à Turing sa conception «imitationelle» de l'intelligence. On trouve un genre de critique semblable dans la célèbre réflexion faite par le scientifique néerlandais Edsger Wybe Dijkstra, l'un des pionniers de l'informatique théorique, réputé pour ses aphorismes: «Se demander si un ordinateur peut penser, c'est aussi intéressant que de se demander si un sous-marin peut nager»!

---

tendument passé le test, le sujet humain imité a l'âge de 13 ans et la langue anglaise n'est pas sa langue maternelle, ce qui permet d'excuser ses fautes.

7. Cfr. Robert FRENCH, *Subcognition and The Limits of the Turing Test*, in «Mind» 99 (1990), pp. 53-66.

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 558

2. Une matière douée d'intelligence: le cas des «machines angéliques».

Les réflexions précédentes reviennent donc à se poser une question plus précise: une machine, c'est-à-dire un ensemble d'éléments matériels organisés, même de façon extrêmement complexe et subtile, si elle imite à la perfection les activités humaines intelligentes les plus élevées, par exemple la connaissance ou le langage, si elle exprime des sentiments ou communique des raisonnements, peut-elle être alors elle-même considérée comme un être doué d'intelligence?

Or il semble justement que l'on peut trouver chez saint Thomas d'Aquin un élément de réponse à cette question, dans un lieu assez inattendu. En effet, dans son traité des anges — considéré souvent comme l'un des chef-d'œuvre de sa carrière théologique, lui ayant d'ailleurs valu le nom de «docteur angélique» — Thomas s'interroge, à la question 51 de la *Prima pars*, sur les rapports que les esprit angéliques entretiennent avec les corps. Dans certains cas, en effet, les anges assument un corps humain, avec toutes ses facultés, au moins en apparence. Et si certains prétendent que toutes ces manifestations ne sont que des visions de l'imagination, Thomas, lui, répond qu'en certains cas, par exemple les anges qui apparaissent à Abraham à Mambré (*Gen.* 18) ou Raphaël qui accompagne le jeune Tobie (*Tob.* 5), il doit s'agir de véritables corps, puisque les actions en question sont vues par tous, indépendamment de leur rôle dans le récit en question.[8] Les anges, en effet, sont de purs esprits, dépourvus de corps, mais ils sont capables d'agir sur la matière pour se constituer des instruments corporels, qui ne leur sont alors pas uni naturellement, mais qu'ils assument seulement pour accomplir certaines fonctions.[9]

Mais comment les anges font-ils donc pour assumer une réalité corporelle? D'où vient-elle et comment agissent-ils sur elle? La réponse de l'Aquinate est fort intéressante: l'intelligence angélique, dit-il en effet, est capable d'utiliser les propriétés de n'importe quelle portion de matière pour la transformer, la modeler, l'organiser, de sorte qu'elle puisse

8. Thomas d'Aquin, *ST* I, q. 51, a. 2, c.

9. Ivi, a. 1, ad 1.

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 559

constituer un corps humain. Dans notre cas, Thomas suggère alors que les anges agissent sur les propriétés de l'air situé dans un lieu, en le solidifiant, pour lui donner consistance, forme et couleur, et le modeler à l'image d'un corps humain.[10]

Le corps ainsi constitué, poursuit Thomas, s'il est véritablement matériel — comme une machine fort complexe, utilisant les propriétés les plus enfouies des atomes et des éléments chimiques de la nature — demeure néanmoins comme extérieur à l'ange qui l'a façonné. Non seulement «l'ange et le corps qu'il assume ne sont pas en rapport de matière à forme», comme une âme avec un corps, puisque l'intelligence angélique n'est pas liée par nature à cet élément corporel; mais de plus l'ange demeure pour le corps en question comme un pilote par rapport au navire qu'il dirige, pour reprendre l'image platonicienne,[11] c'est-à-dire comme une intelligence séparée, agissant de l'extérieur sur une matière donnée mais sans former avec elle l'unité nécessaire pour constituer un être unique.[12] Les qualités intellectuelles de ces corps assumés (comme les bons conseils donnés par l'archange Raphaël à Tobie, ou l'amitié qui naît entre eux) ne peuvent donc en aucun cas, pour l'Aquinate, être attribuées aux corps eux-mêmes. La raison fondamentale en est, explique-t-il, que «l'acte d'intellection [ne peut être] l'acte

10. Ivi, a. 2, ad 3 : «À son degré ordinaire de dilatation, l'air ne retient ni la figure ni la couleur; mais quand il est condensé, il peut revêtir différentes formes et réfléchir des couleurs: on le voit dans les nuages. C'est donc à partir de l'air que les anges forment des corps, avec l'assistance divine, en le solidifiant par la condensation autant qu'il est nécessaire».

11. Cfr. IDEM, *Contra Gentes* II, cap. 57: «D'après Platon et son école, l'âme intellectuelle n'est pas unie au corps comme une forme à sa matière, mais seulement comme un moteur à son mobile: *l'âme serait dans le corps comme un marin sur un navire*. De la sorte, l'union de l'âme et du corps se ramènerait à [un] contact virtuel [...]. Position difficile à tenir! Le contact en question ne saurait, en effet, nous l'avons vu, procurer l'unité pure et simple. Mais l'union de l'âme et du corps constitue l'homme. Il faudrait donc admettre que l'homme n'est pas un, purement et simplement: c'est-à-dire qu'il n'est pas un être pur et simple, mais un composé accidentel».

12. *ST* I, q. 51, a. 2, ad 2: «L'ange est pour le corps comme un moteur que ce corps mobile ne fait que *représenter*. [...] Les anges se façonnent, par la puissance divine, des corps sensibles qui *représentent* leurs propriétés intelligibles. C'est ce qu'on veut exprimer lorsqu'on dit que les anges assument des corps».

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 560

ni d'un corps ni d'une faculté corporelle», et il renvoie à son traité sur l'intellect humain (question 79): pour notre part, nous y reviendrons aussi un peu plus loin.

Dans ce traité sur les «machines angéliques », il nous reste à souligner un élément que Thomas développe en conclusion, et qui peut éclairer encore notre problème: la question est de savoir, en effet, si les activités déployées par les anges à travers ces corps qu'ils se sont formés sont ou non des activités «vitales», c'est-à-dire des activités qui manifestent la vie dans les corps qui les accomplissent: respiration ou digestion, connaissance ou appétit sensible, connaissance intellectuelle et volonté. En d'autres termes, peut-on dire que ces machines corporelles vivent, qu'elles parlent, qu'elles comprennent, qu'elles pensent? Ou encore: ces corps assumés par les anges de l'extérieur, si parfaits en ce qu'ils imitent le corps humain, possèdent-ils eux-mêmes les mêmes capacités sensibles et intellectuelles que les hommes qu'ils imitent? La réponse de Thomas est ici négative: si l'imitation est parfaite, elle n'est qu'une imitation. Les machines ne réalisent pas vraiment ces activités vitales. Il appuie son raisonnement sur la définition aristotélicienne de la vie, à savoir un principe immanent de mouvement — or dans notre cas, les machines angéliques n'ont qu'un principe de mouvement externe ou transcendant, à savoir l'intelligence angélique qui les a conçues.[13]

Au-delà de l'expérimentation théologique que constitue ce passage angélologique, il nous semble qu'on a trouvé ici plusieurs éléments fort instructifs pour notre problème: premièrement (il n'est pas inutile de le souligner), Thomas ne nie pas la possibilité d'intelligences non-humaines, bien au contraire; mais ces intelligences sont toujours nécessairement transcendantes à la matière, même la plus subtile et la plus complexe qui soit. Deuxièmement, les mécaniques corporelles façonnées par les esprits angéliques ne sont jamais elles-mêmes porteuses de la pensée (même si cette solution aurait été plus simple dans le schéma théologique auquel il répond), précisément parce que pour être façon-

13. Ivi, a. 3, c.: «Les anges peuvent donc, par les corps qu'ils assument, exercer les activités des êtres vivants en ce qu'elles ont de commun avec les activités des non vivants, mais non dans ce qu'elles ont de propre. Car, selon Aristote, seul peut produire une action celui qui en a la puissance. Aucun être ne peut donc avoir d'activité vitale s'il n'a pas la vie, qui est le principe potentiel d'une telle action».

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 561

nées, elles ont besoin d'un principe supérieur qui leur communique leur nature. Or pour être vivantes, ces «créatures» devraient être engendrées par des vivants de même nature — ce qui ne saurait évidemment pas être le cas de la pure matière dont il est question ici, inerte par elle-même. Le fait même que ces automates angéliques soient construits à partir de la matière les empêche donc de posséder par eux-mêmes un principe de vie qui en ferait des êtres autonomes, des machines pensantes.

3. L'INTENTIONALITÉ DE LA CONNAISSANCE ET LE PROBLÈME DE LA CHAMBRE CHINOISE

Revenons à l'article d'Alan Turing dont il était question plus haut. Nous avons suggéré qu'imiter l'intelligence n'était pas suffisant pour être au sens propre doué d'intelligence. Mais qu'en est-il des autres «activités vitales» attribuables aux ordinateurs les plus performants? Il vaut la peine de s'arrêter sur la question de la connaissance — en particulier de la connaissance sensible. En effet, une voiture automatique «Google Car» est dotée d'un lidar rotatif (une sorte de radar à laser permettant de générer une carte tri-dimensionnelle de l'entourage de la voiture), d'une caméra avant qui identifie les autres voitures, les cyclistes ou les piétons, de quatre petits radars qui détectent les obstacles proches à l'avant, à l'arrière et sur les côtés, ainsi que d'un senseur qui mesure les mouvements d'une roue motrice pour localiser avec précision le véhicule. Peut-on pour autant en conclure que la voiture «connaît» son environnement, qu'elle «sait» qu'un obstacle traverse devant elle, ou qu'elle «voit» le feu passer au rouge? En d'autres termes, et pour simplifier radicalement la question, quelle est la différence entre le capteur CCD d'une caméra et un œil, ou encore entre un thermomètre bêtement attaché à un mur et ma sensation du chaud et du froid? Est-ce parce qu'une machine est capable de mesurer, reproduire et transmettre une donnée sensible qu'elle peut être assimilée à un organe de connaissance sensible?

La question est loin d'être simple, et elle a occupé, depuis l'Antiquité, de nombreux traités d'épistémologie, autour de l'un des problèmes fondamentaux de la philosophie: qu'est-ce que connaître? Aristote s'oppo-

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 562

sait sur ce thème à la doctrine matérialiste des atomistes Leucippe et Démocrite, selon lesquels l'âme devait être formée en quelque façon d'éléments matériels, les atomes, afin de pouvoir être atteinte et altérée par ces mêmes particules élémentaires, subtiles mais néanmoins matérielles, lorsqu'elles émanaient dans l'air à partir des corps.[14] Mais Aristote répond à cette vue purement matérialiste au moyen de sa doctrine de la puissance et de l'acte: en effet, si les facultés sensibles étaient de la même matière que les objets sensibles, alors les sens seraient continuellement actifs, et ils pourraient même se sentir eux-mêmes — ce qui n'est pas le cas. Au contraire, il faut les concevoir comme des puissances, qui ne sont actualisées que lorsqu'elles reçoivent leur objet sensible qui les informe et les actualise: «la faculté sensitive, dit Aristote, n'existe pas en acte, mais en puissance seulement».[15]

Saint Thomas, lorsqu'il commente ce passage du *De Anima*, va plus loin: il cherche à déterminer la nature de cette actualisation du sens par la *species* sensible. De quelle façon ce qui est reçu dans le sens (un rayon lumineux, une onde sonore) actualise-t-il l'organe du sens pour donner lieu à une sensation? C'est là qu'il propose une distinction fondamentale: pour qu'il y ait sensation, il faut non seulement que l'organe soit affecté

14. Cfr. THOMAS D'AQUIN, *In De An.* I, cap. 2, lect. 5: «[Démocrite] voulait en effet [...] que toutes choses soient composées d'atomes. Or bien que, d'après lui, des atomes de la sorte étaient le principe de toutes les choses, il voulait néanmoins que les atomes de figure ronde soient de la nature du feu; aussi disait-il que l'âme est composée des atomes de figure sphérique. Il soutenait qu'en tant que ce sont les premiers principes, ils détiennent la faculté de connaître, et qu'en tant que ronds, ils détiennent la faculté de mouvoir. C'est pourquoi, dans la mesure où l'âme était composée de corpuscules ronds indivisibles, ils affirmait qu'elle connaît et meut toutes choses».

15. Cfr. ARISTOTE, *De Anima*, chap. 5, 416b 32–417a 7: «Parlons, en général, de toute sensation. La sensation résulte d'un mouvement subi et d'une passion, ainsi que nous l'avons remarqué car, dans l'opinion courante, elle est une sorte d'altération. Certains philosophes disent aussi que le semblable pâtit sous l'action du semblabl ; en quel sens cela est possible ou impossible, c'est ce que nous avons expliqué dans notre discussion générale de l'action et de la passion. Mais voici une difficulté: pourquoi, des organes sensoriels eux-mêmes n'y a-t-il pas sensation, et pourquoi, sans les sensibles extérieurs, les sens ne produisent-ils pas de sensation, alors qu'ils contiennent pourtant le feu, la terre et les autres éléments, lesquels sont objets de sensation soit en eux-mêmes, soit dans leurs accidents? C'est donc évidemment que la faculté sensitive n'existe pas en acte, mais en puissance seulement».

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 563

dans son être physique par une chose extérieure (le tympan qui vibre sous l'effet de l'ondulation de l'air, les capteurs rétiniens qui sont activés par les photons reçus à travers la pupille…) — ce qu'il appelle *immutatio naturalis* — mais il faut de plus que le sens soit affecté de façon «intentionnelle», c'est à dire en recevant la qualité sensible non pas comme telle, mais en tant qu'elle se rapporte à l'objet senti: c'est cette *immutatio spiritualis* qui fait que mon œil voie la fleur, et non pas les rayons lumineux venant de la fleur, ou que mon ouïe perçoive la cloche qui sonne, et non pas l'air agité près de mon oreille.[16]

Dans la *Somme de Théologie*, il justifie l'importance de cette distinction pour faire la part entre une sensation et une pure réception matérielle et physique: «Pour l'action du sens, une modification spirituelle est requise selon laquelle la forme *intentionnelle* de l'objet sensible est produite dans l'organe du sens. Autrement, si la seule modification physique suffisait à produire la sensation, tous les corps physiques en éprouveraient lorsqu'ils subissent un changement qualitatif».[17] En d'autres termes, ce qui distingue une caméra ou un microphone du sens de la vue ou du sens de l'ouïe, c'est précisément cette intentionalité de la perception, qui s'ajoute à la stimulation de l'organe et renvoie le sens, de façon active, vers l'objet extérieur, qui devient ainsi objet connu.

---

16. THOMAS D'AQUIN, *In De An.* II, cap. 7, lect. 14: «Je parle d'affection physique (*immutatio naturalis*) dans la mesure où la qualité est reçue en son être physique chez celui qui en est affecté, comme quand on se refroidit ou se réchauffe ou se déplace. Par contre, une affection spirituelle (*immutatio spiritualis*) consiste à ce que l'espèce soit reçue dans l'organe du sens ou dans le milieu sous forme intentionnelle, et non sous forme physique. En effet, l'espèce sensible ne se reçoit pas dans le sens selon l'être qu'elle a dans la chose sensible». Cfr. aussi *In IV Sent.*, dist. 44, q. 2, a. 1, ad qla. 3: «*Sciendum, quod organa sentiendi immutantur a rebus quae sunt extra animam, dupliciter. Uno modo immutatione naturali, quando scilicet organum disponitur eadem qualitate naturali qua disponitur res extra animam quae agit in ipsum; sicut cum manus fit calida et adusta ex tactu rei calidae, vel odorifera ex tactu rei odoriferae. Alio modo immutatione spirituali, quando recipitur qualitas sensibilis in instrumento secundum esse spirituale, idest species sive intentio qualitatis, et non ipsa qualitas; sicut pupilla recipit speciem albedinis, et tamen ipsa non efficitur alba. Prima ergo receptio non causat sensum per se loquendo, quia sensus est susceptivus specierum in materia praeter materiam, idest praeter esse materiale quod habebant extra animam, ut dicitur in 2 de anima*».

17. IDEM, *ST* I, q. 78, a. 3, c.

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 564

Cette notion médiévale d'intentionalité, telle qu'on la trouve exposée chez Thomas, a d'ailleurs été récemment remise à jour par un philosophe américain majeur dans le domaine de la philosophie des états mentaux: John Rogers Searle. S'inscrivant dans la lignée des travaux de Brentano, à la fin du 19ème siècle, qui y voyait une caractéristique des «phénomènes psychiques»,[18] puis des recherches de Husserl ou de Heidegger développant la notion de «directionalité intentionnelle»,[19] Searle a développé une théorie complète de l'intentionalité en général, qu'il définit comme «la propriété en vertu de laquelle toutes sortes d'états ou d'évènements mentaux renvoient à ou concernent ou portent sur des objets et des états de choses du monde».[20] C'est donc la caractéristique de certains états mentaux (langage, sensation, pensée, volonté), qui possèdent une capacité de renvoyer le sujet conscient à un objet du monde situé en-dehors de lui-même. Ce trait distingue donc fondamentalement l'esprit (esprit — *mind* en anglais — est entendu ici au sens le plus large, et incluant donc la connaissance sensible chez les animaux par exemple) des objets physiques, dont l'identité n'inclut pas autre chose qu'eux-mêmes. Dans le cas de la sensation, en particulier, cette notion d'intentionalité est fondamentale: «Quand je vois une voiture ou autre chose, écrit Searle, j'ai une expérience visuelle déterminée. Dans la perception visuelle de la voiture, je ne *vois* pas l'expérience visuelle: ce que je vois, c'est la voiture».[21] Afin de prouver cette propriété particulière et non-réductible de l'esprit humain, Searle résonne par l'absurde: la nier reviendrait en effet à affirmer que le langage (et toute forme d'expression) est toujours incapable d'entrer en relation avec le réel, et donc finalement d'exprimer un quelconque état de chose externe

18. Cfr. Franz BRENTANO, *Psychologie du point de vue empirique*, [¹1874; ²1924] tr. fr. M. de Gandillac, Vrin, Paris, 2008.

19. Cfr. Edmond HUSSERL, *Recherches logiques*, V, §13, tr. fr. Hubert Élie, Lothar Kelkel & René Schérer PUF, Paris, t. II, vol. 2, 1962. Cfr. par exemple P. MCCORMICK, *Sur le développement du concept de l'intentionnalité chez Brentano et Husserl*, in «Philosophiques» 8 (1981), pp. 227–237.

20. John R. SEARLE, *Intentionalité. Essai de philosophie des états mentaux*, tr. fr. C. Pichevin, Ed. de minuit, Paris, 1985, p. 15.

21. Ivi, p. 57.

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 565

au sujet. En ce sens, l'intentionalité est une condition nécessaire de la compréhension.

C'est précisément ce que Searle a manifesté, par ailleurs, dans le célèbre exemple de la «chambre chinoise»:[22] dans un article de 1980, il s'oppose frontalement à la doctrine de Turing sur les machines pensantes, à l'aide d'une expérience de pensée ayant fait date dans la philosophie de l'IA. Dans cette expérience de pensée, Searle imagine qu'il est enfermé dans une pièce — n'ayant lui-même aucune connaissance du chinois. On lui fournit alors une histoire écrite en chinois, puis un catalogue de règles, écrites en anglais, dans lesquelles sont associées des questions en chinois avec des réponses en chinois. Ces règles sont parfaitement claires pour lui, basées uniquement sur la syntaxe des phrases, sans pourtant qu'il ne comprenne rien ni aux questions ni aux réponses. Il reçoit alors des questions écrites en chinois de la part d'un interlocuteur sinophone, situé à l'extérieur de la chambre, et, en appliquant méthodiquement les règles qu'il a à sa disposition, il répond à son interlocuteur en lui renvoyant des expressions chinoises, que l'autre perçoit comme de véritables réponses à ses questions. L'interlocuteur chinois conclut donc de cette expérience que celui qui est enfermé dans la chambre parle parfaitement le chinois — puisqu'il est capable de donner une réponse à toutes ses questions — alors que ce n'est absolument pas le cas. Searle en conclut qu'il ne suffit pas d'être capable de reproduire exactement des comportements linguistiques pour «parler chinois», car «parler chinois» implique non seulement un usage maîtrisé du langage, mais aussi une conscience «intentionnelle» du sens de ce que l'on dit.

Cette expérience de pensée, transposée au cas des programmes informatiques, démontre donc l'insuffisance du test de Turing: une machine qui saurait imiter parfaitement l'intelligence humaine, simplement par l'application systématique de règles, même extrêmement précises, et éventuellement même de façon plus performantes qu'un être humain, ne saurait être dite «intelligente»: ce qui lui manque, précisément, c'est la compréhension du sens des questions et des réponses, une conscience

22. IDEM, *Minds, brains and programs*, in «The behavioral and brain sciences» 3 (1980), pp. 417–457.

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 566

«intentionnelle» des objets du monde qui l'entoure. Searle montre ainsi de façon évidente l'insuffisance de tous les modèles fonctionnalistes ou computationnalistes de la psychologie, qui réduisent l'esprit humain à une fonction ou à un algorithme du cerveau. On peut en effet imaginer un système automatique, purement mécanique, reproduisant artificiellement des comportements conscients: il serait pourtant faux de déduire qu'un tel système possède un esprit, au sens d'une conscience intentionnelle.

En d'autres termes, si la «Google car» est capable d'identifier un passage piéton ou une bicyclette, peut-être avec une plus grande acuité et une plus grande probabilité qu'un œil humain, on ne peut pas pour autant dire qu'elle les «connaît», ni même qu'elle les «voit»: lui fait défaut ce rapport intentionnel de la conscience à l'objet. Un indice, d'ailleurs, de cette différence essentielle est l'accumulation de données dont l'IA a besoin pour s'orienter dans un environnement — lorsqu'un simple coup d'œil me fait prendre conscience, sans analyse, que le feu est passé au rouge, la Google car a emmagasiné des Gigabits de données, a représenté en trois dimensions l'ensemble de son environnement, et à mis en œuvre une multitude de processus d'analyse et de fonctions…

4. Une machine peut-elle apprendre? Abstraction et accumulation extensive

Cette question de l'accumulation des données nous amène à réfléchir enfin — de façon seulement esquissée — sur la nature de la connaissance intellectuelle, et son applicabilité au cas d'une «machine pensante». En effet, l'une des intuitions les plus remarquables de l'article de Turing publié en 1950 — et ce qui a fait de lui un jalon fondamental dans l'histoire de l'IA — est l'exposition, à l'issue de son étude, d'un projet d'apprentissage automatique, qui a donné lieu à ce qu'on appelle aujourd'hui le «*machine learning*», et qui regroupe désormais la majeure partie de l'activité de recherche en IA: apprentissage supervisé et non-supervisé, apprentissage profond (*deep learning*), apprentissage par renforcement (*Q-learning*), etc.

Le raisonnement de Turing, ici, est simple: puisqu'il est difficile d'imiter un esprit humain adulte, en raison de la complexité du cerveau et

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 567

de ses nombreux aspects encore inexplorés et même inaccessibles, pourquoi ne pas plutôt essayer de produire une machine simulant le cerveau d'un enfant, et la soumettre à un processus d'apprentissage semblable à celui duquel résulte un cerveau adulte?[23] L'avantage de cette approche est double: non seulement le cerveau d'un enfant (doté de rares mécanismes et de nombreuses cases blanches) sera bien plus facile à imiter que celui d'un adulte; mais de plus, en le soumettant à des processus d'éducation et d'apprentissage, il ne sera pas nécessaire que l'enseignant sache précisément ce qui advient dans le cerveau de son enfant-machine. C'est là, sans doute, le point décisif du succès de la méthode suggérée par Turing: il ne s'agit plus d'imposer à la machine un nombre élevé de règles de calcul et d'algorithmes, dont le déroulement est maîtrisé à chaque étape du processus, en vue de lui faire effectuer une opération prévue à l'avance; il s'agit ici de penser l'apprentissage comme une collection d'expériences, pourvues de récompenses et de punitions, au terme desquelles la machine aura par elle-même «appris» comment atteindre le bon résultat au moindre coût, sans que l'instructeur puisse clairement déterminer quels sont les critères qui ont été pris en compte principalement ou accessoirement par la machine.[24]

---

23. A. TURING, *op. cit.*, pp. 455–456: «Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child».

24. Ivi, pp. 458–459: «An important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside, although he may still be. able to some extent to predict his pupil's behaviour. This should apply most strongly to the later education of a machine arising from a child-machine of well-tried design (or programme). This is in clear contrast with normal procedure when using a machine to do computations: one's object is then to have a clear mental picture of the state of the machine at each moment in the computation. This object can only be achieved with a struggle. The view that the machine can only do what we know how to order it to do, appears strange in face of this. Most of the programmes which we can put into the machine will result in its doing something that we cannot make sense of

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 568

Prenons un exemple: les processus d'apprentissage profond sont utilisés le plus souvent aujourd'hui pour la reconnaissance de sons ou d'images (reconnaissance faciale, reconnaissance vocale, vision par ordinateur, traitement automatisé du langage, recommandations d'achat sur internet…). A l'aide d'un empilement de couches de neurones électroniques (imitant les neurones biologiques), dont chacune prend ses entrées sur les sorties de la précédente, assorties d'opérateurs linaires simples et de fonctions d'activation bien choisies, un réseau de neurones artificiel est en un premier temps entraîné sur un grand nombre d'exemples, en faisant varier (selon des méthodes statistiques assez basiques) les poids neuronaux, jusqu'à obtenir des résultats satisfaisants. Par exemple, jusqu'à obtenir l'identification par la machine de toutes (ou presque toutes) les images de chat dans une base de données. Cette phase dite «d'apprentissage» ou «d'entraînement» est alors suivie d'une seconde phase, de «mise en production»: le modèle étant déterminé, de nouvelles images peuvent alors être soumises au réseau de neurones déjà entraîné, afin d'obtenir le résultat correspondant à la tâche souhaitée (reconnaître une image de chat parmi une collection de nouvelles images). On dira alors, en un certain sens, que l'IA «a appris» ce qu'est un chat, qu'elle en a saisi la nature, puisqu'elle est capable désormais d'en reconnaître les différentes représentations, quelles que soient la couleur, la position, la dimension ou l'attitude du chat représenté. Qu'en est-il? Peut-on véritablement parler ici d'apprentissage?

Au point de vue technique d'abord, si l'on explore le contenu de cette «connaissance», on découvre qu'elle n'est en fait qu'une «moyenne» (au sens d'un calcul statistique) des innombrables expériences accumulées en peu de temps par l'IA. C'est d'ailleurs précisément l'accès à des bases de données de plus en plus étendues, et à des temps de calcul de plus en plus courts qui ont permis le développement de ces techniques d'apprentissage. Ce qui est appris par la machine, sa «connaissance», dans ce cas, n'est donc au final que le calcul d'une somme pondérée: une formule composée uniquement d'éléments simples, mais extrêmement complexe en elle-même. Le plus souvent, d'ailleurs, les règles de calcul retenues par la machine sont totalement incompréhensibles (et

---

at all, or which we regard as completely random behaviour».

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 569

même inaccessibles) pour le scientifique qui a lancé le processus: dans la recherche en IA, on appelle désormais le «black box problem» l'étude de ces processus cachés mis en œuvre par un système pour parvenir à des résultats satisfaisants.

Au point de vue philosophique, ce qui est notable, c'est que la reconnaissance d'un objet — son identification, donc l'accès à sa nature — ne se fait pas en termes de «compréhension », mais seulement en termes «d'extension», pour reprendre les termes de la logique scolastique. On se souvient en effet qu'un concept est dit posséder un «contenu objectif», lequel est formé des perfections au moyen desquelles il se distingue des autres : on appelle cela les «notes» de ce concept. Par exemple, le concept d'homme exprime les notes: corporel, vivant, doué de sensibilité et de raison. La *compréhension* d'un concept est alors *l'ensemble des notes qui en constitue le contenu objectif*; tandis que *l'extension* du concept est *l'ensemble des sujets auxquels convient le contenu objectif de cette idée.*[25] Prenons un exemple: on peut définir en compréhension le concept d' «étudiant à l'Angelicum» par l'ensemble des notes: «humain, cherchant à acquérir des connaissances en sciences religieuses, inscrit auprès du secrétariat de l'Université». Mais on peut aussi définir le même concept en faisant une énumération de tous les noms qui apparaissent sur la liste des inscrit. Les deux définitions désignent le même concept, et permettent de reconnaître infailliblement un «étudiant à l'Angelicum». Dans notre cas, une IA programmée pour savoir à l'avance si un jeune homme est un étudiant à l'Angelicum serait «éduquée» en assimilant tous les profils des étudiants actuels. Et l'IA calculerait alors sa propre formule, sur cette base extensive, afin de reconnaître son objet: il est alors probable que le point commun moyen entre tous les étudiants de l'Angelicum — avec un risque faible d'erreur — sera celui d'être passé à au moins 4 reprises durant les 24 dernières heures devant le guéridon d'entrée et d'avoir salué la jolie fille qui se trouve à l'accueil. S'il s'agit d'un critère d'identification efficace, on comprend néanmoins qu'il ne s'agit pas là d'une définition de l'étudiant.

25. Cfr. F.-J. THONNARD, *Précis de philosophie, en harmonie avec les Sciences modernes*, Desclée de Brouwer, Paris, 1950, I. Logique, § 27.

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 570

Nous sommes donc ici confrontés au problème épistémologique fondamental qu'est la question de l'accès à l'universel: pour Aristote, la différence essentielle entre connaissance sensible et connaissance intellectuelle se trouve dans la nature de l'objet connu: «la sensation en acte ne s'applique qu'aux choses particulières, tandis que la science s'applique aux choses universelles».[26] En commentant ce passage, Thomas justifie métaphysiquement cette division. Toute connaissance, en effet, est une certaine ressemblance de la chose connue dans la puissance de connaître: le sens, parce qu'il est lié à un organe matériel, ne peut donc recevoir que ce qui est matériel dans la chose, c'est-à-dire ce qui est individuel et déterminé, puisque la matière est ce qui individualise la forme en la contenant sous des dimensions délimitées; l'intelligence, elle, parce qu'elle est immatérielle, peut recevoir ce qui est universel dans la chose, en abstrayant la forme de ses conditions matérielles individuantes: «il est donc manifeste, conclut Thomas, que la similitude de la chose reçue dans le sens représente la chose en ce qu'elle est singulière, tandis que celle reçue dans l'intelligence la représente sous la conception de sa nature universelle».[27]

En d'autres termes, une machine «intelligente », parce qu'elle est toujours purement et intégralement matérielle, n'est jamais en mesure d'accéder à un contenu abstrait en tant que tel, et donc à un contenu universel. Si une IA a accès au concept, c'est donc toujours exclusivement par accumulation de connaissances individuelles, donc au moyen de l'extension du concept, jamais par l'entremise de sa compréhension, qui porte, elle, sur des universaux. Une IA peut donc être plus performante qu'une intelligence humaine, en tant qu'elle a accès a davantage de données; mais elle ne peut pas, en raison de sa matérialité, fonctionner de la même façon que l'intelligence humaine: cette dernière, dépourvue de ressources exceptionnelles en extension, possède néanmoins cette qualité spirituelle lui donnant, par abstraction, un accès immédiat à la compréhension des concepts.

26. ARISTOTE, *De Anima*, II, 5, 417b 20.

27. THOMAS D'AQUIN, *In De An.* II, 5, lect. 12, § 377.

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 571

CONCLUSION: IA FORTE ET IA FAIBLE

En guise de brève conclusion, il semble donc essentiel de faire nôtre une distinction fondamentale et désormais classique en philosophie de l'IA, et que John Searle proposait dès l'introduction de son article cité plus haut de 1980, la distinction entre IA faible et IA forte: la première regroupant tout processus visant à *simuler* l'intelligence, ou toute machine visant à agir *comme si* elle était intelligente; la seconde, au contraire, désigne des machines qui seraient douées au sens propre d'intelligence, de capacité de raisonnement, donc de liberté, capables d'avoir une réelle conscience d'elles-mêmes, et éventuellement aussi d'éprouver des sentiments.[28]

Au sens de l'IA faible, des «machines intelligentes», où «intelligence» est pris dans un sens analogique, ont donc largement leur place dans le cadre d'une philosophie de l'activité humaine, telle qu'on la trouve chez saint Thomas d'Aquin: en tant qu'elles imitent et même dépassent certaines capacités du cerveau humain, de telles machines jouent le rôle d'outils, qui prolongent et amplifient les capacités humaines, au même titre que tous les autres instruments que l'intelligence humaine a sans cesse mis au point. Les robots autonomes les plus performants sont donc réalisables, avec des perspectives extraordinaires — et aussi avec les dangers inhérents à de telles inventions, comme pour tout instrument mis au point par l'homme. Mais si l'on prétend, au sens de l'IA forte, donner vie à une machine qui possèderait en tant que telle des capacités d'ordre intellectuel et spirituel, une faculté de compréhension et d'abstraction, et donc finalement une autonomie et une liberté caractéristiques des êtres spirituels, la science dépasse là ses limites, et même sa propre cohérence.

Finalement, l'étude de l'intelligence artificielle permettra, il me sem-

28. John SEARLE, *Minds, brains and programs*, *art. cit.*, p. 417: «Selon l'IA faible (*weak IA*), la valeur principale d'un ordinateur dans l'étude de l'esprit (*mind*) est qu'il nous fournit un outil très puissant. […] Mais selon l'IA forte (*strong AI*), l'ordinateur n'est plus seulement un instrument dans l'étude de l'esprit; au contraire, un ordinateur convenablement programmé *est* réellement un esprit, dans le sens où des ordinateurs à qui l'on donnerait les bons programmes pourraient être dit littéralement *comprendre* et posséder des états mentaux».

ANG 96-4 (2019).qxp_ANGELICUM 13/01/20 14:21 Pagina 572

ble, d'approfondir la compréhension de notre propre psychisme, et de mettre en évidence plusieurs de ses caractéristiques essentielles, trop souvent négligées ou rejetées par les systèmes philosophiques actuels: en premier lieu la spiritualité de l'esprit humain, contre les matérialismes contemporains, tout aussi bien que la profonde unité de l'âme avec le corps, contre le dualisme cartésien. Un retour à la sagesse philosophique médiévale est donc peut-être, en ce sens, un allié nécessaire pour qui veut affronter avec sérénité la guerre à venir de l'homme contre les robots…