Ce qui échappe à l'intelligence artificielle

Sous la direction de
François Levin et Étienne Ollion

hermann
Depuis 1876

III

L'invention de l'humain. Pour en finir avec l'opposition homme-machine

par Marcello Vitali-Rosati

Performer l'humain : *L'imitation game*

Qu'est-ce que l'être humain ? Qu'est-ce que la machine ? Est-ce qu'il y a quelque chose qui « échappe à la machine » et qui donc – c'est ce qu'on semble sous-entendre – n'échappe pas à l'humain ? Est-ce qu'il y a quelque chose qui permet de caractériser l'un et l'autre, et de les différencier clairement ?

La thèse que ce texte essayera de démontrer est que ces questions sont mal posées ou, mieux, que leur fonction n'est pas de permettre le repérage de différences entre deux entités préexistantes – d'une part l'humain et, de l'autre, la machine – mais plutôt de produire par différenciation les définitions mêmes de ces deux pôles. En d'autres mots, il n'y a pas quelque chose comme l'humain et quelque chose comme la machine, mais en s'interrogeant sur cette relation, nous produisons ces deux « objets ». Nous allons montrer qu'avec une question comme « Qu'est-ce qui échappe à l'intelligence artificielle ? » nous essayons en réalité non pas de comprendre les limites éventuelles des machines, mais de définir ce qu'est l'humain.

Pour démontrer cette thèse, il est nécessaire de revenir à une des formulations les plus connues de nos questions, celle développée dans

le fameux article de Turing de 1950[1], où le mathématicien a imaginé le test qui permettrait d'évaluer l'« intelligence » des machines et de répondre à la question : « Les machines peuvent-elles penser ? » (« Can machines think ? »)

Il est d'abord important de rappeler qu'il ne s'agit pas là d'un texte technique, mais plutôt d'un *Gedankenexperiment*, d'une expérience de pensée, qui fait appel à notre façon de concevoir l'intelligence dans nos pratiques quotidiennes. Turing ne donne pas de définition formelle, il ne commence pas par établir ce qu'est l'intelligence, ce qu'est l'être humain et ce qu'est la machine : il imagine un jeu. L'approche est très intéressante, car la question ontologique (qu'est-ce que cela ?) est posée à partir d'une pratique, d'une action, d'une performance. On pourrait dire – et c'est ce que ce texte entend démontrer – que l'essence dérive des interactions, ou plutôt, comme nous le verrons, des « intra-actions ».

Turing démarre donc en affirmant qu'il serait inutile de tenter de donner une définition formelle de « pensée » et il propose de passer par un jeu : le jeu de l'imitation (*imitation game*). Il s'agit d'un jeu de société dans lequel chaque participant doit jouer un rôle. Il n'est donc pas tant question d'être mais de performer.

Le jeu dans sa version originale mérite une attention particulière : les personnes qui jouent doivent performer un genre, féminin ou masculin. Il y a donc un homme (A), une femme (B) et un interrogateur (C). A essaie de faire croire à C qu'il est une femme, il essaie donc de le tromper. B essaie d'aider C, elle dit donc la vérité, elle joue la femme. C doit essayer de deviner le genre de A et de B. La question du genre est fondamentale ici, car elle fait immédiatement comprendre les enjeux éthiques et politiques qui se cachent derrière le jeu : il s'agit de définir ce qu'est une femme et ce qu'est un homme en jouant ces rôles. Dans les questions-réponses de A, de B et de C, il est question d'établir quelles sont les caractéristiques d'un homme et celles d'une femme en jouant avec les *a priori* des autres. C'est une sorte de négociation d'une définition, une négociation qui devra faire les comptes avec les préjugés sociaux existants et les convictions de chacun.

Le premier exemple de Turing pour illustrer le jeu est parlant : C questionne sur la longueur des cheveux en présupposant, on imagine,

1. Alan Turing, « Computing Machinery and Intelligence ». *Mind*, 59 (236), 1950, p. 433-60.

qu'une femme doit avoir les cheveux longs et un homme, courts. Peu importe donc si B a les cheveux longs ou pas, si elle porte une jupe ou pas, et si A a les cheveux courts et qu'il porte des pantalons. Ce qui compte est comment les deux jouent. Et les deux jouent la femme : A pour tromper – car il est un homme – et B pour « dire la vérité ». B doit en réalité essayer d'être ce que C pense être une femme, et donc la « vérité » qu'elle dira n'est pas une description fidèle de sa façon d'être une femme, mais une liste des caractéristiques dont elle pense que C pense qu'elles sont féminines.

L'enjeu est évident : le point n'est pas tellement de partir de deux objets bien définis, avec leur essence établie et stable, pour les identifier ; il s'agit plutôt de jouer à produire ces deux essences. La définition de « femme » n'est pas le point de départ du jeu, mais plutôt son point d'arrivée : une femme sera l'ensemble des caractéristiques qui sont arrivées à convaincre C que la personne qui les a est une femme.

C'est cette structure qui est ensuite appliquée à la question « Les machines peuvent-elles penser ? » Malgré une certaine ambiguïté dans le texte de Turing[2], on peut décrire le jeu imaginé par le mathématicien comme suit : le jeu ne change que très peu : on remplace A par une machine et on se demande si la machine sera autant capable de tromper C en lui faisant croire qu'elle est un être humain. On aura donc une machine A, un être humain B et l'interrogateur C : A essaiera de tromper C, B essaiera de l'aider. Si A est capable de faire croire à C qu'elle est un être humain, alors on pourra dire que A « pense ». Il s'agit donc non pas de découvrir les caractéristiques de la machine et de l'être humain, mais de définir l'un et l'autre à partir de leurs interactions.

A pourra essayer de tromper C, par exemple en laissant passer du temps avant de répondre à une question mathématique. Si C se laisse tromper, cela signifie qu'on est en train de rajouter la caractéristique « lent à calculer » aux caractéristiques définissant l'être humain. Mais si C avait la conviction que les êtres humains peuvent calculer aussi vite que les machines, alors A, pour le tromper, devrait répondre plus rapidement. Encore une fois, dans le test de Turing on ne « découvre » pas qui est l'être humain et qui est la machine : on produit la définition de ce qu'est l'être humain et de ce qu'est une machine.

---

2. Sterrett, Susan G. « Turing's Two Tests for Intelligence ». *Minds and Machines* 10, nº 4 (2000): 541-59.

À partir de ce jeu, et après avoir dédié une bonne partie de son texte à discuter de la manière de définir une « machine », Turing conclut qu'il n'y a aucune raison d'affirmer qu'une machine ne pourrait pas tromper C : les machines donc peuvent penser. La stratégie de Turing consiste à réduire le fait de penser au fait d'avoir un comportement qui induise un observateur à considérer que l'on pense. Le fait que la machine trompe C signifie que C, en observant le comportement de la machine, croit que la machine est un être humain, et donc qu'elle pense. Mais pour Turing il n'y a aucune manière de distinguer un comportement qui semble impliquer de la pensée avec la pensée elle-même.

Préjugés anthropocentriques : la chambre chinoise

Turing prévoit une série d'objections possibles à son argument. Toutes ces objections se basent finalement sur un *a priori* : celui d'une certaine supériorité de l'être humain. Il y aurait des choses dont les êtres humains sont capables et pas les machines. Cela est très révélateur : en réalité, la question à propos de l'intelligence des machines n'a pas comme but celui d'identifier leurs potentialités, mais plutôt de justifier la supériorité humaine.

La préoccupation qui oriente l'investigation sur l'intelligence artificielle est en réalité celle de définir l'être humain comme quelque chose qui se détache du reste du monde. En ce sens, cette question peut être comprise dans une longue tradition qui, à partir de l'Antiquité, a opposé les êtres humains aux autres animaux ou aux automates (que l'on pense à Descartes), ou certaines ethnies à d'autres (sous prétexte qu'elle seraient plus « humaines »). Le préjugé anthropocentrique est donc le point de départ des objections à l'argument de Turing.

Nous allons nous arrêter plus particulièrement sur la quatrième objection identifiée par Turing : l'argument de la conscience (*argument from consciousness*). Cet argument est aussi celui qui sera repris plus tard par John Searle, dans son célèbre *Gedankenexperiment* de la chambre chinoise[3].

3. John Searle, « Minds, Brains, and Programs ». *Behavioral and Brain Sciences* 3 (3), 1980, p. 417-24.

Nous allons montrer comment, loin d'être une véritable objection au raisonnement de Turing, cet argument est une *petitio principii* dont l'objectif n'est pas de répondre à la question de la capacité d'une machine à penser, mais plutôt de donner une définition de l'être humain comme quelque chose de supérieur par rapport aux autres choses.

L'objection consiste à dire que, pour déterminer si une machine peut penser, il n'est pas suffisant de savoir si elle est capable de se comporter comme si elle était intelligente, mais plutôt de ressentir et d'avoir conscience de ce comportement. La machine peut peut-être écrire un sonnet, mais elle ne pourra pas ressentir son sens et ses implications. La machine, en d'autres termes, n'a pas de conscience. Cet argument refuse de fait le point de départ de Turing qui est, par principe, réductionniste : à défaut d'être capable de définir ce qu'est l'intelligence, Turing affirme qu'elle doit être réduite à sa manifestation externe dans un comportement objectivable.

Pour exprimer son idée, Searle imagine un *Gedankenexperiment* qu'il formule à la première personne – ce qui est, comme nous le verrons, fondamental à sa réussite.

Voici l'expérience imaginée par Searle :

> Supposons que l'on m'enferme dans une pièce et que l'on me donne un grand nombre d'écritures chinoises. Supposons en outre (comme c'est le cas) que je ne connaisse pas le chinois, que ce soit à l'écrit ou à l'oral, et que je ne sois même pas sûr de pouvoir reconnaître une écriture chinoise comme une écriture chinoise distincte, par exemple, d'une écriture japonaise ou de gribouillis sans signification. Pour moi, l'écriture chinoise n'est qu'un tas de gribouillis sans signification. Supposons maintenant qu'après cette première série d'écritures chinoises, on me donne une deuxième série d'écritures chinoises accompagnée d'un ensemble de règles permettant de mettre en relation la deuxième série avec la première. Ces règles sont en anglais, et je les comprends aussi bien que n'importe quel autre anglophone. Elles me permettent de corréler un ensemble de symboles formels avec un autre ensemble de symboles formels, et tout ce que le terme « formel » signifie ici, c'est que je peux identifier les symboles uniquement par leur forme. Supposons maintenant que l'on me donne un troisième lot de symboles chinois ainsi que des instructions, toujours en anglais, qui me permettent de mettre en corrélation des éléments de ce troisième lot avec les deux premiers lots, et que ces instructions me disent comment rendre certains symboles chinois avec certaines formes en réponse à certaines formes qui m'ont été données dans le troisième lot. À mon insu,

> les personnes qui me donnent tous ces symboles appellent le premier lot « un script », le deuxième lot « une histoire » et le troisième lot « des questions ». En outre, ils appellent les symboles que je leur renvoie en réponse au troisième lot « réponses aux questions » et l'ensemble des règles en anglais qu'ils m'ont donné, ils l'appellent « le programme ».
> Maintenant, pour compliquer un peu l'histoire, imaginez que ces personnes me donnent aussi des histoires en anglais, que je comprends, et qu'elles me posent ensuite des questions en anglais sur ces histoires, et que je leur donne des réponses en anglais. Supposons également qu'au bout d'un certain temps, je parvienne à suivre les instructions pour manipuler les symboles chinois et que les programmeurs parviennent à écrire les programmes si bien que, du point de vue extérieur – c'est-à-dire du point de vue de quelqu'un qui se trouve en dehors de la pièce dans laquelle je suis enfermé – mes réponses aux questions sont absolument impossibles à distinguer de celles de locuteurs chinois natifs. Personne, en regardant mes réponses, ne peut dire que je ne parle pas un mot de chinois. Supposons également que mes réponses aux questions en anglais soient, comme elles le seraient sans aucun doute, impossibles à distinguer de celles d'autres personnes de langue maternelle anglaise, pour la simple raison que je suis de langue maternelle anglaise. D'un point de vue externe – du point de vue de quelqu'un qui lit mes « réponses » – les réponses aux questions chinoises et aux questions anglaises sont également bonnes. Mais dans le cas du chinois, contrairement au cas de l'anglais, je produis les réponses en manipulant des symboles formels non interprétés. En ce qui concerne le chinois, je me comporte simplement comme un ordinateur ; j'effectue des opérations de calcul sur des éléments formellement spécifiés. En ce qui concerne le chinois, je suis simplement une instanciation du programme informatique[4].

Le point de Searle est clair. Il fait appel au ressenti que nous pouvons avoir d'une telle situation : nous pouvons manipuler des symboles sans comprendre ce qu'ils signifient, ou nous pouvons les manipuler en les « comprenant ». Cela permet de faire une distinction nette entre syntaxe et sémantique : le sens ne peut pas être réduit à la syntaxe car, dans notre expérience, le sens est quelque chose qui s'ajoute à la syntaxe.

---

4. John Searle, « Minds, Brains, and Programs ». *Behavioral and Brain Sciences* 3 (3), 1980, p. 417-24.

Cela permet à Searle de conclure que les êtres humains peuvent comprendre et que la compréhension est liée au sens, tandis que les machines ne peuvent que manipuler de la syntaxe et que donc :

Ma voiture et ma calculatrice, en revanche, ne comprennent rien : elles ne font pas partie de ce secteur d'activité.

Les machines peuvent donc réussir le test de Turing en se comportant comme si elles étaient intelligentes – ou pour être plus précis, comme si elles pensaient – mais, en réalité, elles ne le sont pas.

Or le problème fondamental de cet argument est qu'il n'est valable – comme le soulignait déjà Turing – que d'un point de vue subjectif : je suis le seul à pouvoir dire que « je comprends ». Je suis le seul à pouvoir dire que « je ressens quelque chose ». Qu'est-ce que le « sens » ici ? Comment peut-il être défini ? Il semblerait que le sens est justement ce qui excède la définition que nous avons donnée de machine : la machine manipule de la syntaxe tandis que l'être humain fait quelque chose de plus. Mais ce plus est juste un sentiment subjectif que Searle ne prend pas la peine de spécifier. En d'autres termes, si le sens est irréductible à la syntaxe, c'est parce qu'il a été défini de cette manière : il s'agit donc d'une *petitio principii*. L'argument anti-réductionniste (la conscience est irréductible à un comportement externe) est utilisé comme un dogme pour prouver une supériorité. Il s'agit bel et bien d'un dogme, car il n'est défini par rien d'autre qu'un ressenti subjectif dont on affirme que la machine ne peut pas l'avoir.

S'il est vrai, en effet, que dans l'expérience mentale suggérée, le sujet pourra faire la différence entre des symboles dont il ne connaît pas le sens et des symboles dont il connaît le sens, rien ne dit, dans cette expérience, que le « ressenti » des machines ressemblera plus au premier qu'au deuxième cas. Voici donc la *petitio principii* :

1. On définit la machine comme une chose qui manipule des symboles sans en comprendre le sens, exactement comme un sujet qui ferait l'expérience des signes chinois.

2. On cherche, parmi les expériences de ce sujet, quelque chose qui ne corresponde pas à cette première expérience.

3. On affirme que l'être humain a des capacités supérieures à la machine parce qu'il sait faire quelque chose en plus par rapport à ce qu'on a défini comme les limites de la machine.

Pour mieux nous expliquer, essayons de changer l'expérience de Searle et de l'analyser autrement. Prenons en considération deux symboles pris dans un ensemble défini : *a* et *b*. Admettons qu'il existe

une table qui précise que *a* doit être suivi de *b*. Il s'agit d'une instruction semblable à celles décrites par Searle. Si on définit une machine comme une entité capable de manipuler de la syntaxe, alors cette machine saura réaliser cette opération : lorsque l'entrée est *a*, elle donnera la sortie *b*. Ce n'est que de la syntaxe. Il n'y a pas de sens.

L'exemple suivant est presque identique. Il existe deux symboles composés d'un ensemble de caractères : *Qu'est-ce que le soleil?* et *Une étoile.* Il existe ensuite une table qui précise que *Qu'est-ce que le soleil?* doit être suivi de *Une étoile*.

Quelle est la différence entre ces deux exemples ? D'un point de vue syntaxique, il n'y a aucune différence. Mais il y a une différence s'il n'y a plus de table de correspondance : une machine ne sera pas capable de combiner le premier symbole avec le second. Mais quelqu'un – ou quelque chose – qui est capable d'aller au-delà de la syntaxe et de comprendre le sens, sans avoir besoin d'une table, sera capable de combiner les deux. En effet, il « comprendra » ce que signifient « soleil » et « étoile ». Pour renforcer cet exemple, il est possible de faire appel à un pronom et, en supposant qu'il se réfère à une structure autonome et bien définie, de dire : « Ce que je ressens quand je lis "a" puis "b" est différent de ce que je ressens quand je lis "Qu'est-ce que le soleil?" puis "Une étoile" ».

Le problème est que cet exemple, exactement comme celui de Searle, est conçu pour définir clairement l'ensemble des éléments disponibles dans le premier cas et pour laisser ambigus les éléments disponibles dans le second cas. Dans le premier cas, il n'y a que les deux symboles et la table. Dans le second, il y a autre chose. Mais qu'est-ce que ce quelque chose d'autre ? La réponse que l'exemple tente de susciter est la suivante : « le sens », en supposant que le sens est ce qui reste après avoir considéré tous les éléments déclarés. Mais le problème est que ces éléments n'ont pas été déclarés. Et s'il existait de nombreuses tables, chacune définissant des instructions pour établir des relations entre différents symboles ? Une table établissant une relation entre le symbole « Soleil » et le symbole « étoile », une autre établissant une relation entre « Soleil » et une image, une autre établissant une relation entre « étoile » et « brillant », une autre entre « brillant » et « chaud », et entre « étoile » et « objet astronomique constitué d'un sphéroïde lumineux de plasma maintenu par sa propre gravité », et ainsi de suite. Imaginez des millions, des milliards, des trillions de ces relations.

Ces tables de relations peuvent produire des liens entre des symboles, des images, d'autres symboles, des textes, des personnages, et toutes sortes de matériaux. Elles peuvent être à l'origine de « sentiments », que l'on pourrait décrire comme des relations complexes entre de très longues chaînes de textes et d'autres matériaux. La phrase « Je pense que le soleil est beau » est le résultat de relations définies entre les caractères s, o, l, e, i, et l, des entrées de dictionnaire les associant à une définition (comme « le soleil est une étoile »), d'autres définitions, des poèmes sur le soleil, des phrases scientifiques à son sujet et des textes définissant ce qu'est la beauté et un ensemble de comportements humains en réponse à la beauté (expressions faciales, comportements linguistiques et ainsi de suite). Selon cette description, il est très difficile – voire impossible – de faire la distinction entre ce qui se passe à l'intérieur par exemple du cerveau d'un être humain et ce qui se passe à l'extérieur.

On pourrait donc interpréter les comportements neuronaux et l'apprentissage comme le résultat d'un entraînement. Cet entraînement serait rendu possible par la mise en place d'une série de relations, qui pourraient être représentées avec une grande table. Cet « apprentissage » ne serait donc rien d'autre qu'une transcription matérielle de ces relations inscrites sur la table. Une idée d'un tel type est semblable à celle proposée par Giulio Tononi[5], qui propose une compréhension réductionniste de la conscience en tant que réseau de relations : plus le réseau est large, plus il y aura de conscience. En ce sens, la conscience, au lieu d'être une qualité discrète (on est conscient ou on ne l'est pas), est continue (on est plus ou moins conscient).

L'argument de Searle consiste donc à faire appel à une expérience subjective qu'on appelle, sans la définir, « compréhension » et qu'on lie à un autre concept qui n'est pas défini et qu'on appelle « sens ». Ensuite, Searle définit, cette fois de façon précise, ce qu'est la manipulation de la syntaxe : des symboles et des tables de correspondance. Cette définition permet de délimiter la machine. Or, si on soustrait la syntaxe à notre vécu, il reste quelque chose, justement ce qui n'a pas été défini. L'être humain est donc défini à partir d'une soustraction de la définition de machine. Mais la définition de machine a été concoctée justement

5. Giulio Tononi, *Phi: A Voyage From the Brain to the Soul*. Pantheon, 2012.

pour pouvoir laisser quelque chose en excédence : la définition de la machine a été créée ad hoc pour démontrer la supériorité humaine.

L'objectif du *Gedankenexperiment* n'est donc pas de savoir si la machine peut penser, mais de définir ce qu'est l'humain par opposition à quelque chose d'autre, dans ce cas une machine. Ce n'est pas seulement une opposition, par contre, mais aussi, et surtout, une hiérarchisation. En effet, puisque l'humain est défini par soustraction, il est construit de manière à être « quelque chose de plus ».

En analysant de façon plus approfondie nos exemples et notre lecture de Searle, nous pouvons par ailleurs remarquer que la nature de ce « quelque chose de plus » n'est pas la chose la plus importante. Cela peut être le « sens », la « compréhension », dans le cas de Searle, ou, plus précisément, la « conscience ». Cela peut tout aussi bien être l'« intelligence », dans la longue tradition qui définit les êtres humains par soustraction par rapport aux animaux : animal intelligent. On pourrait continuer : le sens politique, les habiletés sociales, les sentiments…

Dans la totalité de ces cas, il s'agit de performer une différence en définissant un objet – la machine, l'animal, l'automate… – et en faisant en sorte que cette définition laisse de côté un petit quelque chose d'indéfinissable qui nous permettra ensuite de caractériser l'être humain comme étant « cela plus cette petite chose » et d'affirmer que, finalement, la petite chose est la plus importante, ce qui garantit la supériorité de l'être humain par rapport au reste du monde. Dans un cas l'être humain sera une machine avec en plus l'attribut de l'intelligence, qui échapperait à la machine ; dans un autre cas l'être humain serait un animal, mais avec en plus l'attribut disons du sens politique qui manquerait aux autres animaux ; dans un autre l'être humain serait comme un grand ordinateur, mais doté, en plus, de sentiments, que la machine ne pourrait pas éprouver.

C'est pour cette raison qu'il faut prendre très au sérieux une boutade que Turing glisse dans son fameux texte. Lorsqu'il analyse l'argument de la conscience, Turing souligne qu'étant donné que la conscience ne peut être qu'une expérience subjective, il est impossible de savoir objectivement non seulement si une machine est consciente, mais aussi si un autre être humain l'est. Je ne peux qu'affirmer que moi, je suis conscient. Le fait de croire que l'autre – qu'il soit humain ou machine – pense devient donc une question éthique. Turing affirme :

« Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks[6]. »

C'est une boutade qui réduit la question éthique à une question de politesse mais, en réalité, cette affirmation est très profonde et nous renvoie aux enjeux politiques cachés derrière une définition de l'humain par soustraction. Le fait que l'être humain soit « cela plus une petite chose » produit une hiérarchisation qui n'est maîtrisable que par celui qui l'affirme. C'est donc le point de vue du Même (pour reprendre la notion levinassienne). C'est Moi qui décide ce qui est humain, car il n'y a que moi qui puisse ressentir « cette petite chose indéfinissable ». La supériorité est affirmée et devient un argument de domination.

De ce point de vue, l'exemple donné par Searle révèle aussi son racisme latent : ce n'est pas un hasard qu'on oppose l'anglais – qui joue le rôle de la langue du sens – au chinois – qui, selon les lieux communs racistes, est la langue de l'incompréhensibilité – dans des phrases comme « Est-ce que je parle chinois ? » par exemple.

Posthuman Studies

Une définition forte de l'être humain, en plus d'être difficilement justifiable si ce n'est pas par des *petitiones principii*, est donc aussi potentiellement très dangereuse d'un point de vue éthique et politique. Cette idée est à la base de plusieurs théories que l'on peut regrouper sous l'étiquette de *Posthuman Studies*[7].

Dans son texte *The Posthuman*, Rose Braidotti[8] souligne le fait que, justement à cause de définitions essentialistes de ce qu'est l'être humain, plusieurs personnes n'ont pas été ou ne sont pas considérées comme humaines, ou sont considérées comme moins humaines que d'autres. Le modèle de l'être humain à partir duquel la définition de ce qui est humain est produite est toujours un homme blanc, ce qui

6. Alan Turing, « Computing Machinery and Intelligence ». *Mind*, 59 (236), 1950, p. 433-460.

7. Karen Barad, *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning.* Duke University Press Books, 2007. Rosi Braidotti, *The Posthuman*. Polity Press, 2013. Cary Wolfe, *What Is Posthumanism? Posthumanities series*, University of Minnesota Press, 2010.

8. Rosi Braidotti, *The Posthuman*. Cambridge, Polity Press, 2013.

a été la cause d'une série de sexismes et de racismes. À partir de ces constats, les Posthuman Studies se questionnent sur ce qu'est l'humain et essayent de mettre entre parenthèses toute prétendue « essence » humaine pour se concentrer sur les dynamiques discursives, sociales, culturelles et politiques à partir desquelles des définitions particulières de l'humain émergent.

Une précision est ici très importante : il faut distinguer, comme le fait de façon précise, entre autres, Carry Wolfe[9], l'approche posthumaniste des théories transhumanistes. Le transhumanisme a comme objectif d'aller au-delà de l'humain en l'augmentant. Le transhumanisme est donc fortement essentialiste quant à la définition d'être humain, et c'est justement à partir d'une définition forte de ce qui est humain que les transhumanistes pensent la possibilité d'une augmentation et d'une multiplication de cette essence. Pour les transhumanistes, il s'agit donc de radicaliser une prétendue essence humaine, de devenir toujours « plus humain » – en créant évidemment des différences toujours plus grandes entre des êtres humains prétendument plus humains et des êtres humains prétendument moins humains. Pour l'approche posthumaniste, il s'agit de faire tout le contraire : il s'agit de remettre en question le fait qu'une définition de l'humain existe, puisse être stable et puisse correspondre à une essence.

Parmi les différentes théories posthumanistes, il me semble que celle développée par Karen Barad dans son *Meeting the Universe Halfway*[10] est particulièrement adaptée pour nous aider à mieux saisir la question qui nous intéresse ici. En suivant l'argumentation de Barad, nous entendons montrer que le fait de se poser des questions comme « Les machines peuvent-elles penser ? » ou encore « Qu'est-ce qui échappe à l'intelligence artificielle ? » est en réalité un mode de production de l'humain. Avec ces questions, nous essayons en réalité non pas de comprendre les limites éventuelles des machines, mais de définir ce qu'est l'humain.

Les considérations de Karen Barad qui nous semblent utiles ici sont au nombre de deux : en premier lieu, ses analyses de la possibilité

---

9. Cary Wolfe, *What Is Posthumanism? Posthumanities series,* University of Minnesota Press, 2010.

10. Karen Barad, *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning. Second Printing edition.* Duke University Press Books, 2007, p. 143 *sq*.

d'isoler des objets et d'en identifier des frontières précises, et, en second lieu, sa théorie de l'« intra-activité ».

Barad souligne le fait que l'identification des objets en tant qu'entités physiquement délimitées par des contours nets est toujours problématique. Ses considérations se basent en premier lieu sur une analyse physique des frontières des corps[11]. Barad montre que l'idée selon laquelle les frontières entre les corps seraient bien définies et identifiables est tout simplement fausse. Le fait d'identifier une personne en tant que personne ne correspond pas avec l'existence d'une frontière physiquement bien saisissable : les corps ne sont pas délimités par des lignes nettes et peuvent difficilement être isolés. En ce sens, il est notamment difficile de décider ce qui fait partie et ce qui ne fait pas partie d'un corps – et c'est ce qui intéresse Barad – dans le cadre d'une expérience de laboratoire. Barad critique l'idée « humaniste » d'un sujet qui observe et qui est clairement séparé du phénomène observé. Les résultats expérimentaux des recherches en physique quantique – particulièrement si on fait référence aux interprétations de Bohr – démontrent qu'il n'est pas possible de séparer l'observateur, le dispositif d'observation et l'objet observé. L'ensemble de ces éléments est en jeu, et les frontières entre ces « choses » ne sont pas données avant l'expérience, mais en sont plutôt le résultat.

Le récit d'une expérience réalisée en 1922 par Otto Stern et Walther Gerlach sert à Barad d'exemple : pendant l'expérience, un des deux chercheurs était en train de fumer un cigare de mauvaise qualité dont la mauvaise combustion a interagit avec l'instrument d'observation en faisant ressortir des résultats qui auraient été invisibles sans cette interaction. Le fait que l'un des deux physiciens soit en train de fumer un mauvais cigare ne devait en principe avoir rien à faire avec l'observation, cependant les résultats auraient été impossibles sans cette fumée. En ce sens, le cigare fait partie de l'observateur et intervient dans l'émergence du phénomène. Ici, non seulement les matériaux impliqués – le corps du chercheur, le cigare – tiennent un rôle fondamental, mais aussi la condition sociale du chercheur en question, qui détermine son choix d'un cigare bon marché. Est-ce que le cigare fait partie du corps de l'observateur ? Est-ce que sa condition sociale en fait partie ? Est-ce que cigare, corps de l'observateur, les instruments de mesure

11. *Ibid.*

et les autres éléments en jeu dans l'expérience peuvent être séparés ? La conclusion de Barad est que cela est impossible : tous ces éléments ensemble constituent le phénomène.

On peut appliquer le même type de raisonnement à notre question sur l'intelligence artificielle : il est impossible de donner des frontières nettes entre être humain et machine pour définir ensuite ce qui relèverait de l'intelligence humaine et ce qui relèverait de l'intelligence artificielle. Turing lui-même se sent obligé de dédier une bonne partie de son fameux texte à la définition formelle de la machine et il reste tout à fait conscient de l'aspect arbitraire de son résultat final. En effet, qu'est-ce qui ferait partie de la machine, et qu'est-ce qui n'en ferait pas partie ? Comment isoler la « machine » ? Prenons l'exemple d'un ordinateur portable qui fait tourner un réseau de neurones pour découvrir des similarités sémantiques dans un corpus de textes. Qu'est-ce qui définirait ses frontières ? Est-ce que, par exemple, l'électricité nécessaire pour le faire fonctionner en fait partie ? Les modes de production de cette électricité ? Les centrales hydroélectriques et donc les masses d'eau qui se déplacent pour produire l'énergie ? Les conditions géopolitiques et économiques qui permettent l'existence de ces centrales ? L'ensemble des contextes techniques, sociaux, culturels qui font en sorte qu'un matériau particulier soit disponible à un prix particulier ? L'ensemble d'articles, de livres et de modes divers de production de la connaissance qui ont fait émerger un modèle particulier de vectorialisation des mots ? Les corpus sur lesquels ces modèles ont été entraînés ? Les personnes qui ont écrit ces textes, ou celles qui les ont transcrits, traduits ? Les différents supports sur lesquels, avant d'être numérisés, ces textes ont circulé ? Les algorithmes d'océrisation utilisés ?

Il ne s'agit pas ici seulement de dire que, quand un ordinateur exécute un programme, ce programme a été pensé et produit par des êtres humains. Il s'agit plus profondément de se rendre compte qu'il est impossible d'identifier quelque chose comme des êtres humains et des ordinateurs si ce n'est d'après leurs interactions. En ce sens il faut reformuler la question « qu'est-ce qui échappe à l'Intelligence artificielle » car il n'est pas possible de séparer l'intelligence qui serait propre à la machine de celle qui serait propre à l'être humain.

Ce constat est à la base de la proposition théorique fondamentale de Barad : ce qu'elle appelle le réalisme agentiel. Barad ne nie pas qu'il y ait quelque chose comme le réel. Elle ne nie pas non plus, au contraire, que l'on puisse faire de la science et que cette science ait une

valeur objective. Elle propose par contre que l'ontologie n'est pas un discours sur des essences stables, mais une analyse de la manière dont, dynamiquement et performativement, le réel est phénomène. Le phénomène acquiert donc une valeur ontologique et il est objectif. Mais cette objectivité consiste dans le fait que les mêmes « intra-actions » – c'est ainsi que Barad les appelle – donnent lieu aux mêmes phénomènes. Il n'y a pas un observateur et un observé, mais des intra-actions qui viennent avant les choses que ces intra-actions relient.

Le concept d'intra-action sert ainsi à Barad pour remettre en question la logique classique de l'interaction, selon laquelle il y aurait d'abord deux objets, deux choses et ensuite l'interaction entre ces deux choses. L'intra-action vient avant les choses, car les choses en sont le résultat.

Ainsi, il n'y a pas un ordinateur, un être humain et ensuite l'interaction entre les deux. Il y a des intra-actions – des forces en jeu, des dynamiques qui impliquent toute une série très large de facteurs qui ne sont identifiables en tant que tels qu'après-coup, après ces intra-actions. Il ne s'agit pas ici seulement d'un changement de perspective, mais d'un véritable modèle ontologique alternatif : l'« essence » ne définit pas des objets, car les intra-actions sont ontologiquement antérieures à l'essence. De ce fait, ce qu'est l'être humain et ce qu'est la machine, c'est le résultat de l'analyse des intra-actions. Et ces frontières et ces définitions d'« objets » se trouvent dans une négociation toujours ouverte.

Barad affirme :

> My posthumanist account calls into question the givenness of the differential categories of human and nonhuman, examining the practices through which these differential boundaries are stabilized and destabilized[12].

Une précision s'impose ici : l'approche que nous proposons, en suivant la théorie de Barad n'est pas une approche nominaliste ou anti-réaliste. Barad parle en ce sens de « réalisme agentiel » : il y a une réalité qui a une valeur ontologique forte ; il ne s'agit pas de dire que « l'être humain » ou « la machine » ne sont que des noms. Il s'agit de comprendre que l'essence est le fruit d'intra-actions et que ces dernières sont bien réelles et matérielles. Ce qui change c'est que l'essence des

12. Karen Barad, *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning,* Duke University Press Books, 2007, p. 66.

« objets » est toujours dérivée et jamais stable et qu'elle émerge comme résultat d'intra-actions dynamiques qui continuent d'évoluer.

Conclusion

Revenons donc à nos questions de départ. Qu'est-ce que l'être humain ? Qu'est-ce que la machine ? Est-ce qu'il y a quelque chose qui « échappe à la machine » et qui donc n'échappe pas à l'être humain ? Est-ce qu'il y a quelque chose qui permet de caractériser l'un et l'autre et de les différencier clairement ?

Il me semble que ces questions continuent à avoir tout leur sens, mais seulement si on est conscient du fait qu'elles n'ont pas une réponse stable. Le fait de poser ces questions permet justement de s'interroger sur les modes de production des définitions d'humain et de machine. Il ne s'agit donc pas de comprendre ce qu'est l'humain et ce qu'est la machine, mais de saisir les éléments discursifs et matériels qui entrent en jeu lorsque nous essayons de produire de telles définitions.

Notre analyse nous a permis, par exemple, de montrer à quel point des présupposés anthropocentriques influencent la production des frontières entre humain et machine et de quelle manière de tels présupposés ont des conséquences potentiellement néfastes d'un point de vue éthique et politique. En réalité, plusieurs approches pour répondre à ces questions, dont celle de Searle, essaient plutôt de démontrer une supériorité de l'être humain. Le présupposé est donc le point de départ de l'analyse de Searle, qui utilise le rapport entre sens et syntaxe pour pouvoir avoir un rapport différentiel et hiérarchique entre l'humain et tout le reste.

La question « Qu'est-ce qui échappe à la machine ? » est donc transformée en instrument de définition et elle devient une thèse : « Est humain ce qui échappe à la machine. » Le rapport humain-machine est l'intra-action qui déclenche la possibilité de définir et de poser des frontières. L'écart entre humain et machine devient une nécessité de l'argumentation, un point de départ logique sans lequel il ne serait plus possible de définir l'humain comme quelque chose qui se démarque du reste du réel.

S'arrêter, comme Barad, sur ces dynamiques ouvertes de production des définitions et des « essences » est fondamental pour plusieurs raisons. Cela nous rend d'emblée conscients de l'aspect arbitraire et artificiel de

toute frontière entre des choses. S'il est sans doute possible de définir l'être humain en opposition à la machine en tant qu'être qui manipule le sens, il est aussi vrai que cette idée dérive d'une analyse particulière des intra-actions en jeu qui exclut certains paramètres, qui en inclut d'autres et qui est orientée par un ensemble de pratiques discursives.

Mais cela signifie aussi que jamais de telles définitions ne devraient être utilisées comme étant « naturelles » : jamais on ne devrait utiliser de telles définitions pour inclure ou exclure quelqu'un ou quelque chose de l'ensemble des humains. « Not all of us can say, with any degree of certainty, that we have always been human, or that we are only that », nous prévient Braidotti en ouvrant son livre sur le posthumanisme[13]. La définition d'être humain a été souvent utilisée pour tracer des frontières qui excluent des personnes, des genres, des communautés. Ces mêmes définitions ont été utilisées pour justifier plusieurs formes de violence envers des animaux, par exemple, ou envers l'environnement.

La question « Qu'est-ce qui échappe à l'IA ? » doit donc toujours rester ouverte et elle doit être une expérience de laboratoire qui nous permette toujours à nouveau de nous questionner sur qui nous sommes, sur notre place dans le réel et, surtout, sur ce que peut signifier ce « nous ». Si elle reste ouverte et si elle ne demande pas une réponse essentialiste, cette question nous oblige à ne pas considérer l'être humain comme une essence établie et stable, mais plutôt comme le résultat d'une négociation de frontières, dont celle entre machine et humain fait partie.

13. Rosi Braidotti, *The Posthuman,* Polity Press, 2013.