Intelligence artificielle

Esprit es-tu là ?

Giovanni Cucci sj, Rome
professeur de philosophie et de psychologie à l'Université grégorienne de Rome.

PHILOSOPHIE

En plus d'avoir un « cerveau », une machine peut-elle avoir une conscience ? La question renvoie à l'éternel et inextricable problème du rapport entre l'esprit et le cerveau, et à l'autre problème, tout aussi complexe, du rapport entre le corps et l'esprit. L'étude du langage et l'expérience du dilemme éthique éclairent le débat.

En plus de faire appel à des disciplines très différentes (philosophie, linguistique, psychologie, psychanalyse, neurologie, génétique, physique, chimie, neurosciences), ces questions ont donné lieu à des hypothèses et à des théories variées et contradictoires, confirmant la difficulté à parvenir à des conclusions définitives et universellement partagées. Il a même été tenté d'éliminer l'un des deux termes - l'esprit - en vain.

Contrairement au cerveau, l'esprit revêt une multitude de sens (conscience, âme) difficiles à préciser, qui, s'ils ne peuvent être reliés au cerveau, ont en revanche chez l'humain un rapport avec l'organisme entier. En ce qui concerne l'intelligence artificielle (IA), on est en droit de douter que la machine puisse avoir un « esprit ».

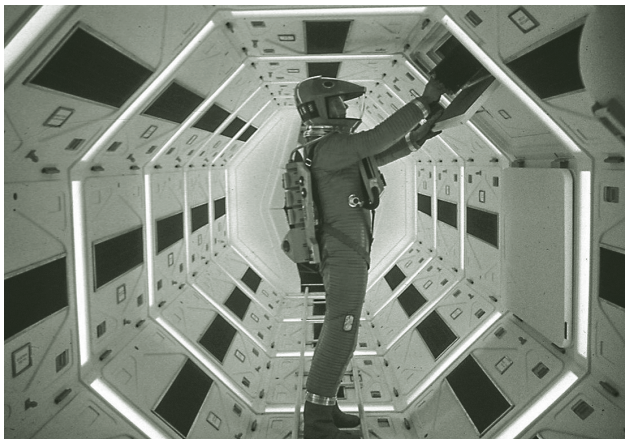
Communiquer n'est pas comprendre

Pour comprendre cette différence fondamentale, le philosophe américain de l'esprit John Searle a imaginé une expérience mentale devenue célèbre : *la chambre chinoise*. Un volontaire se place dans une pièce où se trouvent des lignes de texte en chinois, une langue qu'il ne connaît pas. Il reçoit un manuel d'instructions indiquant les symboles qu'il devra utiliser pour répondre à chaque ligne. Ses réponses seront correctes, mais il ne comprendra toujours pas le chinois.

Cette expérience montre la différence entre le langage humain et un programme informatique qui reçoit et envoie des lignes d'informations sans les comprendre. De ce point de vue, une machine ou un robot ne pourra jamais « parler » comme un être humain. Le programme utilise une procédure, tandis que le langage fait surtout référence au sens des mots et utilise des symboles. En linguistique, on appelle cela la « sémantique ».

Le rapport entre le langage et la santé mentale en dit long sur la dimension biologique, corporelle, vivante du langage humain, dont la sémantique présente un ensemble de règles extrêmement complexes et non codifiables, et pourtant connues de tous. Or la sémantique est absente des programmes informatiques justement parce qu'elle n'est pas « programmable », et surtout parce qu'elle présente des connotations biologiques et affectives. C'est

Adaptation et traduction d'un article paru le 4 janvier 2020 dans la *Civiltà Cattolica*.



2001, L'odyssée de l'espace
© Metro-Goldwyn-Mayer Studios Inc

là d'ailleurs l'une des grandes énigmes de la linguistique. À la différence des machines, les êtres humains ne sont pas programmés par des algorithmes. On utilise volontiers le terme « d'émergence » (quelque chose qui survient) pour indiquer cette dimension de l'esprit humain que l'on ne peut réduire à un algorithme ni à une dimension purement matérielle.

Searle résume ainsi son avis sur la question : les programmes sont complètement syntaxiques ; l'esprit a une sémantique ; la syntaxe n'est pas la même chose et elle n'est pas suffisante en elle-même pour la sémantique. En d'autres termes, il existe un écart de qualité entre le programme et le sens.

Cet écart a été illustré de manière exemplaire en 1968 dans le célèbre film *2001, L'odyssée de l'espace* de Stanley Kubrick, où figure un dialogue entre un astronaute et un superordinateur, Hal 9000, qui a un contrôle absolu sur toutes les opérations à bord du vaisseau spatial. Lorsque le commandant décide de le désinstaller à la suite d'une erreur de calcul, Hal commence à tuer les membres de l'équipage. L'astronaute survivant essaie en vain de lui montrer le sens de sa mission et la valeur de la vie des astronautes ; Hal est incapable de le comprendre et continue de répéter ce pour quoi il a été programmé, jusqu'à interrompre la liaison.

Une logique inadaptée

Comme le langage, l'expérience morale présente, elle aussi, des complexités telles qu'il est impossible de la ramener à un système ou à une théorie exhaustive. De la même façon que pour l'apprentissage d'une langue, il y a une sorte de « grammaire universelle » de la morale, qui est davantage liée aux sentiments qu'à la compétence.

C'est ainsi que les personnes dépourvues de sentiments, les psychopathes, présentent des déficits dans le domaine de la morale. Les recherches menées par le médecin portugais Antonio Damasio dans le domaine neurocérébral ont montré comment la lobotomie, c'est-à-dire l'ablation des lobes frontaux du néocortex cérébral (siège des émotions), induit de graves déficits dans l'évaluation. Ces résultats démentent l'idée reçue selon laquelle un esprit froid et dépourvu d'émotions se trouve dans les conditions optimales pour prendre de bonnes décisions.

Damasio évoque le cas d'un patient, Elliot, à qui les médecins avaient retiré une tumeur en recourant à une lobotomie. Elliot avait gardé intactes ses capacités intellectuelles, linguistiques et de communication, mais n'éprouvait plus aucune émotion. Cette privation avait fait de lui « l'être humain intelligent le plus froid et le moins émotif que l'on puisse imaginer, mais dont la raison pratique était tellement altérée qu'elle l'amenait à commettre plusieurs erreurs dans son quotidien, en violation perpétuelle de ce que vous et moi pourrions considérer comme socialement approprié et avantageux du point de vue personnel ».[1]

Intelligence artificielle

Esprit es-tu là ?

Damasio ajoute que si les lésions adviennent précocement, à l'âge du développement d'une personne, celle-ci sera incapable d'apprendre les règles d'éthique les plus élémentaires. La conclusion du neurologue est que l'action morale est essentiellement liée à l'affect et que, en son absence, elle ne peut être compensée par aucun autre type d'instruction : « Cela ne veut pas dire que ce sont les sentiments (quand ils induisent une action) qui décident pour nous, ou que nous ne sommes pas des êtres rationnels. Je suggère seulement que certains aspects du processus de l'émotion et du sentiment sont indispensables pour la rationalité. »[2]

Le dilemme éthique ne peut être résolu par un algorithme visant à obtenir le résultat maximal pour un coût minimal. Pour les êtres humains, la décision repose sur bien plus.

Le dilemme éthique

Les recherches de Damasio peuvent nous aider à identifier une différence importante dans la façon de procéder, selon que l'on est un humain ou une IA, face à un « dilemme éthique ». Ce concept a été rendu célèbre par la philosophe anglaise Philippa Ruth Foot.[3] Elle a recouru à l'exemple d'un train hors de contrôle qui risque de finir sur un groupe d'ouvriers en pleine activité ; il est impossible de prévenir ceux-ci, mais il est possible d'actionner un levier qui dirigera le train sur une voie désaffectée, sur laquelle se trouve néanmoins un ouvrier. C'est le choix, terrible mais inévitable, du moindre mal. Ce problème n'est pas seulement quantitatif (un groupe de personnes contre une seule personne), il se pose aussi en termes de responsabilité car, dans ce cas, la personne dirige volontairement le train vers la voie désaffectée, entraînant la mort de l'ouvrier. Une version encore plus tragique du dilemme précise qu'il s'agit d'une personne avec laquelle le conducteur a un lien affectif (un ami ou un parent).

Comment les machines réagissent-elles face à un dilemme éthique ? Une voiture sans chauffeur ferait, elle aussi, le choix du moindre mal, mais elle ne se sentirait pas coupable d'avoir tué. Chez l'humain, par contre, le manque d'alternatives n'atténue pas la peine et le remords d'avoir pris une terrible décision.

Le roman de William Clark Styron, *Le choix de Sophie* (1979), adapté trois ans plus tard au cinéma par Alan Pakula, le montre bien. Lors de sa déportation à Auschwitz, la protagoniste, Sophie, est confrontée à un horrible choix : un garde sadique lui ordonne de décider lequel de ses deux enfants ira dans la chambre à gaz. Si elle refuse, les deux mourront. Désespérée, Sophie choisit de garder son fils, espérant que son bourreau changera d'avis, ce qui ne se produit pas. Elle gardera au fond d'elle-même le poids de ce drame pendant de longues années, puis se suicidera.

Le dilemme éthique ne peut être résolu par un algorithme visant à obtenir le résultat maximal pour un coût minimal. Pour les êtres humains, la décision repose sur bien plus. Le sentiment de responsabilité

renvoie à quelque chose de différent sur le plan qualitatif, comme le remords, la repentance, la tristesse, la culpabilité, la rédemption.

« L'éthique utilitariste (conséquentialisme) nie l'existence d'authentiques dilemmes moraux, écrivent le philosophe Julian Nida-Rümelin et l'écrivain Nathalie Weidenfeld. La raison de cette négation est évidente : si l'action est évaluée d'après un critère d'optimisation, il ne peut y avoir aucun conflit [...]. Les ordinateurs numériques sont par définition des machines de Turing et fournissent des résultats univoques. Ils ne sauraient être un modèle de raison pratique, ne serait-ce que pour cette raison. »[4] Face au dilemme, une machine restera dans l'incertitude mais ne se sentira jamais coupable. Et, contrairement à Sophie, elle n'en viendra pas au suicide.

Pour un humanisme numérique

En excluant toute comparaison avec la dimension sapientielle de la vie, la mentalité technologique risque de s'approcher dangereusement de la folie et de la perte de sens. Dans sa célèbre analyse de la domination de la technique dans l'époque moderne, Martin Heidegger remarquait en 1953 déjà que le problème central ne réside pas dans la mesure de cette domination, mais plutôt dans le fait que l'homme n'est pas préparé à la vivre de manière critique et consciente, en soupesant les avantages et les pertes possibles.

Commentant le dialogue entre l'astronaute et Hal 9000 dans *2001, L'odyssée de l'espace*, Julian Nida-Rümelin et Nathalie Weidenfeld remarquent que ce n'est pas un hasard si l'ordinateur ressemble à un œil de verre rouge et noir, les couleurs que l'imaginaire chrétien attribue à l'enfer : « L'enfer est un lieu dans lequel l'homme a donné le pouvoir de vie et de mort à des ordinateurs programmés de manière conséquentialiste, qui sont incapables de penser vraiment. »[5]

La quantité accrue de données et de ressources requiert aussi une évaluation, que l'IA peut au maximum suggérer mais dont elle ne pourra jamais être la dernière instance. L'enjeu a été bien résumé par l'écrivain mexicain Naief Yehya : « Avec un ordinateur, nous pouvons transformer pratiquement tous les problèmes humains en statistiques, en graphiques, en équations. Le plus inquiétant est que, ce faisant, nous créons l'illusion que ces problèmes peuvent être résolus avec les ordinateurs. »[6]

Un dialogue des plus attentifs entre les innovations technologiques et les sciences humaines, en vue d'un « humanisme numérique », est donc une nécessité. Le débat sur l'enjeu et la décision à mettre en œuvre doivent toujours rester, en dernière analyse, dans les mains de l'homme qui est, depuis toujours, un *homo sapiens*. ■

1 **Antonio Damasio**, *L'erreur de Descartes. La raison des émotions*, Paris, Odile Jacob 1995.

2 *Idem.*

3 **Philippa Ruth Foot**, « The Problem of Abortion and the Doctrine of double Effect », in *Oxford Review 5*, Oxford 1967, pp. 5-15.

4 **Julian Nida-Rümelin**, **Nathalie Weidenfeld**, *Digitaler Humanismus*, Piper Verlag Munich 2018, 224 p.

5 *Idem.*

6 **Naief Yehya**, *Homo cyborg. Il corpo postumano tra realtà e fantascienza*, Milano, Elèuthera 2005, p. 15.