Daniel Andler

Intelligence artificielle, intelligence humaine : la double énigme

Gallimard

Introduction

Une femelle dauphin, enceinte, signale à une chercheuse qui l'étudie qu'elle est elle-même enceinte, ce que la jeune femme ignorait. Cette anecdote, dont nous admettrons la véracité pour les besoins de cette entrée en matière[1], n'a rien à voir avec l'intelligence artificielle. Elle rapporte une manifestation d'intelligence naturelle, animale en l'occurrence, qui est un exemple de ce que je propose d'appeler une énigme. Je distingue l'énigme à la fois du mystère et du problème. Le mystère nous dépasse, la distance qui le sépare de notre entendement semble trop grande pour que nous imaginions pouvoir la combler. Le problème quant à lui se présente comme une tâche à notre portée. L'énigme se situe entre les deux : elle nous sidère, nous paralyse, mais nous met au défi de la résoudre. Et ce que nous espérons, ce sont deux explications et non une seule : la première fournit la clé de l'énigme, la dénoue ; la seconde nous fait comprendre ce qui en faisait une énigme.

En l'espèce, la première explication met au jour une chaîne causale, allant de l'écholocalisation à la communication entre les deux créatures, en passant par plusieurs étapes dont certaines restent à combler. Et ce sont nos conceptions de l'intelligence d'un animal, de ce dont il est capable en matière de détection, de compréhension, de rapport à soi et à autrui, de communication qui nous présentent l'histoire sous forme d'une énigme. C'est la deuxième explication : nous ne voyons pas comment une chose pareille est seulement possible, tant que l'éthologie cognitive n'est pas venue à notre aide.

Le titre de ce livre, *Intelligence artificielle, intelligence humaine : la double énigme*, est une façon — un peu énigmatique

peut-être — d'annoncer plusieurs thèses : que l'intelligence artificielle constitue une énigme, que l'intelligence humaine en est une autre, que ces deux énigmes sont étroitement liées, et enfin qu'elles ne sont pas des mystères — de fait, je prétends les résoudre.

Commençons par une surprise. Selon le principe « *verum-factum* » de Vico, on connaît ce qu'on fabrique ; mieux : la seule manière de s'assurer qu'on comprend vraiment un objet, un phénomène quelconque, qu'on sait comment il marche, c'est d'en fabriquer un — s'agissant d'un objet, ou de le produire — s'agissant d'un phénomène. L'ordinateur, comme tout objet fabriqué, semble tomber sous ce principe : nous le connaissons dans ses moindres détails. Et pourtant — c'est la surprise — mis à l'ouvrage il devient imprévisible. Bien évidemment, ce qu'il fait d'un programme donné sur une entrée particulière n'est généralement pas connu d'avance — sans quoi on s'en passerait. Ce qui est gênant, c'est qu'on n'a pas toujours le moyen de s'assurer que le résultat est conforme aux spécifications qu'on a posées : on ne sait pas, même approximativement, ce que fait l'ordinateur en toute circonstance. Cette violation du « *verum-factum* » fait de l'informatique une science à la fois formelle et empirique : une fois l'algorithme conçu et codé, il faut observer son comportement et voir s'il est conforme à nos attentes. L'intelligence artificielle hérite de cette dualité, et c'est pourquoi l'étude de ses fondements est de première importance pour ses applications technologiques.

On pourrait penser que l'intelligence artificielle ne pose pas de problème théorique pressant, mais seulement des problèmes techniques ou bien sociaux, juridiques, économiques. La raison en serait qu'elle n'est une simple copie — sans doute imparfaite — de l'intelligence humaine. L'imprédictibilité dont il vient d'être question montre que c'est une erreur. Un autre mauvais argument serait que la réflexion sur l'intelligence humaine, dont personne n'imagine qu'elle soit dépourvue d'intérêt théorique, a pu se développer pendant de longs siècles avec profit, comme en témoignent l'immense corpus de la philosophie de la connaissance et la tradition rationaliste dans son ensemble, sans se référer à l'intelligence artificielle, qui jusqu'à récemment n'existait pas. Ce serait ignorer que les idées clés de l'intelligence artificielle figurent déjà *in ovo* dans la philosophie des siècles précédents, notamment chez Hobbes

et Leibniz, et chez les logiciens d'Aristote à Boole et Frege. Alan Turing recueille cet héritage et le fait fructifier en ce qui est de loin la plus importante mutation technique et scientifique de l'époque contemporaine, l'informatique et sa branche avancée, l'intelligence artificielle[2].

En un sens, l'intelligence artificielle est donc la philosophie de la connaissance — des bases de l'intelligence — poursuivie par d'autres moyens. Ses succès comme ses échecs sont autant d'éléments de réflexion sur l'intelligence humaine. Il est vrai que la profession (je la désigne par le sigle IA, en la distinguant de ce qu'elle produit, que j'écrirai « intelligence artificielle » en toutes lettres), dans l'ensemble, ne fait appel à l'intelligence humaine que pour désigner les processus à (re)créer dans l'ordinateur. Elle ne cherche pas à les analyser ou à en dégager les propriétés générales. Il est vrai également que la modestie de ses premiers résultats a incité beaucoup de chercheurs, en dehors de la profession, à s'en détourner et à s'intéresser à la psychologie cognitive et autres branches d'étude de l'intelligence naturelle (ce fut mon cas). Mais le vent a tourné : devenue puissante, l'intelligence artificielle suscite de nouvelles interrogations. Elle s'est muée, particulièrement depuis le tournant du siècle, en une vaste entreprise dont les ramifications s'étendent partout ; l'intelligence artificielle incorporée dans les systèmes qu'elle construit est presque infiniment plus puissante — en un sens à préciser — que celle des époques antérieures ; en sorte que ses effets économiques, sociaux, culturels sont sans commune mesure avec ceux d'alors.

Tout cela constitue le contexte du présent ouvrage, et contribue à en justifier l'existence. Mais ce n'est pas son objet : ce sont des bases conceptuelles de l'intelligence artificielle d'aujourd'hui qu'il va être question. Il s'agit de comprendre quelles sources théoriques ont alimenté cette vertigineuse croissance du domaine, quels progrès elles ont fait faire dans la compréhension de l'intelligence naturelle, à quelles limites se heurte aujourd'hui le projet, quelles bornes devront lui être imposées… Bref, il va être question de l'IA d'aujourd'hui. Elle est dominée par les techniques du deep learning (ou apprentissage profond), avatar du connexionnisme, elle ne s'y réduit pas. L'approche symbolique des commencements, à laquelle elle a succédé, n'a pas disparu. Aucun de ces deux paradigmes, selon le jugement des meilleurs spécialistes, ne permettra à

l'IA de réaliser ses plus hautes ambitions. Il n'empêche que dès à présent elle alimente la société en systèmes aux capacités étonnantes, mal expliquées mais incontestables, dont il est crucial de mieux comprendre le fonctionnement, les limites et les risques qu'ils comportent, ce qui exige de revenir sur les questions théoriques posées depuis des décennies.

Il est temps de signaler le flou qui entoure désormais le terme « intelligence artificielle ». Certains pensent que ce flou cache une confusion si grande qu'il faudrait y renoncer et changer d'appellation : pour le public, le moindre algorithme, l'app la plus simple, l'objet connecté le plus ordinaire, tout cela c'est « de l'IA ». Je propose pour ma part de conserver l'expression — nous n'avons d'ailleurs pas le choix : elle est consacrée — mais de distinguer le sens strict d'un sens large : au sens strict, l'IA consiste en un ensemble relativement bien défini d'institutions, de projets et de réalisations d'ordre théorique et technologique, s'inscrivant dans une tradition intellectuelle précise. C'est à elle qu'est consacré ce livre pour l'essentiel. L'intelligence artificielle au sens large inclut l'intelligence artificielle au sens strict mais également tout ce qui contribue au perfectionnement de la « numérisphère » (l'ensemble des processus et systèmes numériques, y compris internet, qui forment désormais une strate supplémentaire du monde humain). L'IA dont il est dit, avec raison, qu'elle exerce de puissants effets sur la société doit être comprise au sens large. Mais les idées de l'IA au sens strict y jouent un rôle déterminant et qui ne fait que croître.

L'IA n'est pas la seule technologie universelle au sens où elle pénètre peu à peu la quasi-totalité des sphères d'activité : l'électricité est l'autre exemple fréquemment cité. Mais il y a une différence essentielle. On comprend assez bien ce dont l'électricité est capable et la manière dont elle affectera un domaine donné avant qu'elle ne l'investisse. Ce n'est pas le cas de l'IA. Il est très difficile de deviner ce dont un système d'IA donné (ce que je noterai un SAI — système artificiel intelligent), déployé dans un certain domaine, est capable, et pourquoi il réussit ou non à faire ce qu'on attend de lui. Il peut être trop faible, mal conçu, mal adapté, et ne causer que du gâchis. Il peut être puissant mais faire ce qu'on n'attend pas de lui, provoquant des dégâts qui peuvent être graves. Il peut donner satisfaction (ce qui est manifestement assez souvent le cas, sans quoi l'IA

serait aujourd'hui rangée au magasin des fausses bonnes idées, comme il lui est presque arrivé à plusieurs reprises), mais sans qu'on comprenne vraiment pourquoi ni dans quel domaine — pour quelles valeurs des entrées — on peut compter sur un résultat correct.

Venons-en aux thèses cachées derrière le titre.

La première est que l'intelligence artificielle est une énigme. Elle contrevient au jugement courant, qui considère qu'elle est produite par des choses que nous fabriquons — ordinateurs et logiciels ou algorithmes — et qu'elle consiste en copies de fragments de l'intelligence humaine — où par « copie » il faut comprendre « version automatisée ». Que la construction d'ordinateurs et l'automatisation de procédures intellectuelles sous forme d'algorithmes posent des problèmes techniques n'est pas douteux. Mais où serait l'énigme ? Elle réside en ceci : la cible de l'IA est une intelligence artificielle qui soit l'égale de l'intelligence humaine. Or cette cible ne semble jamais se rapprocher, alors même que l'IA progresse constamment.

Voici les deux explications que je propose pour lever cette énigme. La première est que l'intelligence, au sens humain du terme, n'est pas une fonction définie, à l'image d'une fonction mécanique ou biologique, dont on pourrait doter un mécanisme, si complexe qu'il soit. Chaque pas accompli par l'intelligence artificielle semble consister à découvrir que l'intelligence n'est pas là où elle pensait en trouver une trace — c'est la « malédiction de l'IA[3] ». La cible de l'IA est une chimère, telle est la clé que je propose. Quant à ce qui explique notre sidération initiale, c'est que l'intelligence se manifeste dans chaque situation particulière par le déploiement de capacités dont chacune semble à la portée de l'intelligence artificielle : cela crée l'illusion que l'intelligence n'est rien d'autre que l'ensemble de ces capacités, et que l'intelligence artificielle accédera à la pleine intelligence lorsqu'elle sera effectivement parvenue à les reproduire toutes. En attendant, elle devrait donc se rapprocher du but, et ce n'est pas le cas.

Selon la deuxième thèse, l'intelligence (humaine) est une énigme. Contrairement à la première, elle ressemble à un poncif : l'intelligence est réputée « insaisissable », elle est un mystère. Poncif auquel s'oppose une autre thèse que la mienne, à savoir que l'intelligence ne serait qu'un simple problème, que

les sciences cognitives sont en train de résoudre. Or l'intelligence n'est selon moi ni un mystère, ni un phénomène en voie d'être expliqué à la manière dont le sont, par exemple, la perception de la profondeur, l'accès lexical, la mémoire autobiographique ou les erreurs systématiques de raisonnement. Loin d'être insaisissable, elle est constamment maniée par la pensée et le langage. Elle est également mesurée, selon des procédures qui survivent aux critiques dont elles sont abreuvées. Peu de gens doutent qu'elle ait des manifestations caractéristiques et qu'elle s'attache de manière régulière et durable, à des degrés différents, à différentes personnes dans différentes situations. Il est donc largement admis qu'il s'agit d'une propriété réelle qui se repère à ses manifestations. Sans qu'elle constitue un mystère, elle résiste à nos tentatives pour la caractériser à l'aide de termes clairs. Son caractère énigmatique fait d'ailleurs partie du concept qu'en ont la plupart des gens : dire de quelqu'un qu'il est intelligent, ou qu'il ne l'est pas, qu'il l'est plus ou qu'il l'est moins, c'est indiquer qu'en le disant, tout est dit ; qu'on ne saurait le dire autrement.

L'énigme se résout en deux temps. Premier temps : l'intelligence n'est pas une chose, phénomène, processus, fonction, mais une norme qui s'applique au comportement : elle qualifie le rapport entre un individu et son monde, d'une manière qui n'est jamais objective et finale comme l'est la mesure en centimètres d'un bout de ficelle, sans être purement subjective pour autant. Deuxième temps : du fait que l'intelligence (humaine) se manifeste dans chaque situation par le déploiement de certaines facultés cognitives ou mentales, on tend à l'assimiler à l'ensemble de ces facultés ; mais on a beau examiner ces facultés une à une, par introspection ou observation ordinaire, ou de manière scientifique, on ne trouve aucune trace d'intelligence. C'est ce qu'on a observé pour l'intelligence artificielle.

La troisième thèse est que les deux énigmes sont étroitement liées. Elle semble de l'ordre de l'évidence : ne partagent-elles pas la référence au concept d'intelligence ? Mais si l'on conçoit l'intelligence artificielle comme une simple copie de l'intelligence humaine, il n'y a qu'une seule énigme et non deux. Il faut chercher un peu plus loin pour trouver le rapport. Une première piste réside dans le terme « copie » : qu'est-ce qu'une copie de l'intelligence ? Pour répondre à la question, il faut tenir en mains à la fois l'original et la copie, ou plus préci-

sément l'original et le matériau dont doit être faite la copie, la toile sur laquelle elle sera peinte. La discussion des deux thèses précédentes fournit une deuxième piste. Les premières explications proposées aux deux énigmes se ressemblent, et les secondes se correspondent étroitement : elles reposent sur une même conception componentielle faisant de l'intelligence artificielle comme de l'intelligence humaine un répertoire de facultés spécialisées.

Mais la preuve la plus massive de la connexion entre les deux énigmes est fournie par le fait que le projet de l'intelligence artificielle est historiquement associé — mieux, identifié — au projet de ce qui s'appelle désormais les sciences cognitives, et qui était compris à l'époque comme l'étude de l'intelligence humaine. Si les deux entreprises ont pris depuis leur indépendance, l'intelligence humaine continue de dicter l'ordre du jour de l'intelligence artificielle, qui consiste à accomplir, une à une, les tâches qui mobilisent chez nous l'intelligence mais en s'en passant. Inversement, l'intelligence artificielle dans ses réalisations concrètes éclaire, parfois en creux, tel ou tel aspect de l'intelligence humaine. Toute enquête d'ordre philosophique sur l'intelligence artificielle implique une réflexion sur l'intelligence humaine, soit directement dans le programme de la philosophie de l'esprit (l'étude des fondements de la psychologie), soit dans celui de la branche de la philosophie des sciences consacrée aux sciences cognitives — les deux entreprises étant d'ailleurs très proches.

Tout cela est trop vite dit. La lecture du livre le rendra j'espère plus clair. On y trouvera plusieurs conclusions, dont l'exposé résumé serait fastidieux. Je n'en retiens ici que deux.

La première est que la poursuite d'une intelligence artificielle dotée d'une intelligence humaine est sans objet : selon la conception de l'intelligence que je défends, l'intelligence au sens humain ne peut s'attacher qu'à un être humain. L'intelligence artificielle, sous quelque forme et à quelque degré de développement qu'elle soit, a pour vocation de résoudre des problèmes, ce qui n'est pour l'intelligence humaine qu'une mission secondaire.

La seconde conclusion concerne les efforts que consacre l'IA, avec une énergie désormais décuplée, à la conception de systèmes toujours plus intelligents, c'est-à-dire, selon elle, tou-

jours plus proches de l'intelligence humaine. Elle vise aussi à conférer à ces systèmes une autonomie aussi large que possible, et à la limite une autonomie totale. Ce double objectif est à la fois incohérent, dangereux et inutile. Incohérent parce que l'autonomie véritable, celle des humains, est de l'ordre du mystère : un SAI ou un robot autonome en ce sens est un concept inintelligible. Dangereux, parce qu'il ne peut conduire qu'à des systèmes difficiles à comprendre et plus encore, par définition, à contrôler. Inutile, parce que ce dont nous avons vraiment besoin, ce qui nous aiderait de manière peut-être décisive à résoudre certains problèmes pressants, c'est d'outils dociles, puissants et versatiles, et non de pseudo-personnes munies d'une forme inhumaine de cognition. Les ingénieurs devraient se consacrer à construire des attelages robustes d'humains et de systèmes artificiels, conçus en sorte que le rôle principal soit confié aux premiers, non seulement en théorie mais par construction, en sorte d'empêcher qu'ils l'abandonnent sous la pression conjuguée de l'habitude, de la paresse et des exigences de productivité.

Enfin, loin de chercher à introduire des SAI aussi puissants que possible partout où c'est possible, nous devrions épouser un principe de modération — par quoi j'entends non un principe de précaution, mais l'analogue du principe de subsidiarité : ne recourir aux SAI que lorsque les avantages, toutes choses considérées, l'emportent largement sur les inconvénients et dangers prévisibles ; et quand cette condition est remplie, nous limiter à des SAI juste assez puissants pour assumer les tâches que nous leur assignons.

Paris, le 1er août 2022