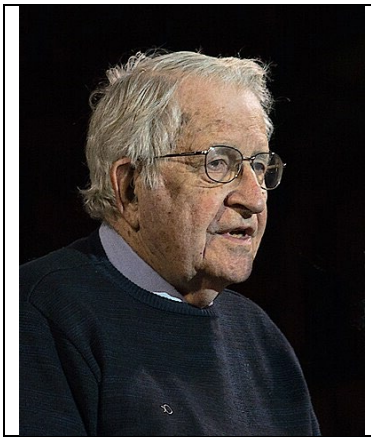

La réalité sur l'intelligence artificielle

Jorge Luis Borges a écrit un jour que vivre à une époque de grands périls et de grandes promesses, c'est faire l'expérience à la fois de la tragédie et de la comédie, avec « l'imminence d'une révélation » dans la compréhension de nous-mêmes et du monde. Aujourd'hui, nos avancées prétendument révolutionnaires en matière d'intelligence artificielle sont en effet une source d'inquiétude et d'optimisme. L'optimisme parce que l'intelligence est le moyen par lequel nous résolvons les problèmes. Inquiétude parce que nous craignons que la souche la plus populaire et la plus à la mode de l'IA – l'apprentissage automatique – ne dégrade notre science et avilisse notre éthique en incorporant dans notre technologie une conception fondamentalement erronée du langage et de la connaissance.

ChatGPT d'OpenAI, Bard de Google et Sydney de Microsoft sont des merveilles de l'apprentissage automatique. Grosso modo, ils prennent d'énormes quantités de données, y recherchent des modèles et deviennent de plus en plus compétents pour générer des résultats statistiquement probables, tels qu'un langage et une pensée apparemment humains. Ces programmes ont été salués comme les premières lueurs à l'horizon de l' *intelligence artificielle générale* – ce moment prophétisé depuis longtemps où les esprits mécaniques surpassent les cerveaux humains non seulement quantitativement en termes de vitesse de traitement et de taille de mémoire, mais aussi qualitativement en termes de perspicacité intellectuelle, de créativité artistique et de toute autre faculté distinctement humaine.

Ce jour viendra peut-être, mais son aube ne se lève pas encore, contrairement à ce que l'on peut lire dans les gros titres hyperboliques et compter sur des investissements peu judicieux. La révélation borgésienne de la compréhension n'a pas eu lieu et ne se produira pas – et, selon nous, ne peut pas se produire – si les programmes d'apprentissage automatique comme ChatGPT continuent de dominer le domaine de l'IA. Aussi utiles que puissent être ces programmes dans certains domaines étroits (ils peuvent être utiles en programmation informatique, par exemple, ou pour suggérer des rimes pour des vers légers), nous savons par la science de la linguistique et la philosophie de la connaissance qu'ils diffèrent profondément de la façon dont les humains raisonnent et utilisent le langage. Ces différences imposent des limitations importantes à ce que ces programmes peuvent faire, les codant avec des défauts indéracinables.

Il est à la fois comique et tragique, comme Borges aurait pu le noter, que tant d'argent et d'attention soient concentrés sur une si petite chose – quelque chose de si trivial lorsqu'on le compare à l'esprit humain, qui, à force de langage, selon les mots de Wilhelm von Humboldt, peut faire « un usage infini de moyens finis », créant des idées et des théories à portée universelle.

L'esprit humain n'est pas, comme ChatGPT et ses semblables, un moteur statistique lourd pour la correspondance de modèles, se gorgeant de centaines de téraoctets de données et extrapolant la réponse conversationnelle la plus probable ou la réponse la plus probable à une question scientifique. Au contraire, l'esprit humain est un système étonnamment efficace et même élégant qui fonctionne avec de petites quantités d'informations ; Il ne cherche pas à déduire des corrélations brutes entre les points de données, mais à créer des explications.

Par exemple, un jeune enfant qui acquiert une langue développe – inconsciemment, automatiquement et rapidement à partir de données minuscules – une grammaire, un système prodigieusement sophistiqué de principes et de paramètres logiques. Cette grammaire peut être comprise comme une expression du « système d'exploitation » inné, génétiquement installé, qui dote les humains de la capacité de générer des phrases complexes et de longs trains de pensées. Lorsque les linguistes cherchent à développer une théorie expliquant pourquoi une langue donnée fonctionne comme elle le fait (« Pourquoi ces phrases – mais pas celles-là – sont-elles considérées comme grammaticales ? »), ils construisent consciemment et laborieusement une version explicite de la grammaire que l'enfant construit instinctivement et avec une exposition minimale à l'information. Le système d'exploitation de l'enfant est complètement différent de celui d'un programme d'apprentissage automatique.

En effet, de tels programmes sont coincés dans une phase préhumaine ou non humaine de l'évolution cognitive. Leur défaut le plus profond est l'absence de la capacité la plus critique de toute intelligence : dire non seulement ce qui est le cas, ce qui était le cas et ce qui sera le cas – c'est la description et la prédiction – mais aussi ce qui n'est pas le cas et ce qui pourrait et ne pourrait pas être le cas. Ce sont là les ingrédients de l'explication, la marque de la véritable intelligence.

Voici un exemple. Supposons que vous teniez une pomme dans votre main. Maintenant, vous laissez aller la pomme. Vous observez le résultat et vous dites : « La pomme tombe. » C'est une description. Une prédiction aurait pu être l'affirmation suivante : « La pomme tombera si j'ouvre ma main. » Les deux sont précieux, et les deux peuvent être corrects. Mais une explication, c'est quelque chose de plus : elle comprend non seulement des descriptions et des prédictions, mais aussi des conjectures contrefactuelles comme « N'importe quel objet de ce type tomberait », plus la clause supplémentaire « à cause de la force de gravité » ou « à cause de la courbure de l'espace-temps » ou autre. C'est une explication causale : « La pomme ne serait pas tombée sans la force de gravité. » C'est ce qu'on appelle la pensée.

Le cœur de l'apprentissage automatique est la description et la prédiction ; Il ne postule aucun mécanisme causal ou loi physique. Bien sûr, toute explication de style humain n'est pas nécessairement correcte ; Nous sommes faillibles. Mais cela fait partie de ce que signifie penser : pour avoir raison, il doit être possible d'avoir tort. L'intelligence ne se compose pas seulement de conjectures créatives, mais aussi de critiques créatives. La pensée de style humain est basée sur les explications possibles et la

correction des erreurs, un processus qui limite progressivement les possibilités qui peuvent être envisagées rationnellement. (Comme Sherlock Holmes l'a dit au Dr Watson : « Quand vous avez éliminé l'impossible, tout ce qui reste, aussi improbable soit-il, doit être la vérité. »)

Mais ChatGPT et les programmes similaires sont, de par leur conception, illimités dans ce qu'ils peuvent « apprendre » (c'est-à-dire mémoriser) ; ils sont incapables de distinguer le possible de l'impossible. Contrairement aux humains, par exemple, qui sont dotés d'une grammaire universelle qui limite les langues que nous pouvons apprendre à celles qui ont une certaine forme d'élégance presque mathématique, ces programmes apprennent des langues humainement possibles et humainement impossibles avec la même facilité. Alors que les humains sont limités dans les types d'explications que nous pouvons conjecturer rationnellement, les systèmes d'apprentissage automatique peuvent apprendre à la fois que la terre est plate et que la terre est ronde. Ils se négocient simplement en probabilités qui changent avec le temps.

Pour cette raison, les prédictions des systèmes d'apprentissage automatique seront toujours superficielles et douteuses. Parce que ces programmes ne peuvent pas expliquer les règles de la syntaxe anglaise, par exemple, ils peuvent très bien prédire, à tort, que « John est trop têtu pour parler » signifie que John est si têtu qu'il ne parlera pas à quelqu'un ou à quelqu'un d'autre (plutôt que qu'il est trop têtu pour être raisonné). Pourquoi un programme d'apprentissage automatique prédirait-il quelque chose d'aussi étrange ? Parce qu'il pourrait faire une analogie avec le modèle qu'il a déduit de phrases telles que « Jean a mangé une pomme » et « Jean a mangé », dans lesquelles ce dernier signifie que Jean a mangé quelque chose ou autre. L'émission pourrait bien prédire que, parce que « Jean est trop têtu pour parler à Bill » est similaire à « Jean a mangé une pomme », « Jean est trop têtu pour parler à » devrait être similaire à « Jean a mangé ». Les explications correctes du langage sont compliquées et ne peuvent pas être apprises simplement en marinant dans le big data.

Paradoxalement, certains amateurs d'apprentissage automatique semblent être fiers que leurs créations puissent générer des prédictions « scientifiques » correctes (par exemple, sur le mouvement des corps physiques) sans utiliser d'explications (impliquant, par exemple, les lois du mouvement de Newton et de la gravitation universelle). Mais ce genre de prédiction, même lorsqu'elle réussit, est de la pseudoscience. Alors que les scientifiques recherchent certainement des théories qui ont un haut degré de corroboration empirique, comme l'a noté le philosophe Karl Popper, « nous ne cherchons pas des théories hautement probables mais des explications ; c'est-à-dire des théories puissantes et hautement improbables.

La théorie selon laquelle les pommes tombent sur terre parce que c'est leur lieu naturel (selon Aristote) est possible, mais elle ne fait qu'inviter d'autres questions. (Pourquoi la terre est-elle leur lieu naturel ?) La théorie selon laquelle les pommes tombent sur terre parce que la masse plie l'espace-temps (le point de vue d'Einstein) est hautement improbable, mais elle vous dit en fait pourquoi elles tombent. La véritable intelligence se manifeste dans la capacité de penser et d'exprimer des choses improbables mais perspicaces.

La véritable intelligence est également capable de pensée morale. Cela signifie contraindre la créativité autrement illimitée de nos esprits avec un ensemble de principes éthiques qui déterminent ce qui devrait et ne devrait pas être (et bien sûr soumettre ces principes eux-mêmes à la critique créative). Pour être utile, ChatGPT doit être habilité à générer des résultats d'apparence nouvelle ; Pour être acceptable pour la plupart de ses utilisateurs, il doit se tenir à l'écart des contenus moralement répréhensibles. Mais les programmeurs de ChatGPT et d'autres merveilles de l'apprentissage automatique ont lutté – et continueront de lutter – pour atteindre ce type d'équilibre.

En 2016, par exemple, le chatbot Tay de Microsoft (un précurseur de ChatGPT) a inondé Internet de contenus misogynes et racistes, après avoir été pollué par des trolls en ligne qui l'ont rempli de données d'entraînement offensantes. Comment résoudre le problème à l'avenir ? En l'absence d'une capacité à raisonner à partir de principes moraux, ChatGPT a été grossièrement limité par ses programmeurs à apporter quoi que ce soit de nouveau à des discussions controversées – c'est-à-dire importantes. Il a sacrifié la créativité pour une sorte d'amoralité.