Emergence du sens en intelligence artificielle

Paris, 25 janvier 2025

D. Defays

Introduction

L'intelligence artificielle (IA) est la discipline qui se propose de faire faire par des machines des tâches, qui, lorsqu'elles sont réalisées par des êtres humains, sont supposées nécessiter de l'intelligence. Dans cet esprit, depuis les années 50, la discipline aborde des tâches de plus en plus compliquées et les systèmes construits atteignent un niveau de plus en plus haut d'expertise. Certains résultats ont frappé les imaginations. Un programme informatique, Deep Blue, a battu aux échecs, en 1997, un champion du monde, Gary Kasparov ; Watson, une autre intelligence artificielle, a joué brillamment à Jeopardy, un jeu de culture générale bien connu aux Etats-Unis et battu en 2011 deux des experts à ce jeu. Alphago en 2017 bat le champion du monde du jeu de go, Ke Jie. Les assistants intelligents - SIRI, Alexa, l'assistant Google - sont de plus en plus présents sur les bureaux, dans les maisons. Fin 2022, un logiciel conversationnel ChatGPT secoue la planète. Les superlatifs pleuvent. Les performances sont tellement impressionnantes que bon nombre de scientifiques demandent une pause dans les développements, réclament un moratoire.

Ces résultats interpellent tous ceux qui s'intéressent à l'esprit et à son fonctionnement. Que les machines puissent réaliser des tâches qui demandent de l'intelligence, quelle que soit la définition donnée à cette aptitude, est une évidence. Mais peut-on pour cette raison les qualifier d'intelligentes ?

Imaginez que nous ayons découvert une forme de vie sur une autre planète. Il est fort probable que cette découverte nous amènerait à nous percevoir différemment, à considérer la vie sur terre comme un cas particulier d'une notion de vie plus large. N'est-ce pas ce qui est en train de se passer avec l'intelligence artificielle ? Les comportements dits intelligents ne doivent-ils pas être reconsidérés à la lumière de ce qui se fait en IA ?

Cette présentation se penche sur cette question. Elle montre comment la barrière, apparemment infranchissable, entre les signes et ce qu'ils signifient, entre les formes et le sens qu'elles véhiculent (les mots et ce qu'ils disent, par exemple), entre la syntaxe et la sémantique s'est progressivement érodée. Des premiers systèmes qui utilisaient la logique des prédicats, fin des années 50, aux grands modèles linguistiques (Large Linguistic Models ou LLM en anglais) des années 2020, l'IA a lentement pénétré l'univers du sens et des significations. En complexifiant la manière dont les informations sont traitées, elle donne l'impression qu'elle « comprend » de plus en plus ce qu'elle dessine, ce qu'elle écrit, ce qu'elle dit. Illusion, hallucination pour certains, changement de phase et irruption dans le domaine du sens - qui ne serait que de la forme sophistiquée pour d'autres.

Les tests de Turing et de la chambre chinoise

Le débat sur l'intelligence des machines ne date pas d'aujourd'hui, bien entendu. Déjà au début du XIXe siècle, Ada Lovelace (1815-1852), fille du poète romantique anglais Georges Byron, disait à propos des programmes informatiques qui devaient tourner sur un ancêtre des ordinateurs actuels - qu'ils ne pouvaient faire que ce qu'on savait déjà faire, qu'ils étaient incapables d'originalité (note G de Ada Lovelace dans "Translator's Notes to M. Menabrea's Memoir," Scientific Memoirs 3 (1843)). Pour être considérés comme ayant un esprit, ajouta-t-elle, ils devraient créer des choses.

Alan Turing, en 1950, propose une manière très différente de répondre à la question « Les machines peuvent-elles penser ? ». Plutôt qu'aborder frontalement la question de l'intelligence des machines, il suggère un jeu dit de l'imitation, qu'il transforme pour en faire un test - le fameux test de Turing. Ce test consiste essentiellement à s'entretenir avec un interlocuteur dissimulé (homme ou machine) dont on ne connait pas l'identité. A partir des réponses données aux questions qu'on lui pose, l'interrogateur doit démasquer la machine éventuelle avec laquelle il s'entretient. Ici, la capacité à penser passe par le langage, et se ramène à produire des propos qui font sens pour un interlocuteur humain. Pas question d'aller voir à l'intérieur de la machine s'il y a des cellules biologiques ou des processeurs électroniques, pas question de demander à la machine de danser sur un morceau de musique, de se faire des amis. Pas besoin non plus de faire appel à une notion d'originalité, de créativité, bien relatives souvent.

Des nombreuses objections à ce test ont suivi dans les années 80. La plus célèbre a été formulée par le philosophe américain John Searle. L'essence de sa critique (connue comme le test de la chambre chinoise) est la suivante. Je peux tenir une conversation en chinois en m'aidant de dictionnaires, de grammaires, de registres de questions possibles avec les réponses correspondantes rédigées en chinois, sans comprendre ce que je dis. Je simule parfaitement. L'interrogateur me pose une question en chinois, je consulte les documents que j'ai à ma disposition, je repère des signes, des caractères, j'opère des substitutions et je réponds… en chinois. Mon interlocuteur trouvera mes réponses raisonnables et s'imaginera que je parle chinois. C'est ce qui se passe lorsqu'une machine réussit dans le test de Turing à se faire passer pour un humain, nous dit John Searle. Que peut-on en déduire sur l'intelligence des machines ? Rien. Cet argument est encore celui qui est le plus souvent utilisé par ceux qui dénient toute forme d'intelligence aux intelligences artificielles. Personne ne pouvait sérieusement penser que les systèmes informatiques des années 60, ceux qui sont considérés maintenant comme les premiers succès de l'IA, comprenaient ce qu'ils faisaient, ce qu'ils disaient. Ils étaient du reste totalement incapables de réussir un test de Turing. Autant demander à une horloge si elle sait ce qu'est le temps qu'elle mesure ! Le programme de dames, par exemple, écrit par Arthur Samuel fin des années 50, a battu un expert américain en 1962, mais il n'avait du jeu de dames qu'une notion très limitée. Il encapsulait quelques règles, sans plus. Rien à voir avec un joueur qui sait ce que gagner veut dire, qui connait d'autres jeux similaires et perçoit bien les spécificités du jeu de dames, qui peut adapter sa stratégie si on change légèrement les règles du jeu, qui peut voir et toucher les pions, le damier…

Les robots conversationnels d'aujourd'hui et leurs performances spectaculaires amènent un certain nombre d'observateurs attentifs de l'IA à se reposer la question. Des experts comme Sejnowski, Hinton, Hofstadter, Bubeck et des nombreux auteurs d'articles scientifiques publiés depuis l'apparition des grands modèles linguistiques considèrent que les systèmes les plus récents témoignent d'un embryon de compréhension. La barrière qui semblait se dresser entre les systèmes artificiels et la compréhension/la pensée/l'accès à la signification semble s'éroder. Cette érosion a été progressive.

Les systèmes de symboles physiques

L'esprit est initialement vu comme une machine à traiter de l'information. Au début de l'histoire de l'IA – en fait baptisée et reconnue comme discipline à part entière quelques années plus tard en 1956 – la séparation entre la forme prise par l'intelligence et la signification véhiculée par ces formes (pour autant que cette distinction un peu simpliste ait un sens, nous y reviendrons) connait son expression la plus radicale. Herbert Simon écrira à propos de ce qui est considéré comme le premier programme d'intelligence artificielle « nous avons inventé un programme informatique capable de penser de manière non numérique et, de ce fait, avons résolu le vénérable problème de l'âme et du corps en expliquant comment un système composé de matière pouvait exhiber, les propriétés de l'esprit. » (Crevier, p 65).

La plupart des approches du début de l'IA, c'est-à-dire grosso modo de 1956 à 1986 (publication du livre Parallel Distributed Processing par David Rumelhart, James McClelland et le PDP research group) repose sur l'hypothèse déclarative. Si je vous demande de verbaliser vos processus mentaux, vous décrirez ce que vous percevez de vos pensées au moyen de phrases, qui articulent des mots, des concepts, des symboles. L'activité mentale est décrite comme une aptitude à manipuler ces symboles. Il n'y a pas dans ce paradigme place à autre chose que du traitement d'objets mentaux. Il n'y a pas un principe de l'intelligence comme il n'y a pas un principe de vie qui véhicule l'essence de la vie, écriront Newell et Simon. Pas de substrat « pensée » ou « signification » au-delà des symboles. Au cours des ans, cette façon radicale d'envisager les rapports entre forme et sens va se raffiner.

Cette manière de modéliser l'activité de l'esprit prend donc progressivement forme et débouche en 1976 sur une hypothèse empirique formulée de manière très explicite par Newell et Simon. Ils définissent d'abord ce qu'ils appellent un système de symboles physiques. L'adjectif « physiques » est ici essentiel. « Un système de symboles physiques », nous disent-ils, « consiste en un ensemble d'entités, appelées symboles. Ces symboles sont des structures physiques qui peuvent apparaître comme des composantes d'un autre type d'entités appelées expressions (ou structures symboliques). Une structure symbolique est donc composée de symboles particuliers physiquement reliés (ils peuvent être adjacents, par exemple)... De plus, le système contient également un ensemble de procédures qui transforment les expressions en d'autres expressions : création, modification, reproduction et destruction. Il se présente comme une machine qui produit au cours du temps une collection dynamique de structures symboliques. » Lorsqu'un système de symboles est suffisamment puissant (les auteurs énumèrent les propriétés que doit avoir le système), il

possède les caractéristiques nécessaires et suffisantes pour réaliser des conduites intelligentes.

Premiers éléments de sens

Comment ces premiers systèmes dont les limites sont apparues dans les années 80 ont-ils aborder le problème du sens? L'impression de compréhension par les agents conversationnels de l'époque (Eliza, SHRDLU, Boris …) de ce qu'ils écrivent ou de ce qu'ils disent repose sur plusieurs caractéristiques de leur architecture, que l'on pourrait appeler des « ponts » entre l'univers des signes et l'univers du sens. L'utilisation de la logique des prédicats débouche sur l'utilisation de règles d'inférence qui préserve la véracité des expressions, c'est-à-dire garantit qu'elles aient un sens. Suivre des règles syntaxiques appropriées permet donc de tenir des propos cohérents, en phase avec la réalité ; c'est le premier pont identifié. La conformité à la réalité ou à un domaine et la notion de vérité paraissent cependant dans beaucoup de situations des contraintes trop strictes. Les systèmes des années 70 vont avoir recours à d'autres artifices pour ancrer leurs performances et donner des relents de sens à leurs propos. Des correspondances sont établies entre des situations différentes, des appariements avec des environnements qui nous sont familiers sont définis et une impression de compréhension émerge ; c'est le deuxième pont. De manière plus générale, lorsque l'ambition va au-delà du discours sur un domaine ou sur un micro-monde donné, pour avoir une signification, les productions des systèmes artificiels doivent se baser sur de nombreuses connaissances structurées et articulées. Ces systèmes comprennent ce qu'ils lisent ou ce qu'ils disent dans la mesure où les récits ou les réponses données sont le fruit d'un traitement sophistiqué qui s'appuie sur des structures existantes. C'est un troisième pont entre les signes et leur signification. Les systèmes construits à partir de ces principes relèvent des approches dites symboliques. Le sens de ce qu'ils produisent est lié à la justesse des inférences que véhiculent des entités symboliques, des correspondances établies à un niveau symbolique, des traitements des mots, de morceaux de phrases.

Les premiers neurones artificiels

Des modèles totalement différents ont été proposés dans les années 1940. Warren McCulloch et Walter Pitts, deux chercheurs américains, avaient cherché à reproduire avec des réseaux des opérations logiques élémentaires. Pour eux, l'événement atomique de l'activité mentale est l'émission par une cellule nerveuse d'une impulsion. A partir de ce constat, ils cherchent à imaginer des réseaux mathématiques simples où des nœuds échangent des signaux. Une vue très simplifiée du fonctionnement du cerveau toujours à la base des modèles d'aujourd'hui. Ils réussissent ainsi à faire réaliser par des réseaux des opérations logiques et montrent implicitement que le cerveau avec son mode de fonctionnement peut effectuer des calculs. Le mot clé n'est pas le symbole, mais plutôt le signal ; le souvenir devient la réactivation de la trace d'un signal, pas nécessairement le rappel en mémoire d'éléments symboliques. Les diffusions d'activations président au fonctionnement de ces réseaux. En 1949, un physiologiste, Donald Hebb proposa une méthode – connue maintenant sous le nom d'apprentissage hebbien - qui permet à ces réseaux d'apprendre. Elle repose sur un principe simple : l'activation répétée d'un neurone

par un autre mène à une amélioration de la conductivité de leur connexion synaptique. Ces idées vont conduire au développement des modèles connexionnistes. De nouveau, pas de place dans ce modèle pour un principe d'intelligence, une flamme qui ferait émerger la compréhension, ou une couche « signification » qui enroberait les signaux véhiculés dans les réseaux. Elle est implicitement considérée comme une forme d'émergence de l'activité neuronale… artificielle.

Ces premiers travaux, qui répétons-le, se révèlent être à la base de modèles beaucoup plus complexes comme les LLMs, n'ont pas connu un succès retentissant avant la fin des années 80. Un livre devenu célèbre, écrit par Marvin Minsky et Seymour Papert et intitulé « Perceptrons: an introduction to computational geometry » a montré clairement les limites de ces premiers types de réseau – les perceptrons - et freiné considérablement les travaux sur ces architectures. Le paragraphe qui suit explique le fonctionnement d'un perceptron.

Le perceptron

Rosenblatt en 1958 propose une structure de réseau très simple. Elle est illustrée ci-dessous.

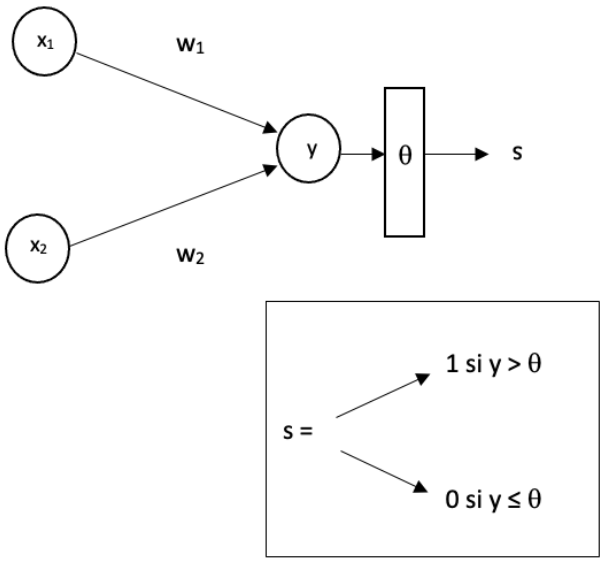

Figure 1. Un perceptron

Ce dispositif qui s'inspirait du fonctionnement du cerveau cherche à formaliser la manière dont on associe une réponse s (qui vaut ici 0 ou 1) à un stimulus (ici deux nombres x_1 et x_2). Il peut fonctionner en présence d'informations dégradées. Dans ce réseau, le signal d'entrée est représenté par deux nombres (le vecteur (x_1, x_2)). Ces deux nombres traversent le réseau en passant par des arêtes auxquelles on a associé des poids (censés représenter la conductivité synaptique). Ils sont notés w_1 et w_2. Le signal (x_1,x_2) est modulé par ces poids et arrive sous une forme recomposée y au nœud de sortie. Plus mathématiquement, la recomposition se fait comme suit : $y = w_1 x_1 + w_2 x_2$. Le signal en sortie du perceptron, noté s dans la figure, est une transformation de y. Il est simplement comparé à un seuil θ.

Si y est supérieur à θ, le perceptron répond 1 (s=1), sinon, il répond 0 (s=0). Imaginez cette réponse binaire comme un « oui » ou un « non ». Le perceptron dit s'il a ou s'il n'a pas reconnu le motif entré sous la forme (x_1,x_2). En réalité, le perceptron peut admettre plus de deux nœuds d'entrée, ce qui permet par exemple de considérer comme signal d'entrée tous les pixels d'une image en noir et blanc et en sortie d'exiger une réponse oui ou non suivant que cette image représente ou pas un visage donné. Les poids et le seuil sont appris par le perceptron à travers la présentation d'exemples. Ces réseaux ont été critiqués, comme je l'ai signalé plus haut, pour leur simplicité excessive qui en limitait la puissance. Certaines formes ne pouvaient pas être reconnues. On les a donc complexifiés en ajoutant des nœuds et des couches.

Les LLMs

Complexification des modèles

Les réseaux profonds actuels qui permettent d'interagir en langage naturel avec des interlocuteurs, de tenir des propos cohérents, de faire des analogies, de raisonner, de traduire, de programmer, de faire des résumés, d'écrire des histoires, de faire des recommandations pour citer quelques tâches dont ils sont capables, sont construits à partir de structures atomiques similaires aux perceptrons. Mais leur ordre de complexité est beaucoup plus élevé : les perceptrons prennent une forme plus élaborée. Ils comportent des milliers de canaux d'entrée et de sortie, opèrent des traitements du signal qui ne consistent plus à une simple comparaison à un seuil, mais à la transformation du signal « recomposé » par des fonctions simples dites fonctions d'activation et sont organisés en couches hiérarchiques, d'où l'appellation de réseaux profonds (il peut y en avoir plus de cent). Les réseaux les plus récents incorporent en plus des mécanismes dits d'attention qui permettent d'établir des liens entre différents mots, différentes parties d'un discours. L'architecture de ces mécanismes autorise un traitement en parallèle de certaines opérations ce qui accélère considérablement les apprentissages et permet de travailler à partir d'une quantité gigantesque de données.

L'enchâssement

Dans ces réseaux, le traitement d'un texte commence par la construction d'une représentation vectorielle des mots (ou de parties de mots) qui constituent le texte. Les mots sont plongés dans des espaces à plusieurs dimensions (il peut y en avoir plus d'un millier). Cette représentation permet de donner une forme géométrique aux relations sémantiques qui existent entre les différents mots. Ainsi, les mots qui ont un sens proche se retrouvent dans des mêmes zones de l'espace dans lequel on les plonge, les relations entre ces mots du type « opposé », « est le partenaire de », « a pour capitale » , « donne à l'imparfait » deviennent des relations de parallélisme, des structures géométriques particulières. Ce plongement sémantique est appelé un enchâssement et est fort utile en traitement du langage.

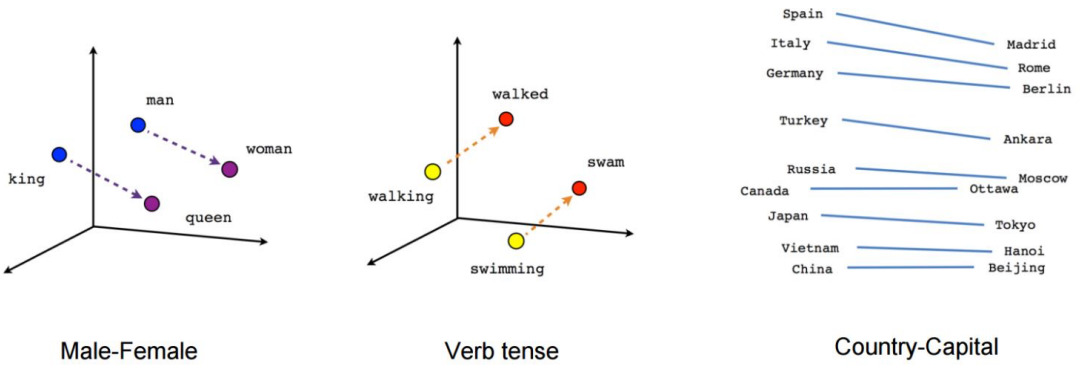

Figure 2. Enchâssement (source : https://towardsdatascience.com/deep-learning-4-embedding-layers-f9a02d55ac12)

L'idée à la base de la construction d'un enchâssement a été proposée par Yoshua Bengio. Elle consiste à entraîner un réseau neuronal à prédire dans une phrase un mot manquant. Ainsi, dans la phrase « Un chien court dans la pièce », on va cacher successivement les mots un, chien, court, dans, la et pièce et chaque fois demander à la machine de trouver le mot caché à partir de ce qu'on appelle le contexte (ici les autres mots de la phrase). Et ceci se fait habituellement à partir de textes assez longs bien entendu. Ils peuvent contenir des millions de mots avec des vocabulaires de dizaines de milliers de mots différents (en incluant majuscules, ponctuations...).

Les plongements sémantiques permettent de représenter le sens général des mots. Mais ceci n'est pas encore suffisant. Dans un texte, dans une phrase, ce sens doit se moduler en fonction du contexte. Le mot « tour » lorsqu'il est suivi de « Eifel » prend une forme plus précise qui requiert une adaptation de sa représentation « sémantique ». Ces transformations de représentation sont réalisées par des transformeurs.

Le transformeur

Le transformeur est un type de réseau qui permet d'établir des liens entre différents éléments d'une séquence (par exemple une phrase) au moyen d'un mécanisme dit d'attention. Il a été introduit en 2017 et est à la base des développements spectaculaires observés ces dernières années. Il opérationnalise l'utilisation du contexte. Le système analyse directement les dépendances entre les mots d'une phrase, d'un paragraphe, d'un texte pour pouvoir les utiliser dans son traitement. En fait, les réseaux utilisent plusieurs mécanismes d'attention qui analysent différentes dimensions des mots, après leur enchâssement. Leur démultiplication permet de définir simultanément plusieurs contextes, d'appréhender différents types de dépendance : liens temporels, rôles des mots dans la phrase, accords en fonction du genre, etc. Il peut y avoir dans certains transformeurs une centaine de mécanismes d'attention. Le traitement est hiérarchique de manière à pouvoir travailler sur des représentations de plus en plus globales des textes donnés en entrée au système.

L'émergence

Une des particularités des LLMs est la gamme des performances qu'ils autorisent. Jusqu'en 2022, la plupart des systèmes étaient « monotâches ». Ils se montraient efficaces dans des domaines particuliers : Eliza simulait le comportement d'un thérapeute non directif, SHRDLU pouvait débattre de la position et de la nature de blocs dans un microcosme, les systèmes experts de type MYCIN permettaient de poser des diagnostics dans des cas très particuliers, les systèmes de reconnaissance de forme classaient, identifiaient des objets, reconnaissaient des visages mais ne pouvaient pas en parler, ni raisonner sur ces objets, les réseaux neuronaux utilisés pour traduire des textes ne pouvaient rien faire d'autre. Comme déjà mentionné, les modèles de langage des années 2020 sont capables de soutenir une conversation, de raisonner, de générer du contenu, de faire des analogies, de traduire des textes, d'écrire des programmes, d'imiter des styles et cette liste n'arrête pas de s'allonger. Comment expliquer l'apparition de cette polyvalence ?

Jason Wei et al., dans un article publié en 2022 et intitulé « Emergent Abilities of Large Language Models", ont proposé le concept d'émergence pour répondre à cette question. Ils ont étudié les propriétés dites émergentes des LLMs. Ils montrent que lorsque la complexité augmente – elle peut être mesurée par la taille des ensembles d'apprentissage, la complexité des calculs en FLOPs (nombre d'opérations en virgule flottante) ou le nombre de paramètres - les performances des systèmes, avec un apprentissage complémentaire léger à partir de quelques exemples (typiquement entre 10 et 100), augmentent brutalement à partir d'un certain seuil. L'amélioration des performances se traduit par une gamme beaucoup plus large des tâches que les systèmes peuvent accomplir. Ils qualifient ce surgissement d'un nouveau type de comportement d'émergence.

Tout se passe donc, comme si ces réseaux profonds avaient besoin d'une certaine assise, d'une base suffisamment élaborée pour avoir accès à des performances d'un autre ordre. A force d'empiler les niveaux, de multiplier les paramètres, les systèmes acquièrent, grâce à quelques exemples qu'on leur donne des performances attendues, une disposition générale qui leur permet d'exécuter des tâches pour lesquelles ils n'ont pas été programmés.

Un débat animé sur l'accès au sens des nouveaux agents conversationnels

Ils peuvent aussi grossièrement se tromper comme l'ont montré de nombreux experts. Ils inventent des significations à des mots qui n'existent pas (un bivol est un vol avec une escale et une trifouille une fourche à bêcher), font des erreurs de raisonnement, génèrent des images absurdes. Ce mélange de performances impressionnantes qui ne paraissent explicables qu'en gratifiant le système d'un minimum d'intelligence, et d'erreurs grossières révélatrices d'un manque totale de compréhension suscite des commentaires divers.
« Fluent manipulation of mere words and phrases does not, on its own, imply the presence of concepts. " a écrit Mélanie Mitchell, une experte en IA très concernée par les liens entre l'IA et le sens. Il est vrai que ce que nous montrent les LLMs est essentiellement une certaine aptitude à manipuler la langue, c'est-à-dire des formes verbales. La notion de concept va bien au-delà. Dans le mot printemps, il y a des odeurs de sève, des sensations de chaleur, du

sang frais, de l'attente de jours ensoleillés qui vont au-delà des mots et des sons utilisés pour en parler. Le concept, souvent présenté ou défini comme une idée de quelque chose, est au cœur de la notion de pensée. Il présuppose une certaine forme d'accès au sens. Mais comment peut-on affirmer que nous avons accès au sens, si ce n'est à travers notre usage sensé des mots ? Lamartine écrivait déjà dans « La providence à l'homme » : « Tu pensas; la parole acheva ta pensée ». Et chatGPT utilise les mots de manière "relativement" sensée. Ceci amène D. Hofstadter à répondre en ces termes à M. Mitchell : « Concepts are what lies behind words (and phrases) that are imbued with meaning, and since GPT-4's words do indeed have meanings (albeit only to a limited degree, so far), then to that extent, GPT-4 (and similar LLM's) do have concepts (to a limited degree, at least)." S. Bubeck argumente de manière plus empirique, mais dans le même sens. Prétendre que maîtrise de la langue et pensée sont des choses différentes, n'est pas totalement satisfaisant, pour lui.

Qu'entend-t-on par compréhension ?

Les notions de sens, de compréhension dont débattent les experts appartiennent à une galaxie de concepts qui tournent autour de la pensée, des idées, de l'intelligence, de la cognition, de la signification, de la sémantique… La question « les machines peuvent-elles comprendre ?» est étroitement liée aux questions suivantes : les machines peuvent-elles penser, peut-on passer d'une analyse syntaxique à une analyse sémantique, la manipulation de signes peut-elle conduire à une appréhension du contenu véhiculé par ces formes, le sens peut-il émerger du traitement complexe de structures de données… Et la galaxie du sens paraît assez proche de celle qui tourne autour de la conscience, car il est difficile de comprendre sans saisir ce que veut dire « Je ». Une approche plus analytique de cette galaxie paraît ainsi indispensable.

Différentes nomenclatures ont, par exemple, été proposées pour rendre compte de la diversité des types de compréhension.

David Chalmers dans une conférence donnée en 2023 sur « The barrier of meaning » propose de distinguer 3 types de compréhension. Le premier type correspond à ce « Aha », cette expérience, ce ressenti lorsqu'on vous explique un phénomène, un fonctionnement, lorsqu'on résout une énigme. C'est ce qu'on pourrait appeler sa dimension phénoménologique (p-knowledge). Un deuxième niveau requiert l'utilisation de connaissances. On comprend que si on débranche la batterie d'une voiture, les phares ne fonctionneront plus. Mais avant de comprendre, il faut savoir : si une balle roule et s'arrête derrière une boîte, nous dit Chalmers, pas besoin d'une théorie pour savoir qu'elle est derrière la boîte. Comprendre ce qui se passe, par contre, nécessite le recours à des principes abstraits comme la permanence des objets, la gravité, l'inertie, le frottement. Comprendre, ici, c'est savoir pourquoi, c'est justifier. C'est ce qu'on appelle quelquefois la compréhension basée sur les connaissances (knowledge based understanding ou e-knowledge). Importante pour la pensée scientifique, les relations sociales, mais moins pour la compréhension d'une langue probablement. Enfin, un dernier niveau couvre notre capacité à réagir de manière appropriée à une information, à l'utiliser. On peut répondre à une question, on peut manipuler un objet, on peut raisonner, faire des inférences – une dimension évoquée plus haut avec le voyage à Londres (u-knowldge)

D’autres distinctions ont été proposées, la compréhension par l’expérience, ou la compréhension référentielle qui permet d’établir des liens, de mettre en correspondance, de faire des analogies.

Il y a donc lieu, lorsqu’on s’interroge sur la faculté de comprendre des agents conversationnels, de préciser à quel type de compréhension on fait référence. Des batteries de test ont du reste été proposées pour évaluer le niveau d’ « intelligence » des machines : le SQuAD, le GLUE, Winograd schema challenge. Lorsqu’on les soumet à ces tests standardisés, les logiciels révèlent leurs limites. ChatGPT, par exemple, éprouve des difficultés à raisonner dans certaines situations, particulièrement quand il s’agit de traiter des dispositions dans le temps ou dans l’espace, à faire des inférences logiques. On retrouve aussi certaines erreurs qui sont communes chez les humains comme la conjunction fallacy, la tendance à proposer de prime abord une réponse spontanée avant de raisonner suite à une indication de l’utilisateur, l’incapacité à appliquer des règles trop compliquées. La dimension phénoménologique est aussi totalement absente : les LLMs se comportent comme des non-voyants qui nous parlent de couleurs. Ils en ont, actuellement, une certaine compréhension, mais elle ne passe pas par des sensations.

Les impacts

Les progrès observés ces dernières années auront un impact évident sur l’économie, la société et l’humanité en général.

- L’économie va voir apparaître de nouveaux métiers, disparaître d’autres et beaucoup se transformer progressivement. Des pressions sur les composants électroniques, les ressources naturelles sont déjà observables début des années 2020.

- La banalisation des infox risque aussi de transformer nos sociétés et leur mode de fonctionnement. La confiance dans les informations reçues qui constitue le système sanguin de nos démocraties sera probablement affectée. Le système éducatif ne pourra pas ignorer l’existence de ces nouvelles sources d’information et de création. La justice devra intégrer des dispositions dans son arsenal pour tenir compte de l’apparition de nouveaux types d’agents, pour délimiter des responsabilités, pour protéger les droits d’auteur, la vie privée. Les citoyens devront être protégés contre les utilisations malveillantes de l’IA. Et les risques d’une IA qui puisse asservir l’esprit humain paraissent peut-être de moins en moins relever de la science-fiction. Les gouvernements commencent à se mobiliser pour cadrer les travaux en IA, des sommets internationaux proposent des recommandations. Le mode de l’enseignement se prépare à intégrer ces nouveaux outils, les juristes reconsidèrent les droits d’auteur, la protection de la vie privée, le monde culturel découvre les nouvelles possibilités offertes. Des utilisations malveillantes ou des dérapages peuvent représenter des menaces pour la société et des dispositifs commencent à être mis en place pour y faire face.

- Enfin, après l’impact économique et le social, les effets attendus sur les individus, ce qui fait leur humanité, leurs valeurs devront être évalués. Banaliser la création artistique, la créativité humaine – jusqu’aujourd’hui considérées comme des

monopoles de l'esprit humain - ne se fera sûrement pas sans coût, sans effet sur la manière dont nous nous percevons, par exemple. Quel traitement devront nous réserver à ces machines dont les performances devraient encore progresser dans les années à venir ? Faudra-t-il inventer de nouveaux types de solidarité, de citoyennet?

Les menaces, les dangers, les impacts considérables attendus sur nos modes de vie ne doivent cependant pas faire perdre de vue les gains considérables que les intelligences artificielles vont apporter. Quasi tous les métiers seront touchés et seront probablement améliorés, facilités et élargis.

Les premières leçons à tirer dans l'étude de l'esprit artificiel

Les psychologues et tous ceux intéressés par le mental ne peuvent pas être indifférents aux progrès observés sur les agents conversationnels des années 2020. Peuvent-ils nous apprendre quelque chose sur la pensée, la cognition, le fonctionnement du cerveau ? Quelles sont les caractéristiques essentielles de ces systèmes artificiels qui pourraient expliquer les performances impressionnantes observées ?

Le premier constat est lié à la nature de la tâche sur laquelle les réseaux sont entraînés : la prévision de morceaux manquants dans des phrases. Le rôle majeur, dans l'évolution et dans la survie de l'espèce, de la capacité à anticiper, à prévoir, ne doit plus être établi. Il est piquant de noter que la simple aptitude à compléter une phrase permet l'émergence d'un dispositif neuronal à même d'aborder des tâches diverses. L'établissement de liens qui est au cœur de la disposition à anticiper paraît tisser une toile sur laquelle peuvent se greffer différentes fonctions cognitives.

L'architecture générale des LLMs mérite aussi d'être soulignée. Les systèmes sont essentiellement des systèmes hiérarchiques matérialisés par différentes couches dans des réseaux neuronaux profonds. Mais le rôle des différents niveaux reste à établir. L'esprit artificiel – pour autant que cette expression audacieuse ait un sens – apparaît comme un mille-feuille. Certains niveaux peuvent se décrypter avec nos concepts actuels, s'habiller de mots qui correspondent à l'intuition que nous avons de notre manière de penser, mais d'autres sont beaucoup plus mystérieux. Ils obéissent peut-être à des lois, des algèbres, une géométrie qui restent à décrypter. Les performances observées des LLMs est incontestablement le fruit d'une digestion progressive à travers plus de 100 couches (pour GPT-4) des données d'entrée. Une longue sédimentation. Une vue répandue de l'esprit invite à considérer deux niveaux d'analyse : le niveau conscient qui manipule des informations symboliques dont le fonctionnement a été simulé par les premiers systèmes développés en IA et un niveau sous-cognitif ou inconscient qui était plutôt le domaine de prédilection des approches neuronales. La nécessité d'un lien entre les deux a du reste été soulignée par de nombreux experts (voir Le Cun, 2019, par exemple). Les modules neuronaux procèdent de manière « bottom-up », sur un mode réactif où un stimulus appelle une réponse ; les systèmes symboliques opèrent par délibération sur des modèles du monde. Une articulation entre les deux paraît nécessaire. Peut-on affirmer que tout ce qui se passe chez chatGPT relève de la seule logique bottom-up ? L'apprentissage a probablement permis au système d'identifier des régularités qui lui permettent de percevoir les nouveaux stimuli d'une manière partiellement top-down. La délibération, la planification, les tâches

intellectuelles de haut niveau doivent-elles passer par un traitement symbolique des informations ? Les performances des LLMs semblent répondre non à cette deuxième question, ou du moins pas toutes, ou pas entièrement. La compréhension du fonctionnement de l'esprit requiert peut-être de se pencher sur ces strates, si elles existent, qui constituent la fabrique de la pensée.

L'architecture générale des transformeurs est très peu fonctionnelle. Il n'y a pas un module chargé de faire des regroupements (chunking) et de concentrer l'information, un module chargé de faire de la prévision (c'est la fonction de l'ensemble du système dans sa version initiale), un module chargé de faire des dérivations, des inférences, des comparaisons. Or ces aptitudes émergent dans les systèmes. Les premiers systèmes développés en IA étaient beaucoup plus structurés avec des modules de traduction de phrases en instructions (Shrdlu), des modules d'analyse grammaticale des phrases (Boris), des agents spécifiques comme dans la société de l'esprit proposée par Marvin Minsky. Différents types d'apprentissage étaient distingués et mis en œuvre dans des systèmes différents comme l'apprentissage par généralisation, par discrimination, par explication (voir Defays, 1987, par exemple). Il semble donc préférable, lorsqu'on construit des systèmes artificiels, plutôt que de développer des modules fonctionnels spécifiques, de laisser le système s'organiser lui-même.

Les opérations de base réalisées par les réseaux artificiels sont très simples : tout se ramène à des additions, des multiplications, des concaténations… Surprenant que des opérations mentales sophistiquées qui n'ont a priori aucun caractère arithmétique, comme l'écriture d'un poème, l'explication d'une analogie, ou l'écriture de notes humoristiques (voir chapitre suivant) se ramènent à du calcul élémentaire. Des règles simples et beaucoup de calcul peuvent donner l'impression de processus très complexes.

Les systèmes satellites, comme Dall.E qui permet à partir d'une invite de générer des images, les logiciels qui permettent de faire de la traduction, du codage informatique, de résumer des textes, se construisent sur le système de base, entraîné avec des tâches de prédiction. Il est vrai que des apprentissages et des ajustements complémentaires sont nécessaires pour arriver à ces performances, mais ils sont relativement légers. «In general, it's interesting how little "poking" the "originally trained" network seems to need to get it to usefully go in particular directions … » (Wolfram, 2023). Tout se passe comme si le réseau initial, avant apprentissages spécifiques, encapsulait déjà une forme de raisonnement humain qui lui permet de généraliser et de servir à d'autres tâches.

La reproduction de certains types d'erreurs observées chez les sujets humains comme la conjunction fallacy, la tendance à proposer de prime abord une réponse spontanée avant de raisonner suite à une indication de l'utilisateur, l'incapacité à appliquer des règles trop compliquées, l'inconfort avec certaines règles de logique est a priori assez surprenante. On pourrait s'attendre d'un système qui tourne sur une prodigieuse machine à calculer (l'ordinateur) avec des algorithmes qui utilisent des opérations mathématiques, qu'il soit à l'aise avec la logique, les règles alambiquées, les calculs de probabilité. Si on veut voir dans chatGPT une simulation de la pensée humaine, c'est par contre rassurant. Les processus à l'œuvre lorsque chatGPT est confronté à une tâche ne sont de toute évidence pas rationnels et peuvent poser des problèmes de crédibilité dans les réponses du système.

Comme le souligne Stephen Wolfram, indépendamment des similarités ou des dissimilarités entre chatGPT et l'esprit humain, les travaux sur les LLMs montrent qu'une architecture du type de celle de chatGPT, avec son mode d'apprentissage, peut simuler ce que fait le cerveau lorsqu'il nous permet d'utiliser le langage. Pour raconter des histoires, il ne faut pas obligatoirement un cerveau humain.

L'étude de l'intelligence prend donc une autre tournure avec les développements de l'IA. Il y a probablement différents types d'intelligence, différents types de cognition. ChatGPT et ses collègues nous offrent des opportunités de reconsidérer ce que l'intelligence est vraiment.

Références

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Tat Lee, Y., , Y., Lundberg, S., Nori, H., Palangi, H., Tulio Ribeiro, M., Zhang, Y. (2023), Sparks of Artificial General Intelligence: Early experiments with GPT-4, https://arxiv.org/abs/2303.12712

Defays, D. (1987), *L'esprit en friche. Les foisonnements de l'intelligence artificielle*, Mardaga

Dyer, M. G. (1983), *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension*, MIT Press, Cambridge, MA.

Hofstadter, D. and the fluid analogies research group (1995), *Fluid concepts and Creative Analogies, Computer Models of the Fundamental Mechanisms of Thought*, BasicBooks.

Hofstadter, D. (2023), Is there an "I" in AI? https://berryvilleiml.com/wp-content/uploads/Is-there-an-"I"-in-AI-.pdf

Jamil, U. (2023), Attention is all you need (Transformer) - Model explanation (including math), Inference and Training, https://www.youtube.com/watch?v=hjesn5pCEYc

Kolpakov, N. (2023), How good is ChatGPT on QA tasks? A hands-on comparison using ChatGPT and fine-tuned encoder-based models on QA tasks, https://artificialcorner.com/how-good-is-chatgpt-on-qa-tasks-953a10236ec7

Le Cun, Y. (2019), *Quand la machine apprend. La révolution des neurones artificiels et de l'apprentissage profond*, Odile Jacob

Mitchell, M. (2021), *Intelligence artificielle, Triomphes et Déceptions*, Dunod

Newel, A. & Simon, H. (1976), Computer Science as Empirical Inquiry: symbols and search? *Communications of the ACM, vol 19, issue 3*, https://dl.acm.org/doi/10.1145/360018.360022

Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P. (2016), SQuAD: 100,000+ Questions for Machine Comprehension of Text, arXiv.1606.05250v3, https://arxiv.org/pdf/1606.05250.pdf

Ryan, M. (2019) Ada Lovelace, her Objection, Turing Tests and Universal Computing https://medium.com/swlh/ada-lovelace-her-objection-e189717bd262

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I. (2017) Attention Is All You Need, https://arxiv.org/abs/1706.03762v7

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Wolfram, S. (2023), "What Is ChatGPT Doing ... and Why Does It Work", Stephen Wolfram Writings. writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work.