“ ChatGPT m’a dit que… ”: l’illusion de la discussion avec l’IA nous mène à l’erreur

Frédéric Alexandre

« ChatGPT m'a dit que… » : l'illusion de la discussion avec l'IA nous mène à l'erreur

Frédéric ALEXANDRE, Centre Inria de l'Université de Bordeaux

Le point fort de ChatGPT, c'est de nous donner l'impression de discuter avec un autre être, qui penserait de la même manière que nous. Mais ces apparences nous induisent en erreur, car la conception de l'IA induit un comportement très différent du nôtre. Là où nous croyons lire un discours raisonné, il ne s'agit que d'une production très intuitive.

Certains risques liés à l'usage de l'IA générative et en particulier des agents conversationnels (comme ChatGPT) commencent à être bien décrits et assimilés par les utilisateurs. C'est le cas pour ce qui concerne leurs coûts, financier et environnemental, déraisonnables, le besoin de protection des données (comme évoqué dans cette publication de la CNIL : https://www.cnil.fr/fr/intelligence-artificielle-la-cnil-poursuit-ses-travaux) ou le rôle croissant de ces outils comme porte d'entrée sur Internet (souligné dans le rapport annuel de l'ARCEP : https://www.arcep.fr/actualites/actualites-et-communiques/detail/n/numerique-tome-3-ra-2024-040724.html). Cependant un autre phénomène, probablement plus diffus et plus insidieux, est encore mal décrit et mal pris en compte : De par leur fonction d'agent conversationnel, ces outils prennent de facto un rôle d'agent cognitif dans leurs interactions avec les utilisateurs humains et c'est donc vers les Sciences Cognitives qu'il faut se tourner pour tenter de mieux comprendre leur nature, les interactions qui vont en résulter et les risques associés.

Quel est le profil cognitif d'un agent conversationnel ?

La cognition humaine se caractérise par une dualité bien documentée. D'une part, nous avons la capacité d'aborder un problème en considérant explicitement ses caractéristiques, avec délibération et raisonnement prospectif (qu'est-ce que se passerait si…). Ce traitement est lent et représente un coût cognitif important mais cela permet de s'adapter de façon flexible à une nouvelle situation et d'expliciter sa démarche. D'autre part, si la même situation se reproduit régulièrement, nous allons pouvoir automatiser notre prise de décision et réagir de façon implicite à cette situation sans en traiter tous les détails, à un coût cognitif bien moindre et avec un traitement plus rapide. Cette automatisation n'est possible qu'après de nombreux essais et elle se caractérise donc par un apprentissage très lent. Si cette décision, plus inflexible et moins consciente, aboutit à une erreur, il faudra « reprendre les commandes », c'est-à-dire revenir à l'analyse explicite de la situation pour comprendre ce qui a changé et s'y adapter. Cette dualité se retrouve dans la description de deux modes de pensées (système 1 et système 2 : https://fr.wikipedia.org/wiki/Syst%C3%A8me_1_/_Syst%C3%A8me_2_:_Les_deux_vitesses_de_la_pens%C3%A9e) ou encore en neurosciences cognitives, par la distinction entre mémoires implicite et explicite (https://fr.wikipedia.org/wiki/M%C3%A9moire_%C3%A0_long_terme).

Il convient maintenant de souligner que les fonctions cognitives que les IA génératives sont supposées prendre en charge (comme le langage, le raisonnement ou l'imagination) relèvent chez les humains de la mémoire explicite et du système 2. Elles correspondent à ce que l'on appelle des fonctions exécutives, qui supposent en général un contrôle cognitif explicite du

comportement. On parle aussi de métacognition. Mais quand on regarde maintenant comment ChatGPT a été construit pour réaliser ces fonctions cognitives supérieures (voir par exemple https://theconversation.com/comment-fonctionne-chatgpt-decrypter-son-nom-pour-comprendre-les-modeles-de-langage-206788), on peut se rendre compte que, malgré certains principes astucieux comme son mécanisme attentionnel, sa structure repose sur un réseau profond entrainé par l'algorithme de rétropropagation, qui est fondamentalement de type implicite (système 1). De ceci découlent des différences fondamentales entre la façon dont l'être humain doit implanter ces fonctions cognitives explicites couteuses, pour pouvoir travailler en mode système 2, alors que les modèles artificiels qui restent en mode système 1 peuvent continuer à privilégier un traitement implicite par calcul massif.

Trois limitations fondamentales pour l'être humain

La façon dont le cerveau humain réalise ces traitements explicites de type système 2 découle de trois limitations fondamentales. Premièrement, nous avons une limitation en temps de calcul : L'être humain doit parfois s'adapter rapidement à des situations nouvelles (en particulier si elles sont potentiellement dangereuses). Pour cela, nous avons développé des capacités d'apprentissage rapide par lesquelles un ou deux exemples nous suffisent à nous adapter à une nouvelle situation ou à retenir certaines de ses caractéristiques. Ce n'est pas le cas de ChatGPT qui repose sur un apprentissage massif nécessitant des milliards de mots et pouvant durer des semaines. Il n'a pas cette pression du temps qu'ont les êtres vivants dans un monde changeant et parfois dangereux.

Deuxièmement, nous avons une limitation en puissance de calcul. Nous vivons depuis des milliers d'années avec à peu près le même cerveau, dans un monde qui est devenu de plus en plus complexe. Et ce cerveau à des contraintes énergétiques majeures qui limitent drastiquement notre mémoire de travail ou la fréquence de fonctionnement de nos neurones alors que les ordinateurs modernes, toujours plus efficaces et rapides, enchainent les records de puissance de calcul et de capacité de stockage. Le moyen que notre cognition a trouvé pour faire face à ce problème est d'apprendre à décomposer les problèmes complexes en sous-problèmes plus simples à résoudre, ce qui nous a amené à organiser notre comportement dans le temps et nos connaissances en degrés d'abstraction. ChatGPT n'a pas ce type de contrainte et peut aborder directement des problèmes qui dépassent notre entendement (par exemple faire la synthèse de milliers de textes).

Troisièmement, nous avons des problèmes de communication. Alors qu'avec des techniques comme les modèles de fondation ou l'apprentissage par transfert, un réseau de neurones peut simplement transférer ce qu'il a appris à un autre réseau (par exemple en partageant ses poids), nous n'avons pas la possibilité de nous instruire auprès d'un autre être humain en nous connectant directement à son cerveau. A la place, nous avons développé des « stratégies » comme le langage, l'éducation ou la culture qui nous obligent à apprendre à nous expliquer et à communiquer.

On pourrait bien sûr discuter plus avant pour savoir si ces limitations (qui permettent aussi d'expliquer certains de nos biais cognitifs) en sont vraiment ou plutôt si elles sont parmi les meilleurs atouts de notre cognition (apprentissage rapide, organisation structurée des connaissances et explication de nos acquis). Mais il est cependant important de constater

qu'elles rendent notre façon d'aborder les problèmes fondamentalement différente des traitements implicites et massifs réalisés par ChatGPT. Ceci est à la source d'une ambigüité majeure : alors que ChatGPT est un système complexe comportant des milliards de paramètres entrainés à partir de milliers de milliards de données pour apprendre à prédire le prochain mot le plus probable, il cache cette complexité derrière sa fonction d'agent conversationnel qui nous parle et, par simple projection, comme on le fait quand nous parlons avec nos semblables, nous développons cette illusion d'interagir avec un agent intelligent, qui pense, ressent et comprend comme nous. Ce qui fait que, pour comprendre ces systèmes dont la complexité nous dépasse, nous en sommes à demander à des psychologues d'interpréter leurs raisonnements et de découvrir leurs biais (https://theconversation.com/quand-les-psychologues-decryptent-le-raisonnement-des-intelligences-artificielles-228528).

Quand les sciences humaines et sociales nous permettent de comprendre nos relations avec ChapGPT

Une étude publiée par le Boston Consulting Group auprès de professionnels utilisant ChatGPT (https://www.bcg.com/press/21september2023-ia-generative-quel-impact-sur-la-productivite-au-travail) illustre bien cette dualité. D'une part, elle rapporte que ces professionnels sont plus productifs et plus objectifs (ChatGPT leur permet de traiter rapidement des problèmes complexes et évite certains de nos biais) mais d'autre part, elle souligne des problèmes dus à cette illusion que nous avons que ChatGPT nous est semblable. Les performances chutent drastiquement quand ChatGPT fait des erreurs ou quand le problème dépasse nos compétences ? En effet, en travail collaboratif, nous faisons confiance aux personnes plus expertes que nous. Les productions de ChatGPT tendent à homogénéiser les résultats produits ? Pas surprenant pour un traitement statistique massif, comparé aux vues originales et parfois même divergentes que nous pouvons avoir. Nous sommes satisfaits par une proposition de ChatGPT qui n'est qu'une prédiction étayée par aucun raisonnement ? Là aussi, les sciences cognitives indiquent qu'en dehors de nos domaines d'expertise, des prédictions assénées avec assurance nous donnent l'illusion de comprendre et que nous faisons confiance à des explications réductrices.

Ainsi, ces outils numériques, qui peuvent nous aider à produire plus, peuvent aussi nous entrainer, si on les considère comme des partenaires cognitifs, à moins comprendre et moins pouvoir expliquer ce que l'on fait, à commettre des erreurs, à ne pas nous rendre compte que notre espace de recherche s'est réduit et appauvri. Plus encore, si ces usages se généralisent, on pourrait aller vers une homogénéisation de notre culture, due à des outils créés par des entreprises privées dont les algorithmes aussi bien que les valeurs sont souvent obscurs. Cela pourrait progressivement éliminer la diversité de nos expériences et notre subjectivité alors que les sciences cognitives ont bien montré que des groupes avec de la diversité (de genre, d'ethnie, de culture) sont meilleurs pour résoudre des problèmes et être créatifs.

Il devient donc crucial d'insister sur l'éducation aux outils numériques des jeunes générations et la formation de certains acteurs clés (formateurs, médias, entreprises, politiques) pour les sensibiliser aux principes qui gouvernent ces systèmes. Il est impératif d'apprendre à les

utiliser à bon escient, et de former l’esprit critique de ces acteurs envers certains risques, en particulier dans le domaine cognitif.