L'IA comprend-elle ce qu'elle fait ? Si oui, encore faudrait-il le prouver par des tests appropriés. Pas si simple, d'autant que les réseaux de neurones actuels emploient des raccourcis statistiques qui brouillent les pistes.

Melanie Mitchell

Vous souvenez-vous du jeu Jeopardy diffusé en France de 1989 à 1992 ? Une sorte d'anti-quiz : l'animateur fournissait une réponse, et les candidats imaginaient la question. Aux États-Unis, patrie originelle du programme (la trente-huitième saison est en cours aujourd'hui !), une intelligence artificielle (IA) conçue par IBM et baptisée Watson est devenue un crack à ce jeu en battant, en février 2011, deux anciens champions humains. Une publicité de l'époque prétendait même que « Watson comprend le langage naturel dans toute sa complexité et son ambiguïté ». Cependant, les promesses n'ont pas été tenues, et Watson a échoué de manière spectaculaire dans sa quête pour « révolutionner la médecine avec l'intelligence artificielle ». Preuve qu'une facilité linguistique de surface est bien différente d'une réelle compréhension du langage humain.

Depuis longtemps, le traitement du langage naturel est un des objectifs majeurs de la recherche en IA. Les chercheurs ont d'abord essayé de programmer à la main tout ce dont une machine peut avoir besoin pour comprendre un magazine, un roman ou toute autre production écrite. Cette approche, Watson l'a montré, était vaine – il est impossible de coucher noir sur blanc tous les faits, règles et suppositions qui n'ont pas encore été écrits et sont nécessaires pour comprendre un texte.

Plus récemment, un nouveau paradigme a été établi : au lieu de les abreuver de connaissances explicites et les dresser à prédire des mots, nous laissons les machines comprendre le langage par elles-mêmes en leur faisant ingurgiter d'énormes quantités de textes. Le résultat est ce que les chercheurs appellent un « modèle de langage ». Le GPT-3 de la société OpenAI, inauguré en mai 2020, par exemple, peut produire de la prose (et de la poésie !) humaine à s'y méprendre, et conduire des raisonnements linguistiques apparemment raffinés.

Mais GPT-3, entraîné avec des textes issus de milliers de sites web, de livres et d'encyclopédies, va-t-il au-delà du simple vernis de Watson ? Comprend-il les mots qu'il produit et sur lesquels il raisonne ostensiblement ? C'est là un sujet de vif désaccord au sein de la communauté des chercheurs en IA. De telles discussions étaient l'apanage des philosophes, mais tout au long de la décennie passée, l'IA a jailli hors de sa bulle académique vers le monde réel. La question n'est pas que rhétorique, car le manque de compréhension de ce monde peut avoir des conséquences concrètes et parfois dévastatrices. Dans une étude portant sur les recommandations de Watson pour lutter contre le cancer, des « exemples multiples de traitements incorrects et peu sûrs » ont été pointés. D'autres travaux ont montré que le système de traduction de Google commettait d'importantes erreurs lorsqu'il s'appliquait à des instructions médicales destinées à des patients non anglophones.

Comment juger, en pratique, si une machine a la capacité de comprendre ? En 1950, le pionnier de l'informatique Alan Turing a essayé de répondre à cette question avec son fameux « jeu de l'imitation », aujourd'hui connu sous le nom

En bref

> Donner aux machines la maîtrise du langage naturel est, de longue date, un objectif de l'intelligence artificielle.

> Entraînés de façon adéquate, les réseaux de neurones comme GPT-3 semblent capables de performances spectaculaires. Mais comprennent-ils vraiment ?

> Du test de Turing aux schémas Winograd, il n'est pas aisé d'établir un test prouvant une vraie compréhension, notamment parce que les programmes recourent à des raccourcis statistiques.

> Les machines ne sauront sans doute comprendre le langage que lorsqu'elles sauront aussi comprendre le monde qui les entoure.

de « test de Turing ». Une machine et un humain, tous deux cachés, sont en compétition pour convaincre un juge humain de leur humanité en n'usant que de la conversation écrite. Si le juge est incapable de trancher, pensait Turing, nous devrions alors considérer que la machine pense – et, en effet, comprend.

Malheureusement, Turing a sous-estimé la propension des humains à se laisser duper. Même de simples robots de conversation, tels qu'Eliza – l'ersatz de psychothérapeute conçu par Joseph Weizenbaum en 1960 –, ont réussi à faire croire à des testeurs qu'ils discutaient avec un être doué de compréhension, alors même qu'ils savaient parler à une machine.

MAÎTRE DU MONDE POUR UN PRONOM

Dans un article de 2012, les informaticiens Hector Levesque, Ernest Davis et Leora Morgenstern ont proposé un test plus objectif, qu'ils ont baptisé « challenge du schéma Winograd ». Ce test a, depuis, été adopté par la communauté du langage de l'IA comme moyen, peut-être le meilleur, d'évaluer le degré de compréhension d'une machine – bien que, nous le verrons, il ne soit pas parfait. Un schéma Winograd, nommé d'après le chercheur en linguistique Terry Winograd, de l'université Stanford, consiste en une paire de phrases différant d'un mot exactement, chacune suivie d'une question. Voici deux exemples :

Paire 1, phrase 1 : J'ai versé de l'eau de la bouteille dans la tasse jusqu'à ce qu'elle soit *pleine*.

Question : Qu'est-ce qui était rempli, la bouteille ou la tasse ?

Cet article a d'abord été publié en anglais par ***Quanta Magazine,*** *une publication en ligne indépendante, soutenue par la Simons Foundation afin de favoriser la diffusion des sciences : bit.ly/3MFPKtr*

Paire 1, phrase 2 : J'ai versé de l'eau de la bouteille dans la tasse jusqu'à ce qu'elle soit *vide*.

Question : Qu'est-ce qui était vide, la bouteille ou la tasse ?

Paire 2, phrase 1 : L'oncle de Joe peut encore le battre au tennis, bien qu'il soit trente ans plus *vieux*.

Question : Qui est plus vieux, Joe ou l'oncle de Joe ?

Paire 2, phrase 2 : L'oncle de Joe peut encore le battre au tennis, bien qu'il soit trente ans plus *jeune*.

Question : Qui est plus jeune, Joe ou l'oncle de Joe ?

Dans chaque paire, la différence d'un mot *(en italique)* peut changer la personne ou la chose à laquelle le pronom fait référence. Répondre correctement à ces questions semble nécessiter une compréhension de type sens commun. Les schémas Winograd sont conçus précisément pour évaluer cela, en réduisant la vulnérabilité du test de Turing aux juges humains non fiables et aux astuces des robots conversationnels. En particulier, les auteurs ont conçu quelques centaines de schémas qu'ils pensent être imperméables à Google : une machine ne devrait pas être capable de lancer une recherche Google (ou tout autre moteur de recherche) pour répondre à ces questions correctement.

Ces schémas ont été l'objet d'une compétition en 2016 durant laquelle le programme vainqueur a répondu juste à seulement 58 % des phrases – un résultat à peine meilleur que celui qu'elle aurait obtenu en devinant les réponses au hasard. Ce qui fait dire malicieusement à Oren Etzioni, éminent chercheur en IA : « Quand