

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 différenciant 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 ?

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



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