

→ Plus qu'un simple jouet, Moxie est un robot compagnon destiné aux 5-10 ans, capable, selon son concepteur Embodied, de converser en langage naturel pour instruire les enfants et, plus encore, les inciter à développer leurs relations sociales dans la vie de tous les jours. Connectée à une plateforme de mise à jour, l'IA de Moxie évolue à mesure que l'enfant grandit.



en lui donnant l'impression de converser avec un humain.

Comment évaluer les performances d'un agent conversationnel en matière d'imitation, de conversation, de mémorisation, de raisonnement, de simulation affective, d'anticipation ? L'une des difficultés est que ces capacités sont conférées par des algorithmes et des techniques d'apprentissage profond, qui gardent une forme d'opacité : ces techniques fonctionnent selon un principe transparent et bien compris, mais elles produisent des décisions dont il est impossible de retracer les raisons et le cheminement. Hormis ce problème dit « de boîte noire », nous sommes en droit de nous poser bien d'autres questions sur les *chatbots*, en particulier leurs enjeux éthiques. Sur ce sujet, un avis que j'ai corédigé à la CNPEN (Commission nationale pilote d'éthique du numérique) a été rendu au Premier ministre en novembre 2021.

À TROP FRÉQUENTER LES HUMAINS, LE RISQUE EST DE DEVENIR AGRESSIF

Par exemple, la machine en question respecte-t-elle la vie privée de ses utilisateurs et des autres humains ? Se comporte-t-elle de façon éthiquement responsable, et, si oui, est-ce que ce sera toujours le cas jusqu'à la fin de sa vie ? Peut-on créer des jumeaux numériques de personnes décédées et quelles seraient nos relations avec elles ? Est-ce suffisant de tester la machine avant sa mise sur le marché si elle est amenée à s'adapter en continu ? De telles questions sont particulièrement importantes dans le cas de systèmes capables d'apprendre tout au

long de leur durée d'utilisation. Il est alors difficile de contrôler leur comportement futur et de garantir que ce dernier restera acceptable. Un *chatbot* au contact d'humains pourrait développer des attitudes agressives ou inconvenantes, un algorithme conçu pour faire des offres financières pourrait à terme orienter ses propositions en fonction du groupe social auquel appartient son interlocuteur, etc.

Un exemple d'une telle dérive a été fourni en 2016 par Tay, un système conversationnel développé par Microsoft. Ce *chatbot*, déployé pour converser avec des jeunes sur la plateforme Twitter, était doté d'une capacité d'apprentissage en continu. On a pu constater qu'il pouvait apprendre de ses interactions avec les internautes, puisqu'il adaptait son langage aux habitudes et éléments qu'il repérait dans les conversations. Mais l'essai a tourné court : au bout de vingt-quatre heures seulement, Tay s'est mis à poster sur la plateforme des tweets à caractère raciste et nazi !

Comment faire évoluer le test de Turing pour évaluer les capacités des agents conversationnels ? Rappelons-le, ce test éprouve la faculté d'une machine à imiter la conversation humaine. La mission du juge est de deviner,

d'après les messages échangés, qui est, de l'un ou de l'autre de ces interlocuteurs, une machine. Le contenu des messages est libre : il est possible de parler de tous types de sujets. Alan Turing a prédit que, en l'an 2000, des ordinateurs avec 128 Mo de mémoire seraient capables de tromper environ 30 % des juges humains durant une mise à l'épreuve de cinq minutes. On en est loin.

Des biais ne sont pas exempts du test de Turing. Jean-Paul Delahaye, chercheur en informatique à Lille (et auteur des articles des pages 60 et 76), explique par exemple qu'une simple question de calcul est susceptible de démasquer l'interlocuteur-programme : l'humain sera laborieux, alors que la machine excellera. Mais, pourrait-on rétorquer, rien n'empêche de programmer la machine pour qu'elle fasse exprès de se tromper...

Toujours est-il que faire passer le test de Turing reste un graal pour tous les concepteurs d'agents conversationnels. Depuis 1990, il existe même une compétition annuelle, le prix Hugh Loebner, qui couronne les systèmes satisfaisant le mieux les critères du test de Turing, c'est-à-dire avec lesquels il est le plus difficile de déterminer s'il s'agit d'un humain ou d'une machine. Les meilleurs programmes n'effectuent qu'un simulacre de conversation qui, malgré de récents progrès (systèmes capables de répondre dans des domaines ouverts, adaptables à différents interlocuteurs, dotés de certains traits de personnalité et de détection des émotions...), font rarement illusion très longtemps.

SURVEILLER LA COÉVOLUTION DES HUMAINS ET DES MACHINES

Il reste donc à trouver des façons pertinentes d'évaluer les diverses capacités, tant cognitives qu'émotionnelles, des machines. Pour ce qui est du cognitif, de nombreux chercheurs imaginent des scénarios inspirés du test de Turing, mais composés de plusieurs épreuves dans des contextes différents. Dans des versions scolaires, les machines passent des examens, comme des écoliers. D'autres évaluations portent sur la capacité à démêler des ambiguïtés du langage naturel, sur la préhension d'objets, etc. Quant aux capacités émotionnelles, la conception de tests en la matière constitue un nouveau chantier de la recherche en intelligence artificielle.

Ainsi, avec des *chatbots* munis de facultés que l'on considérerait comme le propre des humains il n'y a pas si longtemps, des variantes

du test de Turing ou des batteries d'examens sont à développer – qui pourront être à la base d'une sorte d'autorisation de mise sur le marché et de contrôle technique pour robots tout au long de leur vie. C'est un axe important des recherches actuelles et de défis pour la normalisation. Enfin, il est nécessaire de surveiller la coévolution et la coadaptation des humains et des machines dans leurs interactions. Les *chatbots* pourront nous manipuler ou nous mentir pour notre bien-être, mais encore faudra-t-il s'assurer que c'est bien le cas. Il faut se préoccuper de l'évolution, à long terme, du comportement des humains vis-à-vis des agents conversationnels, afin d'éviter l'isolement social, un attachement démesuré à la machine ou des manipulations abusives. C'est aussi dans cette direction que doit évoluer le test de Turing. Il doit porter non seulement sur les capacités des machines, mais aussi sur leur propension à modifier nos croyances et notre comportement vis-à-vis d'elles.

Ajoutons, pour finir, qu'une bonne information sur les capacités de ces assistants conversationnels, sur la façon dont ils fonctionnent, sur les biais qu'ils peuvent présenter, permettrait aux utilisateurs de prendre plus de recul et de se comporter de façon adéquate pour relever ce grand défi : bien vivre avec les machines.

— L'autrice —

> **Laurence Devillers**
est professeure en intelligence artificielle à Sorbonne Université et dirige une équipe de recherche au CNRS-LISN (Saclay), où elle est responsable de la chaire IA HUMAINE (<https://humaine-chaireia.fr>).

— À lire —

> **D. Adiwardana et al.**, Towards a Human-like Open-Domain Chatbot, 2020. arxiv.org/abs/2001.09977

> **L. Devillers**, Des robots et des hommes : mythes, fantasmes et réalité, *Plon*, 2017.

> Vivre avec les robots, vidéo réalisée par le *Limsi-CNRS* : youtu.be/p1ID-gvUnWs

> Les robots en quête d'humanité, *Dossier Pour la Science*, n° 87, avril-juin 2015.

> **A. M. Turing**, Computing machinery and intelligence, *Mind*, 1950.