

Le danger manifeste et actuel de l'IA



[Source : aubedigitale.com]

Par Bruce Abramson

L'intelligence artificielle menace-t-elle de conquérir l'humanité ? Au cours des derniers mois, cette question est passée des pages des romans de science-fiction au premier plan de l'attention des médias et des gouvernements. Il n'est toutefois pas certain que beaucoup de ceux qui ont discuté de cette question aient compris les implications de ce bond.

Dans l'esprit du public, la menace se concentre sur la confusion inhérente à des contrefaçons de plus en plus performantes et sur ses conséquences pour le marché de l'emploi, ou bien elle s'oriente vers des directions qui feraient l'objet d'un grand film : Que se passerait-il si les systèmes d'IA décidaient qu'ils sont supérieurs aux humains, prenaient le contrôle et mettaient en œuvre des plans génocidaires ? Cette dernière perspective est évidemment la plus convaincante des deux.

Bien qu'un tel cauchemar soit possible en théorie, il est peu probable qu'il se produise. Le danger clair et présent que représente l'IA nous détruira bien avant que nos serviteurs automatisés rebelles ne se déclarent nos suzerains exterminateurs. Deux mots essentiels résument la menace : valeurs et autorité.

Les valeurs d'abord. Malgré toute leur sophistication et leur mystère, les systèmes d'IA sont essentiellement des détecteurs de modèles. Presque tous les comportements humains – et une part encore plus importante des événements non humains – suivent des schémas prévisibles. Des mécanismes d'enregistrement de plus en plus sophistiqués fournissent aux systèmes d'IA un ensemble croissant de données passées dans lesquelles ils peuvent trouver des modèles. Des algorithmes de plus en plus sophistiqués fournissent aux systèmes d'IA des capacités qui s'améliorent rapidement pour trouver des modèles dans ces données.

L'IA ne devient toutefois intéressante que lorsqu'elle projette ces schémas dans l'avenir. Peu de gens se soucient, par exemple, du fait que l'IA puisse trouver des modèles dans Shakespeare, mais beaucoup seront fascinés lorsque

l'IA composera une « tragédie shakespearienne » se déroulant pendant la guerre de Sécession en utilisant une langue, un style, un idiome et des compétences qui étaient auparavant considérés comme propres au Barde.

C'est dans la projection et la prédiction que les valeurs entrent en jeu. Toute décision implique une certaine évaluation des valeurs. Dans la plupart des contextes de contrôle mécanisé actuellement soumis aux conseils de l'IA, les valeurs fondamentales sont si peu controversées qu'elles échappent à l'attention. Peu de gens (à part peut-être les kamikazes ou les fabricants de munitions) contesteraient les propositions selon lesquelles il est « mauvais » que les moteurs surchauffent, que les chaudières explosent ou que les véhicules s'écrasent. Ces croyances n'ont rien de « naturel » – un météore est indifférent au fait qu'il se précipite dans l'espace ou qu'il s'écrase sur une planète – mais elles sont si intrinsèquement humaines qu'elles transcendent les différences de culture et de temps.

Ce consensus quasi universel s'estompe toutefois lorsque des êtres humains entrent en scène. Prenons l'exemple de Dylan Mulvaney, l'influenceur transgenre des médias sociaux qui a récemment contribué à redorer le blason de Bud Light. (Selon le système de croyances woke contemporain, Mulvaney est une femme qui a eu la malchance de naître dans un corps masculin. Selon tout autre système de croyance ayant jamais existé, Mulvaney est un homme qui a choisi de vivre en tant que femme.

Comment l'IA devrait-elle désigner Mulvaney : « il » ou « elle » ? La réponse démontre une préférence pour un ensemble de croyances et de valeurs plutôt qu'un autre. Comment l'IA a-t-elle pu déduire cette préférence ?

Deux voies sont possibles : elle pourrait s'être formée à partir d'un grand nombre de réflexions éthiques tirées des milliers de systèmes de croyances qui ont vu le jour au cours de l'histoire du monde et en avoir conclu qu'un système spécifique était le meilleur, ou elle pourrait avoir adopté les préférences de valeurs de ses concepteurs et formateurs humains.

Peut-être qu'un jour, un système d'IA suivra la première voie. Aujourd'hui, personne ne doute que les tendances wakes des systèmes d'IA tels que ChatGPT reflètent les tendances wakes des professionnels de la technologie qui les ont développés. Ce que cela signifie pour la plupart des utilisateurs, c'est que les IA sur lesquelles ils pourront bientôt s'appuyer pour émettre des jugements et des recommandations fondés sur des valeurs ne partagent pas leurs propres valeurs.

Cette reconnaissance fait entrer en jeu l'autorité. Le véritable danger imminent de l'IA provient de l'intégration potentielle des systèmes d'IA dans le culte de l'expertise.

Dans de nombreux domaines, nous avons choisi de faire preuve d'une déférence quasi totale à l'égard des experts accrédités. Pendant Covid, par exemple, il était presque hérétique de remettre en question les recommandations du Dr Fauci, de la Dr Birx, des CDC ou de la FDA. Cependant, même à cette époque,

il était encore (à peine) possible de noter que ces experts avaient peut-être indûment ignoré les conséquences économiques négatives de leurs recommandations en matière de santé publique. Imaginez la prochaine épidémie, lorsque le conseiller politique sera une IA qui aura pris en compte des centaines de milliers de variables pour évaluer des quadrillions de scénarios.

Un tel résultat est le véritable avenir cauchemardesque : la dictature effective d'une IA qui ne partage pas vos croyances ou vos valeurs fondamentales.

Bien avant qu'une IA exterminatrice ne prenne le contrôle du monde en son nom propre, nos gouvernements et nos concitoyens accorderont à un système d'IA le pouvoir de déclarer des situations d'urgence, de supprimer les libertés civiles et de contrôler nos vies.

Tant que nous n'aurons pas redéveloppé les mécanismes gouvernementaux dont nous avons besoin pour préserver notre liberté face à l'assaut des experts, nous ne serons jamais à l'abri des privations potentielles de la gouvernance de l'IA.

L'IA elle-même peut être contenue. La combinaison de l'IA, d'une dévotion servile et d'une application généralisée ne le sera pas. Nous devons mettre un frein à notre croyance collective en l'infailibilité d'une classe d'experts, sinon les conseillers en IA pourraient bien mettre fin à notre civilisation.

Bruce Abramson, PhD, JD, est président de la société de conseil stratégique Informationism, Inc. et directeur de l'American Center for Education and Knowledge. Il a été l'un des premiers à utiliser des simulations à grande échelle et des analyses statistiques dans les systèmes d'intelligence artificielle. Il a écrit cinq livres, dont « The New Civil War: Exposing Elites, Fighting Utopian Leftism, and Restoring America » (RealClear Publishing, 2021).

Traduction de Real Clear Wire par Aube Digitale