FÉDÉRATION DES FINANCES

ARTICLE DU JOUR

11 février 2025

Fédération des Finances FORCE OUVRIERE - 46, rue des petites écuries - 75010 Paris • www.financesfo.fr

Le Monde

# Intelligence artificielle : les chercheurs identifient trois types de risques

Le dernier rapport international sur la sécurité des systèmes d'IA développe les « utilisations malveillantes », les « dysfonctionnements » et les « risques systémiques » liés au développement de la technologie, et expose désaccords et tensions entre experts mondiaux.

Par David Larousserie

En matière de sécurité des systèmes d'intelligence artificielle (IA), *« le futur est incertain »*, comme le proclame de façon consensuelle le dernier rapport international sur le sujet, publié le 29 janvier. Ce constat peu tranché peut surprendre, tant la personne qui a présidé les travaux apparaît plus souvent inquiète que sereine sur ce sujet. Yoshua Bengio, professeur à l'université de Montréal, pionnier multirécompensé de cette discipline, a signé en 2023 deux des principaux textes tirant la sonnette d'alarme sur le développement de l'IA. D'abord une tribune qui demandait une pause dans la mise au point d'agents conversationnels comme ChatGPT. Puis une pétition s'inquiétant d'un *« risque d'extinction lié à l'IA »*, comparable aux *« pandémies et guerres nucléaires »* et demandant que la réduction de ce danger soit une *« priorité mondiale »*.

Ce ton catastrophiste n'est pas celui adopté par les 300 pages du rapport, rédigé par une centaine d'experts de 30 pays et riche de plus de 1 360 références scientifiques. M. Bengio, mandaté en novembre 2023 par le gouvernement britannique, a voulu en faire, toutes proportions gardées, un document à la manière des rapports scientifiques du Groupe d'experts intergouvernemental sur l'évolution du climat. En outre, aucune recommandation n'est faite, seulement le souhait que les décideurs s'emparent de cet état des lieux.

Le texte traduit les tensions entre spécialistes. Les plus catastrophistes jouent du marketing de la peur pour convaincre décideurs et investisseurs que leur soutien à ces technologies capables de telle puissance est utile. D'autres regrettent que l'attention sur le pire masque des problèmes réels déjà en cours (désinformation, discrimination, violations de droits…). D'autres encore dédramatisent en minimisant les capacités intelligentes de ces systèmes. Les uns s'inquiètent du court terme, quand les autres voient des problèmes très loin de nous.

## Armes biologiques ou chimiques

La principale vertu du rapport est donc d'éclaircir ce sujet controversé. Il distingue trois catégories de risques : les *« utilisations malveillantes »*, les *« dysfonctionnements »* et les *« risques systémiques »*.

La première catégorie regroupe les tentatives de désinformation, de cyberharcèlement ou de manipulation de l'opinion, que l'IA n'a pas inventées, mais qu'elle amplifie, par la création rapide de contenus, par ses capacités à déformer ou à détourner des images. Une autre facette concerne l'utilisation de l'IA pour fabriquer des armes biologiques ou chimiques. Il ne s'agit pas de robots laborantins produisant diverses toxines ou poisons, mais plutôt de la capacité des IA à aider des terroristes à les élaborer (recette, manière d'échapper à des contrôles…). Les auteurs parlent de technologies duales, pouvant être utilisées à des fins bénéfiques ou, au contraire, négatives. Ils mettent dans cette catégorie le logiciel récompensé du prix Nobel de chimie en 2024, AlphaFold, qui, en prédisant les interactions entre protéines, peut servir à fabriquer des médicaments comme des poisons.

Du côté des dysfonctionnements, le rapport liste les problèmes de fiabilité (erreurs, inventions…), la présence de biais discriminants ou encore la perte de contrôle. Ce dernier point a été très débattu, *« certains considèrent qu'il n'est pas plausible, d'autres qu'il est probable, et d'autres encore qu'il s'agit d'un risque de faible probabilité »*… En ligne de mire, notamment, les scénarios recourant à des agents autonomes, capables de prendre des décisions et d'agir en lançant des programmes ou en cliquant sur un écran, quitte à faire n'importe quoi. Aux journées scientifiques consacrées à l'IA, à l'Institut polytechnique de Paris, les 6 et 7 février, Ece Kamar, directrice de l'IA chez Microsoft Research, a raconté une anecdote pour illustrer que ce risque est à prendre au sérieux. L'agent autonome qu'elle a sollicité pour l'aider à remplir des mots croisés du *New York Times*, a tâtonné jusqu'à lui demander la réinitialisation de son mot de passe. *« Il faut accélérer la recherche sur la sécurité des agents »*, a-t-elle conclu.

## Avancées fulgurantes

Enfin, dans la catégorie des risques systémiques, la liste s'étend du coût environnemental aux pertes d'emplois, en passant par les infractions aux droits d'auteur, les atteintes à la vie privée ou les inégalités de développement entre pays dotés ou non de ces capacités.

La tonalité générale apparaît finalement d'autant plus inquiétante à la lecture des ajouts effectués depuis la version provisoire, publiée en mai 2024, qui tiennent compte des avancées fulgurantes dans le domaine. Qu'il s'agisse de scores en hausse à de nombreuses évaluations, de la multiplication des agents autonomes ou de l'arrivée des systèmes de « raisonnement », tels les modèles o1 et o3 d'OpenAI, ces progrès renforcent les scénarios de risque présentés. Les auteurs y ajoutent un autre volet, également à l'origine de désaccords entre experts, sur les systèmes dits « open source », dont l'accès au code facilite tout autant l'innovation que des réutilisations malveillantes.

*« L'IA n'est pas une fatalité. Les choix des citoyens détermineront son avenir* (…)*. Il est urgent d'œuvrer en faveur d'un accord international et de consacrer des ressources à la compréhension et à la prise en compte des risques liés à cette technologie »*, conclut le rapport.