



IDÉE 4

LES **BIAIS** DANS LES SYSTÈMES **D'IA** POURRAIENT PERDURER

La partialité est une caractéristique de la prise de décision humaine et de l'IA. Comme les données utilisées pour former l'IA sont souvent entachées de biais difficiles à corriger, le recours croissant à l'IA dans les systèmes de prise de décision pourrait répandre les biais et entraîner des dommages importants. Les préjugés ne peuvent jamais être éliminés, en partie à cause des perspectives contradictoires sur l'équité.

AUJOURD'HUI

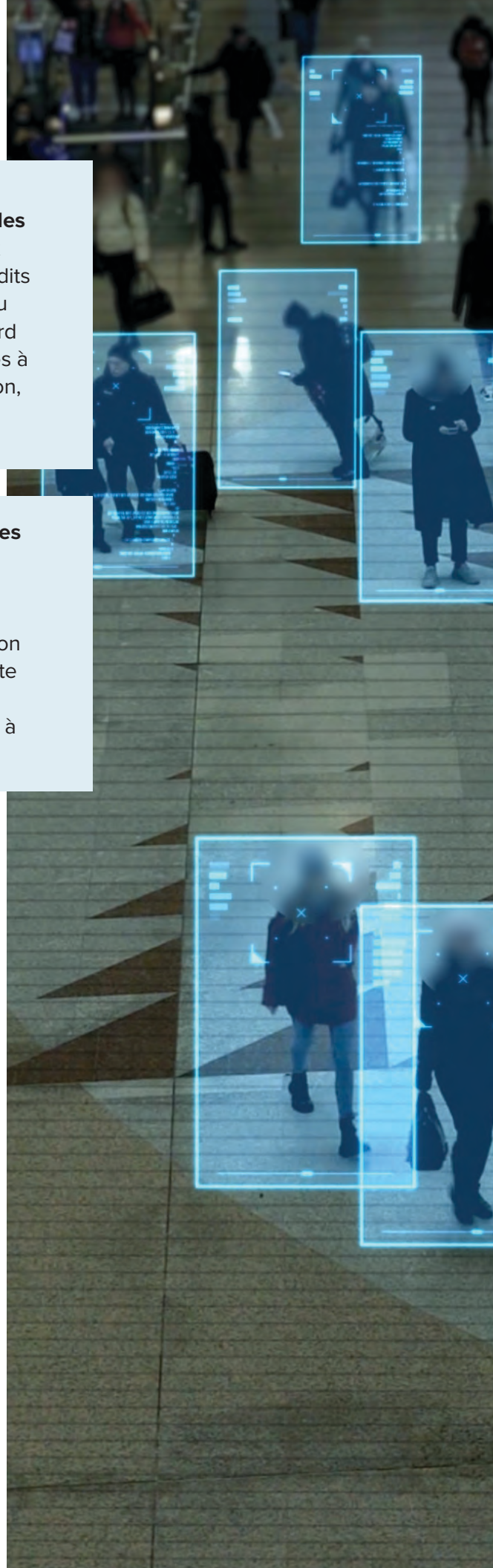
Les préjugés dans l'IA sont considérés comme un problème majeur capable d'automatiser la discrimination à grande échelle d'une manière qui peut être difficile à déterminer. Si les décisions humaines sont également biaisées, l'un des principaux risques de l'automatisation des décisions à fort enjeu est qu'elles se généralisent, augmentant ainsi la possibilité d'erreurs et de préjudices systémiques. Alors qu'un gestionnaire partial pourrait décider d'accorder des notes d'entretien plus élevées aux quelques candidats qui lui ressemblent et parlent comme lui, un modèle d'IA biaisé pourrait avoir un effet semblable sur des milliers de personnes au sein d'organisations, de secteurs ou de pays.



De nombreux produits d'IA prétendent être moins biaisés que les décideurs humains, mais les enquêtes indépendantes ont révélé des échecs et des rejets systématiques¹. Par exemple, un audit de deux outils d'embauche par IA a révélé que les types de personnalité prédits variaient selon que le candidat soumettait son CV en format Word ou en texte brut². Des outils semblables ont été discriminatoires à l'égard des femmes³ ou des personnes handicapées⁴. Les biais sont intégrés à l'IA à de nombreux stades de son cycle de vie : données de formation, développement algorithmique, interaction avec l'utilisateur et retour d'information⁵.

Les biais peuvent être impossibles à éliminer parce que les données utilisées pour former les modèles d'IA sont elles-mêmes souvent biaisées d'une manière qui ne peut pas être facilement corrigée.

Le contrôle des résultats peut également poser des problèmes. Par exemple, un modèle d'IA qui apprend à écarter les mots à connotation raciste pourrait omettre des informations importantes sur l'Holocauste ou l'esclavage⁶. En outre, les algorithmes ne peuvent souvent pas calculer différentes notions d'équité en même temps, ce qui conduit à des résultats constamment différents pour certains groupes^{7, 8, 9}.





AVENIRS

Dans un avenir où les préjugés ne pourront jamais être éliminés, qu'ils soient humains ou algorithmiques, les sociétés devront peut-être repenser les idées actuelles sur l'équité et la meilleure façon d'y parvenir. Les gens ne sont pas nécessairement d'accord sur le sens du mot « équitable ». Par exemple, certains considèrent que la discrimination positive est juste, d'autres non. Les institutions pourraient adopter des normes visant à distribuer les ressources, emplois, subventions, prix ou autres biens, d'une manière qui tente explicitement de réparer les injustices historiques. Les organisations qui cherchent à éviter les préjugés systémiques peuvent utiliser une approche de « pluralisme algorithmique », qui implique divers éléments dans le processus de prise de décision et garantit qu'aucun algorithme ne limite sévèrement les possibilités¹⁰.

Des efforts pourraient être faits pour réduire les biais dans les systèmes d'IA à un niveau acceptable, bien qu'il soit impossible de les éliminer complètement. L'utilisation des technologies de l'IA dans certains domaines sensibles, tels que la police ou l'embauche, pourrait continuer à susciter des réactions négatives. Par ailleurs, ces technologies pourraient continuer à s'améliorer et devenir moins biaisées à l'avenir. Quoi qu'il en soit, il est probable que l'on continuera à s'efforcer de réduire les biais dans les technologies d'IA.



IMPLICATIONS

- **Les préjudices ou les échecs systémiques pourraient être institutionnalisés** dans des contextes où des algorithmes uniques sont autorisés à prendre des décisions globales concernant l'accès des personnes à certaines ressources (p. ex., des emplois, des prêts, des visas)
- **Les préjugés humains pourraient devenir plus importants** parmi ceux qui utilisent les systèmes d'IA, car les gens apprennent et reproduisent les perspectives biaisées de l'IA, portant les préjugés avec eux au-delà de leurs interactions
- **Les désaccords sur la meilleure façon de coder l'équité algorithmique** peuvent résulter de définitions différentes de ce que signifie réellement l'équité. Cela peut conduire à des résultats complètement différents pour des technologies ou des systèmes semblables
- L'incapacité à éliminer les préjugés des algorithmes pourrait **conduire à des divisions politiques, sociales ou économiques**
- Si les décisions sont de plus en plus distribuées, notamment divers algorithmes et humains à différents stades d'un processus, **il pourrait être difficile de faire des réclamations pour discrimination** ou de déterminer une partie responsable de la discrimination
- Des cas très médiatisés de discrimination algorithmique pourraient entraîner **une perte de confiance dans les systèmes décisionnels de l'IA**, en particulier dans les domaines de la police et des soins de santé, ainsi qu'une augmentation des litiges

Notes de fin de texte

- 1 Helhoski , Anna. « [AI Could Prevent Hiring Bias - Unless It Makes It Worse](#) ». NerdWallet, 12 juin 2023.
- 2 Rhea, Alene K.; Markey, Kelsey; D'Arinzo, Lauren; Schellmann, Hilke; Sloane, Mona; Squires, Paul; Arif Khan, Falaah et Stoyanovich, Julia. « [An external stability audit framework to test the validity of personality prediction in AI hiring](#) ». *Data Mining and Knowledge Discovery*, vol. 36, no 6 (2022) : p. 2153-2193.
- 3 Andrew, Lori et Hannah Bucher. « [Automating Discrimination: AI Hiring Practices and Gender Inequality](#) ». *Cardozo Law Review*. Consulté le 15 août 2024.
- 4 Brown, Lydia; Shetty, Ridhi et Richardson, Michelle. « [Algorithm-Driven Hiring Tools: Innovative Recruitment or Expedited Disability Discrimination?](#) », 3 décembre 2020.
- 5 Samuel, Sigal. « [Why It's so Damn Hard to Make AI Fair and Unbiased](#) ». Vox, 19 avril 2022.
- 6 Ferrara, Emilio. « [Éliminer les biais dans l'IA peut être impossible – un informaticien explique comment l'apprivoiser à la place](#) ». La Conversation, 19 juillet 2023.
- 7 Kleinberg, Jon. « [Inherent Trade-Offs in Algorithmic Fairness](#) ». 10 avril 2018.
- 8 Dwork, Cynthia. « [The Emerging Theory of Algorithmic Fairness](#) ». 6 septembre 2018.
- 9 Raghavan, Manish. « [What Should We Do When Our Ideas of Fairness Conflict?](#) » *Communications of the ACM*, vol. 67, no 1 (janvier 2024) : p. 88-97.
- 10 Shomik, Jain; Suriyakumar, Vinith; Creel, Kathleen et Wilson, Ashia. « [Algorithmic Pluralism: A Structural Approach To Equal Opportunity](#) ». arXiv, 21 septembre 2023.

© Sa Majesté le Roi du chef du Canada, 2025

Pour obtenir des informations sur les droits de reproduction : <https://horizons.gc.ca/fr/contactez-nous/>

PDF : PH4-214/2025F-PDF

ISBN : 978-0-660-76888-5

Also available in English under the title: Bias in AI systems could remain forever.

AVERTISSEMENT

Horizons de politiques Canada (Horizons de politiques) est le centre d'excellence en prospective du gouvernement du Canada. Notre mandat est de doter le gouvernement du Canada d'une perspective et d'un état d'esprit tournés vers l'avenir afin de renforcer la prise de décisions. Le contenu de ce document ne représente pas nécessairement le point de vue du gouvernement du Canada ou des agences et ministères participants.