



IDÉE 19

LA COURSE AUX DONNÉES **AXÉE** SUR L'IA

Une course aux données axée sur l'IA augmente la valeur des données, incite à de nouvelles tactiques de collecte de données, de nouvelles façons de générer des données, y compris des données synthétiques, et de nouvelles tentatives d'imposer des restrictions juridiques et techniques sur la façon dont les données peuvent être utilisées.

AUJOURD'HUI

Les entreprises d'IA utilisent des tactiques préoccupantes d'un point de vue éthique pour obtenir des données d'entraînement de leurs modèles d'IA.

Des niveaux d'investissement sans précédent dans l'IA créent une pression sur les entreprises pour accélérer le développement de modèles d'IA¹. Cela nécessite des données de bonne qualité, qui sont de plus en plus rares². Certaines plateformes de médias sociaux réécrivent leurs conditions d'utilisation pour permettre l'utilisation du contenu généré par leurs utilisatrices et utilisateurs dans l'entraînement de modèles d'IA. Certaines entreprises d'IA extraient des données de sites Web sans permission (voir l'encadré 1).

Boîte 1

Le moissonnage du Web est la collecte automatisée de données à partir de sites Web par des programmes appelés robots d'indexation. Les normes Web permettent aux sites Web d'étiqueter les données avec une directive « ne pas effectuer de moissonnage de données ». Dans la pratique, cependant, les robots d'indexation peuvent ou non se conformer à ces directives. Malgré les critiques, certaines entreprises d'IA font du moissonnage de données du contenu personnel ou protégé par le droit d'auteur pour l'entraînement de modèles d'IA.

De plus en plus de sites Web et de créatrices et créateurs de contenu tentent d'empêcher les entreprises d'IA d'utiliser leurs données.

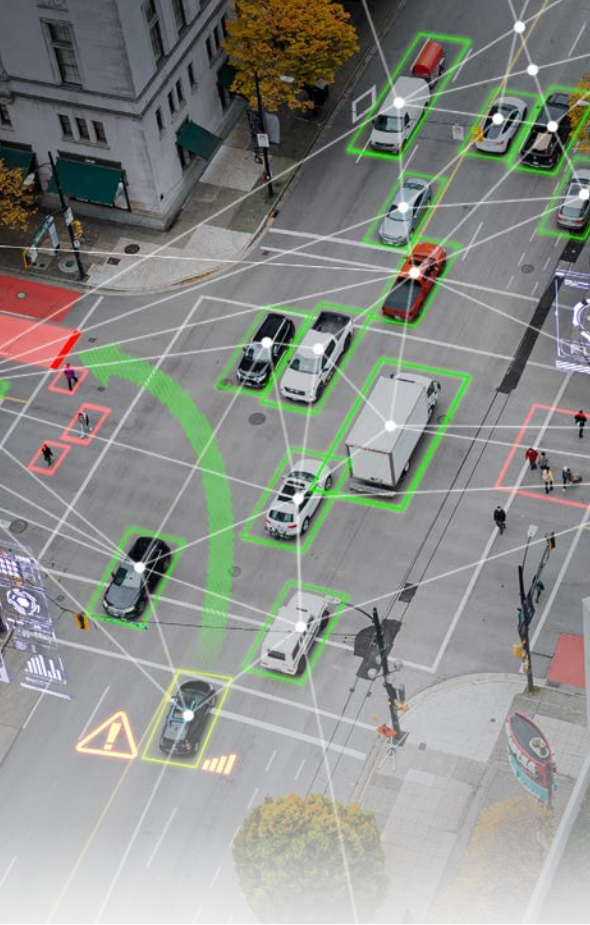
De plus en plus de sites Web modifient leurs conditions d'utilisation, bien que ces changements puissent être difficiles à appliquer et ne s'appliquent pas de façon rétroactive lorsque les données ont déjà été récupérées³. Un audit récent a révélé que 45 % des données d'un ensemble de données d'entraînement Google sont limitées par les conditions d'utilisation des sites Web⁴. De nombreuses poursuites devant les tribunaux visent à restreindre l'utilisation de matériel protégé par le droit d'auteur dans l'entraînement de l'IA⁵⁶. Des outils technologiques sont également déployés : par exemple, Kudurru de Spawning bloque le moissonnage du Web automatisé, tandis que Nightshade « empoisonne » les données d'entraînement de l'IA sans nuire à leur apparence aux yeux des humains⁷⁸.

Les données synthétiques sont de plus en plus intéressantes pour l'entraînement de l'IA.

Ayant plus de difficulté à accéder aux données générées par l'humain, de plus en plus d'entreprises d'IA explorent l'utilisation des données générées par l'IA pour la formation des modèles d'IA.⁹ Des modèles d'IA génératives sont conçus pour créer rapidement des données synthétiques qui peuvent être utilisées pour la formation de modèles plus performants¹⁰. L'utilisation de données synthétiques n'en est qu'à ses débuts. Aujourd'hui, elle est plus couramment utilisée pour des cas d'utilisation particuliers, comme l'entraînement de véhicules autonomes, mais des spécialistes suggèrent qu'elle pourrait devenir plus importante que les données réelles d'ici 2030¹¹². Cependant, elle présente certains défis. Elle est plus utile lorsqu'elle est générée à partir de données réelles, mais même si c'est le cas, elle peut simplifier exagérément les nuances du monde réel ou exagérer les biais trouvés dans les échantillons d'ensembles de données¹³. Elle peut également être vulnérable aux attaques de réidentification, révélant des renseignements personnels contenus dans des échantillons d'ensembles de données utilisés pour construire l'ensemble de données synthétiques.¹⁴

Les progrès scientifiques créent de nouvelles possibilités de générer des données dans des domaines techniques qui pourraient être utilisées pour l'entraînement de l'IA.

Par exemple, l'ADN d'un génome humain peut maintenant être séquencé en une seule journée, comparativement à une décennie au tournant du siècle¹⁵. Des efforts sont en cours pour rendre plus d'ensembles de données scientifiques faciles à trouver, accessibles, interopérables et réutilisables (FAIR) à l'usage des équipes de recherche en IA¹⁶.



AVENIRS

La valeur des *données réelles* pourrait continuer d'augmenter, ce qui stimulerait les investissements dans la collecte et l'étiquetage des données. Le marché de l'annotation et de l'étiquetage des données devrait croître de près de 600 % d'ici 2030¹⁷. Les organisations pourraient se concentrer davantage sur le fait de rendre les données qu'elles recueillent « prêtes pour l'IA » afin qu'elles puissent être plus facilement monétisées¹⁸. L'incitation à recueillir plus de données pourrait réduire les inégalités, par exemple en menant à de nouveaux efforts de collecte de données dans les régions éloignées ou dans les pays du Sud, ce qui pourrait améliorer le développement des services pour ces groupes¹⁹. Les gens pourraient se voir offrir de nouveaux incitatifs pour communiquer leurs données, comme l'accès à des services numériques ou des rabais sur les appareils qui génèrent des données.

Une dépendance excessive aux données synthétiques générées par l'IA pourrait entraîner un « effondrement du modèle ». Si les modèles d'IA sont de plus en plus entraînés sur des données synthétiques de qualité inférieure aux données réelles, cela pourrait entraîner une dégradation autorenforcée des modèles d'IA²⁰. À mesure que le contenu synthétique domine Internet, le risque d'effondrement des modèles pourrait augmenter²¹.

Les décisions judiciaires pourraient restreindre l'utilité des modèles d'IA. Si les tribunaux décident que les données protégées par le droit d'auteur ne peuvent pas être utilisées pour l'entraînement de modèles, certains modèles d'IA pourraient devoir être retirés et réentraînés²². Les nouveaux modèles entraînés uniquement sur des données obtenues avec une permission explicite pourraient avoir des capacités plus limitées. Par exemple, StarCoder2 est un modèle d'IA entraîné exclusivement sur le contenu communiqué par les développeuses et développeurs de codes qui ont choisi d'y participer, et est moins performant que ChatGPT²³.

Boîte 2

Types de biais dans les ensembles de données

Le biais historique résulte des données reflétant les inégalités passées ou existantes.

Le biais d'échantillonnage résulte de la sous-représentation de certains groupes dans les données.

Le biais d'étiquetage résulte du fait que les groupes qui étiquettent les données imposent leurs points de vue subjectifs.

Le biais de mesure résulte des méthodes de collecte de données qui favorisent certains résultats.

Le biais de confirmation résulte de la sélection par les collecteurs de données qui confirment leurs croyances.

De nouvelles façons de restreindre l'utilisation des données pourraient être codées dans les données elles-mêmes. Les conventions sur le moissonnage du Web ne semblant plus suffisantes, de nouvelles approches pourraient voir le jour pour changer la façon dont les agents d'IA interagissent avec les données en ligne. Des entreprises telles que Meta, IBM,²⁴ Google et Microsoft²⁵ proposent une norme²⁶ de provenance des données et une norme²⁷ technique qui pourrait ajouter des filigranes aux données. Le codage des directives d'utilisation des données au sein même des données pourrait améliorer la transparence, aidant les développeuses et développeurs à éviter les problèmes de droits d'auteur ou les biais (voir l'encadré 2)²⁸. Cela pourrait également augmenter la pression sur les développeuses et développeurs d'IA pour trouver de nouvelles sources de données ou des moyens de contourner les nouvelles restrictions.

D'autres techniques d'entraînement de l'IA pourraient donner aux utilisatrices et utilisateurs une plus grande confidentialité et un plus grand contrôle. L'apprentissage fédéré est une technique qui ne repose pas sur un dépôt central de données, mais qui s'appuie sur des sources distribuées de manière à préserver la vie privée et à donner aux propriétaires de données un plus grand contrôle sur les droits d'accès²⁹. De tels modèles décentralisés de l'entraînement de l'IA³⁰ pourraient encourager davantage de personnes et d'organisations à autoriser l'accès à leurs données, ce qui pourrait générer plus de nouvelles perspectives dans des domaines tels que les voitures autonomes³¹, la recherche sur les matériaux³² et les soins de santé³³.

IMPLICATIONS

- Ne pas résoudre le goulot d'étranglement de l'acquisition des données **pourrait ralentir l'innovation dans le domaine de l'IA**, augmenter le coût des données d'entraînement et refroidir l'intérêt des investisseuses et des investisseurs.
- De nouveaux **incitatifs financiers pour l'échange de données** pourraient avoir une incidence sur le comportement des individus et des organisations :
 - › Les groupes vulnérables pourraient être exploités pour leurs données.
 - › Le maintien de la vie privée pourrait devenir plus difficile ou plus coûteux.
 - › Les initiatives de recherche pourraient choisir de monétiser leurs ensembles de données³⁴ plutôt que de les communiquer ouvertement, ce qui nuirait à l'innovation et la science³⁵.
- Les systèmes utilisant des modèles d'IA entraînés sur des données synthétiques peuvent être vulnérables à l'effondrement des modèles, ce qui dégrade leur rendement et **réduit la confiance du public dans l'IA** et dans les institutions qui ont intégré l'IA dans leurs services
- Si les jugements des tribunaux limitent l'accès légal à **de puissants modèles** d'IA qui ont été entraînés sur du matériel protégé par le droit d'auteur, les cybercriminelles et cybercriminels, et d'autres personnes désireuses de continuer à utiliser ces modèles illégalement, pourraient obtenir un avantage tactique.
- **Des normes et des certifications**, comme Fairly Trained, qui établissent les ensembles de données de sources ouvertes, entraînés à l'éthique et exempts d'infractions, pourraient **devenir plus influentes**³⁶.
- Les pays pourraient **se disputer l'accès aux données** et considérer de plus en plus leurs propres données comme une ressource stratégique à protéger et à contrôler. Cela pourrait miner la coopération mondiale dans des domaines tels que la science³⁷.
- Les pays pourraient adopter **des normes et des protocoles de données différents**, qui entraînent une fragmentation des marchés. Par exemple, le *Règlement sur les marchés numériques* de l'Union européenne, qui oblige les entreprises technologiques à assurer l'interopérabilité et l'accès à certains types de données générées sur leurs plateformes³⁸, a conduit Apple à décider de ne pas déployer sa fonctionnalité d'IA « Apple Intelligence » dans l'Union européenne.³⁹

Notes de fin de texte

- 1 Kinding, Beth. « Big Tech Battles On AI, Here's The Winner ». Forbes. Consulté le 8 octobre 2024. <https://www.forbes.com/sites/bethkindig/2024/08/08/microsoft-leads-big-tech-in-ai-monetization-amazon-a-close-second/>.
- 2 Villalobos, Pablo. « Will We Run Out of ML Data? Evidence From Projecting Dataset Size Trends ». Epoch AI, 10 novembre 2022. <https://epochai.org/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset>.
- 3 Roose, Kevin. « The Data That Powers A.I. Is Disappearing Fast ». *The New York Times*, 19 juillet 2024, rubrique Technology. <https://www.nytimes.com/2024/07/19/technology/ai-data-restrictions.html>.
- 4 Longpre, Shayne, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena et coll. « Consent in Crisis: The Rapid Decline of the AI Data Commons ». arXiv, 24 juillet 2024. <https://doi.org/10.48550/arXiv.2407.14933>.
- 5 Reed, Rachel. « Does ChatGPT Violate New York Times' Copyrights? ». Harvard Law School. Consulté le 9 octobre 2024. <https://hls.harvard.edu/today/does-chatgpt-violate-new-york-times-copyrights/>.
- 6 Panettieri, Joe. « Generative AI Lawsuits Timeline: Legal Cases vs. OpenAI, Microsoft, Anthropic, Nvidia, Perplexity, Intel and More ». Sustainable Tech Partner for Green IT Service Providers, 23 janvier 2025. <https://sustainabletechpartner.com/topics/ai/generative-ai-lawsuit-timeline/>.
- 7 spawning.ai. « Data Governance for Generative AI ». Consulté le 5 septembre 2024. <https://spawning.ai/>.
- 8 Heikkilä, Melissa. « This New Data Poisoning Tool Lets Artists Fight Back against Generative AI ». MIT Technology Review. Consulté le 14 janvier 2025. <https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/>.
- 9 Villalobos, Pablo. « Will We Run Out of ML Data? Evidence From Projecting Dataset Size Trends ». Epoch AI, 10 novembre 2022. <https://epochai.org/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset>.
- 10 Sawers, Paul. « How Advex Creates Synthetic Data to Improve Machine Vision for Manufacturers ». *TechCrunch* (blogue), 28 octobre 2024. <https://techcrunch.com/2024/10/28/how-advex-creates-synthetic-data-to-improve-machine-vision-for-manufacturers/>.
- 11 Harris, Steve. « The Rising Role of Synthetic Data in the Automotive Industry ». *Automotive Testing Technology International* (blogue), 5 juin 2023. <https://www.automotivetestingtechnologyinternational.com/industry-opinion/the-rising-role-of-synthetic-data-in-the-automotive-industry.html>.
- 12 Gartner. « Is Synthetic Data the Future of AI? ». Consulté le 14 janvier 2025. <https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai>.
- 13 « Synthetic Data: What Operational Privacy Professionals Need to Know ». IAPP. Consulté le 15 janvier 2025. <https://iapp.org/news/a/synthetic-data-what-operational-privacy-professionals-need-to-know>.
- 14 Commissariat à la protection de la vie privée du Canada. « Blogue savoir techno : Faire du neuf avec du vieux – la réalité des données synthétiques ». 12 octobre 2022. <https://www.priv.gc.ca/fr/blogue/20221012/>.
- 15 Gouvernement du Canada, Instituts de recherche en santé du Canada. « Engagement A – Séquencer l'avenir : Plan stratégique de l'institut de génétique 2022-2027 – IRSC », 9 mai 2022. <https://cihr-irsc.gc.ca/f/52973.html>.
- 16 Huerta, E.A., Ben Blaiszik, L. Catherine Brinson, Kristofer E. Bouchard, Daniel Diaz, Caterina Doglioni, Javier M. Duarte et coll. « FAIR for AI: An Interdisciplinary and International Community Building Perspective ». *Scientific Data*, vol. 10, no 1 (26 juillet 2023), article 487. <https://doi.org/10.1038/s41597-023-02298-6>.
- 17 « Data Annotation and Labelling Market Is Expected to Generate Revenue of USD 29,584.2 Mn by 2033, at 34.4% CAGR: Dimension Market Research ». Yahoo Finance, 4 septembre 2024. <https://finance.yahoo.com/news/data-annotation-labelling-market-expected-144500834.html>.

- 18 Cisco. « Cisco AI Readiness Index ». Consulté le 23 août 2024. https://www.cisco.com/c/m/en_us/solutions/ai/readiness-index.html.
- 19 Mehmood, Hamid. « Data Drought in the Global South ». Our World. Consulté le 10 octobre 2024. <https://ourworld.unu.edu/en/data-drought-in-the-global-south>.
- 20 Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoochi et Richard G. Baraniuk. « Self-Consuming Generative Models Go MAD ». arXiv, 4 juillet 2023. <https://doi.org/10.48550/arXiv.2307.01850>.
- 21 Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson et Yarin Gal. « AI Models Collapse When Trained on Recursively Generated Data ». *Nature*, vol. 631, no 8022 (juillet 2024), p. 755-759. <https://doi.org/10.1038/s41586-024-07566-y>.
- 22 Reed, Rachel. « Does ChatGPT Violate New York Times' Copyrights? ». Harvard Law School. Consulté le 9 octobre 2024. <https://hls.harvard.edu/today/does-chatgpt-violate-new-york-times-copyrights/>.
- 23 ServiceNow Research. « Announcing StarCoder2 and The Stack V2 ». Consulté le 15 janvier 2025. <https://www.servicenow.com/blogs/2024/announcing-starcoder2-stack-v2>.
- 24 Data & Trust Alliance. « The Data & Trust Alliance ». Consulté le 7 août 2024. <https://dataandtrustalliance.org/>.
- 25 « Overview – Coalition for Content Provenance and Authenticity ». Consulté le 5 septembre 2024. <https://c2pa.org/>.
- 26 Schuman, Leslie. « Leading Corporations Introduce Data Provenance Standards ». Businesswire, 30 novembre 2023. <https://www.businesswire.com/news/home/20231130851266/en/Leading-Corporations-Introduce-Data-Provenance-Standards>.
- 27 Heikkilä, Melissa. « Why Big Tech's Watermarking Plans Are Some Welcome Good News ». MIT Technology Review. Consulté le 5 septembre 2024. <https://www.technologyreview.com/2024/02/13/1088103/why-big-techs-watermarking-plans-are-some-welcome-good-news/>.
- 28 Longpre, Shayne, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Jad Kabbara et Alex Pentland. « How to Fix Data Authenticity, Data Consent & Data Provenance for AI ». MIT Initiative on the Digital Economy Policy Brief 2024, vol. 3. https://ide.mit.edu/wp-content/uploads/2024/06/PB_6-21-24.pdf?x21090.
- 29 « What Is Federated Learning? ». IBM Research. Consulté le 5 septembre 2024. <https://research.ibm.com/blog/what-is-federated-learning>.
- 30 Hagemann, Johannes, Sami Ja, Jack Minong, et Vincent Weisser. « INTELLECT-1: Launching the First Decentralized Training of a 10B Parameter Model ». Consulté le 16 octobre 2024. <https://www.primeintellect.ai/blog/intellect-1>.
- 31 Eid Kishawy, Mohab M., Mohamed T. Abd El-Hafez, Retaj Yousri et M. Saeed Darweesh. « Federated Learning System on Autonomous Vehicles for Lane Segmentation ». *Scientific Reports*, vol. 14, no 1 (23 octobre 2024), article 25029. <https://doi.org/10.1038/s41598-024-71187-8>.
- 32 Wang, Ran, Cheng Xu, Shuhao Zhang, Fangwen Ye, Yusen Tang, Sisui Tang, Hangning Zhang, Wendi Du et Xiaotong Zhang. « MatSwarm: Trusted Swarm Transfer Learning Driven Materials Computation for Secure Big Data Sharing ». *Nature Communications*, vol. 15, no 1 (28 octobre 2024), article 9290. <https://doi.org/10.1038/s41467-024-53431-x>.
- 33 Thakur, Anshul, Soheila Molaei, Pafue Christy Nganjimi, Andrew Soltan, Patrick Schwab, Kim Branson et David A. Clifton. « Knowledge Abstraction and Filtering Based Federated Learning over Heterogeneous Data Views in Healthcare ». *Npj Digital Medicine*, vol. 7, no 1 (16 octobre 2024), p. 1-14. <https://doi.org/10.1038/s41746-024-01272-9>.
- 34 Ellencweig, Ben, Guilherme Cruz, et Vishnu Kamalnath. « Intelligence at Scale: Data Monetization in the Age of Gen Ai ». McKinsey & Company. Consulté le 25 août 2025. <https://www.mckinsey.com/capabilities/business-building/our-insights/intelligence-at-scale-data-monetization-in-the-age-of-gen-ai>.

- 35 Science|Business. « AI-Assisted Drug Discovery Held Back by Private Sector Secrecy on Datasets ». Consulté le 10 octobre 2024. <https://sciencebusiness.net/news/ai-assisted-drug-discovery-held-back-private-sector-secrecy-datasets>.
- 36 Fairly Trained. « Fairly Trained ». Consulté le 15 janvier 2025. <https://www.fairlytrained.org>.
- 37 Iper, Alexandra, Kanishka Singh et Kanishka Singh. « Biden Cracks down on US Data Flows to China, Russia ». Reuters, 28 février 2024, rubrique Technology. <https://www.reuters.com/technology/biden-crack-down-us-data-flows-china-russia-2024-02-28/>.
- 38 « Règlement sur les marchés numériques : garantir des marchés numériques équitables et ouverts ». Commission européenne. Consulté le 10 décembre 2024. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_fr.
- 39 Lawler, Richard. « Apple May Delay AI Features in the EU Because of Its Big Tech Law ». The Verge, 21 juin 2024. <https://www.theverge.com/2024/6/21/24183251/apple-eu-delay-ai-screen-mirroring-shareplay-dma>.

© 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 : PH4-229/2025F-PDF

ISBN : 978-0-660-78233-1

Also available in English under the title: The AI-driven data race.

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.