



INSIGHT 19

THE AI-DRIVEN DATA RACE



An AI-driven race for data is driving up the value of data, incentivizing new tactics to acquire data, new ways to generate data – including synthetic data – and new attempts to impose legal and technical restrictions on how data can be used.

PRESENT

AI companies are using ethically concerning tactics to get data for training their AI models. Unprecedented levels of investment in AI are creating pressure on companies to accelerate the development of AI models.¹ This requires good quality data, which is in increasingly short supply.² Some social media platforms are rewriting their terms of service to allow content generated by their users to be used in training AI models. Some AI companies are scraping data from websites without permission (see Box 1).

Box 1

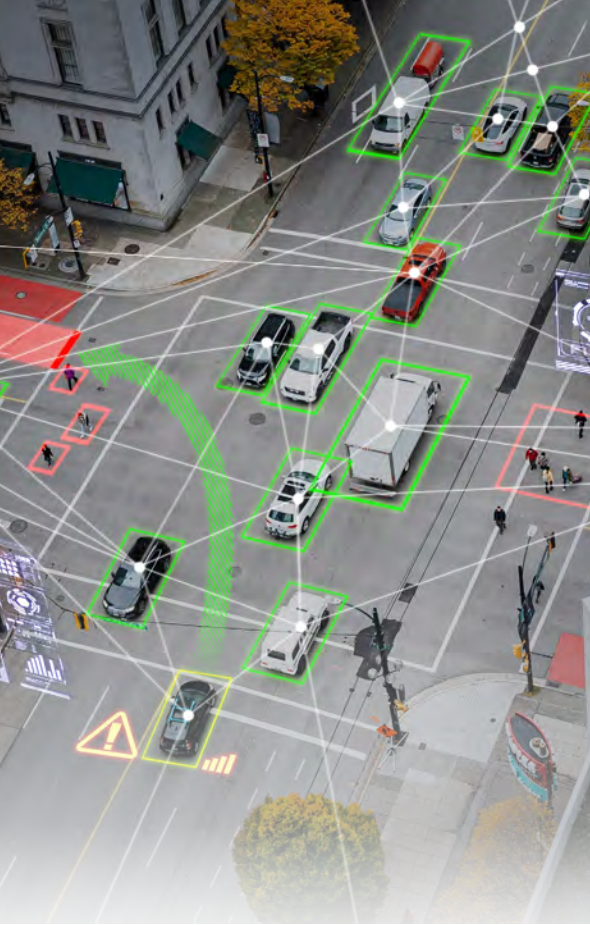
Web scraping is the automated collection of data from websites by programs called web crawlers. Web standards allow websites to tag data with a “do not scrape” instruction. In practice, however, crawlers may or may not abide by these instructions. Despite criticism, some AI firms are scraping personal or copyrighted content for use in training AI models.



More websites and content creators are trying to prevent AI companies from using their data. More websites are changing their terms of service – though such changes may be difficult to enforce, and do not apply retroactively when data has already been scraped.³ One recent audit found that 45% of the data in a Google training dataset is restricted by websites’ terms of service.⁴ Many lawsuits before the courts are seeking to restrict use of copyrighted material in AI training.⁵⁶ Technological tools are also being deployed: for example, Spawning’s Kudurru blocks automated web scraping, while Nightshade “poisons” data for AI training purposes without affecting how it appears to humans.⁷⁸

Synthetic data is increasingly of interest for AI training. Finding it harder to access human-generated data, more AI companies are exploring the use of AI-generated data to train AI models.⁹ Generative AI models are being designed to rapidly create synthetic data that can be used to train more capable models.¹⁰ The use of synthetic data is still in its early stages. Today it is more commonly used for specific use cases, such as training autonomous vehicles, but some experts suggest it could become more important than real data by 2030.¹¹¹² However, it has some challenges. It is most useful when generated from real data, but even if so, it can oversimplify real-world nuance or exaggerate biases found in sample datasets.¹³ It can also be vulnerable to reidentification attacks, revealing personal information contained in sample datasets used to build the synthetic dataset.¹⁴

Scientific advances are creating new opportunities to generate data in technical domains that could be used for AI training. For example, the DNA in a human genome can now be sequenced in a single day, compared to a decade at the turn of the century.¹⁵ Efforts are underway to make more scientific datasets findable, accessible, interoperable, and reusable (FAIR) for use by AI researchers.¹⁶



FUTURES

The value of *real* data could continue to grow, driving investment in data collection and labelling. The data annotation and labelling market is projected to grow by nearly 600% by 2030.¹⁷ Organizations could focus more on making the data they gather “AI ready” so it can more easily be monetized.¹⁸ The incentive to collect more data could reduce inequality – for example, leading to new data collection efforts in remote regions or the Global South, which in turn could improve development of services for these groups.¹⁹ Individuals could be offered new incentives to share their data, such as access to digital services or discounts on devices that generate data.

Overreliance on AI-generated synthetic data could lead to “model collapse”. If AI models are increasingly trained on synthetic data that is of lower quality than real data, this could cause a self-reinforcing degradation of AI models.²⁰ As synthetic content comes to dominate the Internet, the risk of model collapse could grow.²¹

Legal rulings could restrict the usefulness of AI models. If courts rule that copyrighted data cannot be used for training, some AI models could need to be removed and retrained.²² New models trained only on data obtained with explicit permission could have narrower capabilities. For example, StarCoder2 is an AI model trained exclusively on content shared by code developers who opted in, and is less capable than ChatGPT.²³

Box 2

Types of bias in datasets

Historical bias results from data reflecting past or existing inequalities

Sampling bias results from certain groups being underrepresented in data

Labelling bias results from those labelling the data imposing their subjective views

Measurement bias results from data collection methods that favour certain results

Confirmation bias results from collectors selecting data that confirms their beliefs

New ways of restricting data use could be coded into the data itself. With conventions around web scraping no longer seeming sufficient, new approaches could emerge that change how AI agents interact with data online. Companies including Meta, IBM,²⁴ Google and Microsoft²⁵ are proposing a data provenance standard²⁶ and a technical standard²⁷ that could add watermarks to data. Coding data-use instructions within data could improve transparency, helping developers to avoid copyright issues or bias (see Box 2).²⁸ It could also increase pressure on AI developers to find new data sources, or ways to circumvent the new restrictions.

Alternative AI training techniques could give users greater privacy and control. Federated learning is a technique that does not rely on a central repository of data, but draws from distributed sources in a way that maintains privacy and gives data owners more control over access rights.²⁹ Such decentralised models of AI training³⁰ might encourage more individuals and organisations to allow access to their data, potentially generating more insights in areas such as self-driving cars,³¹ materials research³² and healthcare.³³

IMPLICATIONS

- ▶ Failing to solve the data bottleneck **could slow AI innovation**, raising the cost of training data and cooling investor interest.
- ▶ New **financial incentives for data sharing** could affect how individuals and organisations behave:
 - › Vulnerable groups could be exploited for their data.
 - › Maintaining privacy could become harder or more expensive.
 - › Research initiatives could opt to monetize their datasets³⁴ rather than sharing them openly, undermining innovation and science.³⁵
- ▶ Systems using AI models that are trained on synthetic data may be vulnerable to model collapse, degrading their performance and **reducing public trust in AI** and institutions that have integrated AI into their services.
- ▶ If court judgements restrict lawful access to **powerful AI models** that were trained on copyrighted material, cyber criminals and others willing to continue using these models illegally could gain a tactical advantage.
- ▶ **Standards and certifications** like Fairly Trained, which identify datasets that are open-source, ethically trained, and infringement-free, could **become more influential**.³⁶
- ▶ Countries might **compete for access to data** and increasingly view their own data as a strategic resource to be protected and controlled. This could undermine global cooperation in areas such as science.³⁷
- ▶ Countries could adopt **varying data standards and protocols** that lead to fragmentation of markets. For example, the European Union's Digital Markets Act – which requires technology companies to ensure interoperability and access to certain types of data generated on their platforms³⁸ – has led to Apple deciding not to roll-out its Apple Intelligence AI feature in the EU.³⁹

Endnotes

- 1 Kinding, Beth. 'Big Tech Battles On AI, Here's The Winner'. Forbes. Accessed 8 October 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 November 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 July 2024, sec. 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 al. 'Consent in Crisis: The Rapid Decline of the AI Data Commons'. arXiv, 24 July 2024. <https://doi.org/10.48550/arXiv.2407.14933>.
- 5 Reed, Rachel. 'Does ChatGPT Violate New York Times' Copyrights?' Harvard Law School. Accessed 9 October 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 January 2025. <https://sustainabletechpartner.com/topics/ai/generative-ai-lawsuit-timeline/>.
- 7 spawning.ai. 'Data Governance for Genrative AI'. Accessed 5 September 2024. <https://spawning.ai/>.
- 8 Heikkilä, Melissa. 'This New Data Poisoning Tool Lets Artists Fight Back against Generative AI'. MIT Technology Review. Accessed 14 January 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 November 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* (blog), 28 October 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* (blog), June 5, 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?' Accessed 14 January 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'. Accessed 15 January 2025. <https://iapp.org/news/a/synthetic-data-what-operational-privacy-professionals-need-to-know>.
- 14 Canada, Office of the Privacy Commissioner of. 'Privacy Tech-Know Blog: When What Is Old Is New Again – The Reality of Synthetic Data', 12 October 2022. <https://www.priv.gc.ca/en/blog/20221012/>.
- 15 Government of Canada, Canadian Institutes of Health Research. 'Commitment A - Sequencing Our Future: Institute of Genetics Strategic Plan 2022-2027 - CIHR', 9 May 2022. <https://cihr-irsc.gc.ca/e/52973.html>.
- 16 Huerta, E. A., Ben Blaiszik, L. Catherine Brinson, Kristofer E. Bouchard, Daniel Diaz, Caterina Doglioni, Javier M. Duarte, et al. 'FAIR for AI: An Interdisciplinary and International Community Building Perspective'. *Scientific Data* 10, no. 1 (26 July 2023): 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 September 2024. <https://finance.yahoo.com/news/data-annotation-labelling-market-expected-144500834.html>.
- 18 Cisco. 'Cisco AI Readiness Index'. Accessed 23 August 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'. Accessed 10 October 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 Siahkoobi, and Richard G. Baraniuk. 'Self-Consuming Generative Models Go MAD'. arXiv, 4 July 2023. <https://doi.org/10.48550/arXiv.2307.01850>.
- 21 Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 'AI Models Collapse When Trained on Recursively Generated Data'. *Nature* 631, no. 8022 (July 2024): 755–59. <https://doi.org/10.1038/s41586-024-07566-y>.
- 22 Reed, Rachel. 'Does ChatGPT Violate New York Times' Copyrights?' Harvard Law School. Accessed 9 October 2024. <https://hls.harvard.edu/today/does-chatgpt-violate-new-york-times-copyrights/>.
- 23 Research, ServiceNow. 'Announcing StarCoder2 and The Stack V2'. Accessed 15 January 2025. <https://www.servicenow.com/blogs/2024/announcing-starcoder2-stack-v2>.
- 24 The Data & Trust Alliance. 'The Data & Trust Alliance'. Accessed 7 August 2024. <https://dataandtrustalliance.org/>.
- 25 'Overview - Coalition for Provenance and Authenticity'. Accessed 5 September 2024. <https://c2pa.org/>.
- 26 Schuman, Leslie. 'Leading Corporations Introduce Data Provenance Standards'. *Businesswire*, 30 November 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. Accessed 5 September 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 and 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'. Accessed 5 September 2024. <https://research.ibm.com/blog/what-is-federated-learning>.
- 30 Hagemann, Johannes, Sami Ja, Jack Minong, and Vincent Weisser. 'INTELLECT–1: Launching the First Decentralized Training of a 10B Parameter Model'. Accessed 16 October 2024. <https://www.primeintellect.ai/blog/intellect-1>.
- 31 Eid Kishawy, Mohab M., Mohamed T. Abd El-Hafez, Retaj Yousri, and M. Saeed Darweesh. 'Federated Learning System on Autonomous Vehicles for Lane Segmentation'. *Scientific Reports* 14, no. 1 (23 October 2024): 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, and Xiaotong Zhang. 'MatSwarm: Trusted Swarm Transfer Learning Driven Materials Computation for Secure Big Data Sharing'. *Nature Communications* 15, no. 1 (28 October 2024): 9290. <https://doi.org/10.1038/s41467-024-53431-x>.
- 33 Thakur, Anshul, Soheila Molaei, Pafue Christy Nganjimi, Andrew Soltan, Patrick Schwab, Kim Branson, and David A. Clifton. 'Knowledge Abstraction and Filtering Based Federated Learning over Heterogeneous Data Views in Healthcare'. *Npj Digital Medicine* 7, no. 1 (16 October 2024): 1–14. <https://doi.org/10.1038/s41746-024-01272-9>.
- 34 Ellencweig, Ben, Guilherme Cruz, and Vishnu Kamalnath. 'Intelligence at Scale: Data Monetization in the Age of Gen AI'. McKinsey & Company. Accessed 25 August 2025. <https://www.mckinsey.com/capabilities/business-building/our-insights/intelligence-at-scale-data-monetization-in-the-age-of-gen-ai>.

- 35 SciencelBusiness. 'AI-Assisted Drug Discovery Held Back by Private Sector Secrecy on Datasets'. Accessed 10 October 2024. <https://sciencebusiness.net/news/ai-assisted-drug-discovery-held-back-private-sector-secrecy-datasets>.
- 36 Fairly Trained. 'Fairly Trained'. Accessed 15 January 2025. <https://www.fairlytrained.org>.
- 37 Iper, Alexandra, Kanishka Singh, and Kanishka Singh. 'Biden Cracks down on US Data Flows to China, Russia'. *Reuters*, 28 February 2024, sec. Technology. <https://www.reuters.com/technology/biden-crack-down-us-data-flows-china-russia-2024-02-28/>.
- 38 'The Digital Markets Act: Ensuring Fair and Open Digital Markets - European Commission'. Accessed 10 December 2024. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en.
- 39 Lawler, Richard. 'Apple May Delay AI Features in the EU Because of Its Big Tech Law'. *The Verge*, 21 June 2024. <https://www.theverge.com/2024/6/21/24183251/apple-eu-delay-ai-screen-mirroring-shareplay-dma>.

© His Majesty the King in Right of Canada, 2025

For information regarding reproduction rights: <https://horizons.gc.ca/en/contact-us/>

PDF: PH4-229/2025E-PDF

ISBN: 978-0-660-78232-4

Aussi disponible en français sous le titre : La course aux données axée sur l'IA.

DISCLAIMER

Policy Horizons Canada (Policy Horizons) is the Government of Canada's centre of excellence in foresight. Our mandate is to empower the Government of Canada with a future-oriented mindset and outlook to strengthen decision making. The content of this document does not necessarily represent the views of the Government of Canada, or participating departments and agencies.