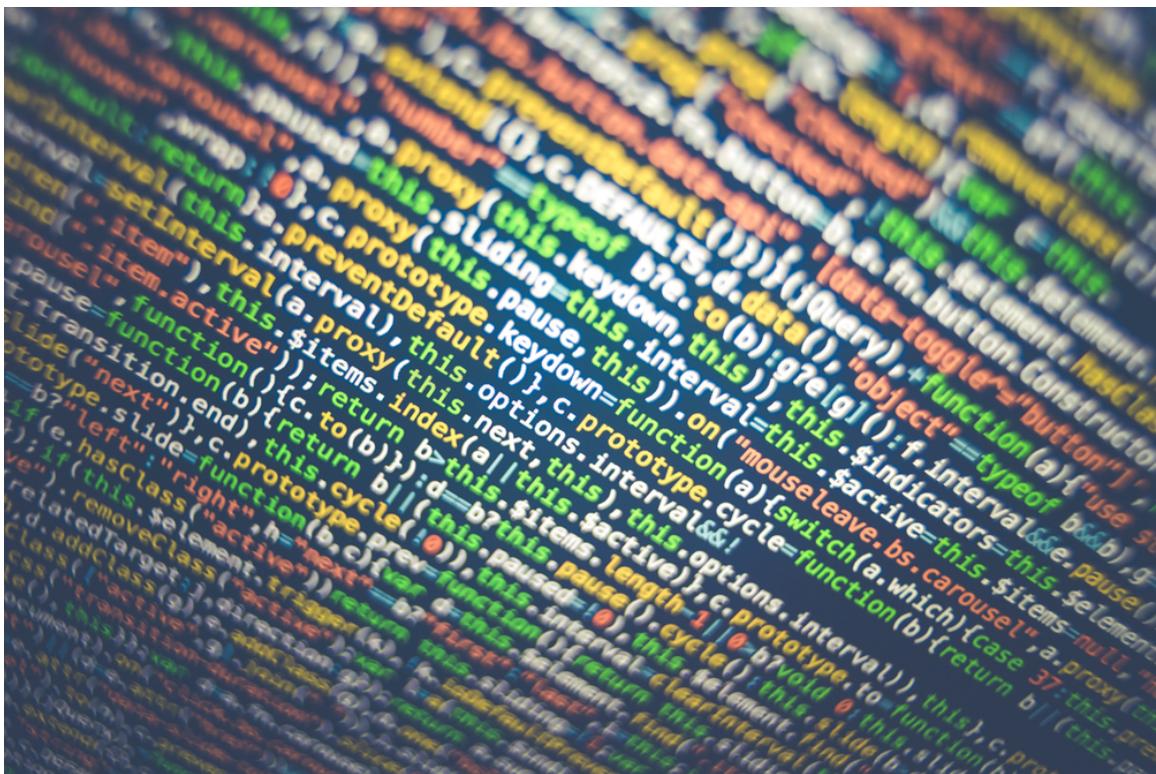




AI Safety, Cybersecurity, and Open Scholarship

by Britt Amell | 15 September 2025 | English, Insights and Signals Reports



This insights and signals report was written by Brittany Amell, with thanks to INKE Partners Aaron Mauro and James MacGregor for their review and comments.

Lisez-le en français

At a Glance

Topic area	AI Safety, Cybersecurity, and Open Scholarship
Key participants	Policy Horizons Canada, Canadian Centre for Cyber Security, the AI Safety Institute, Global Affairs Canada, the University of New Brunswick
Timeframe	2025
Keywords	AI safety / Sécurité de l'IA; AI governance / Gouvernance de l'IA; Cybersecurity / Cybersécurité; CRKN / RCDR; AI bots / Robots d'indexation IA; open access / libre accès; open infrastructure / infrastructure ouverte; open science / science ouverte; open social scholarship / approches sociales des savoirs ouverts; Canada

Summary

This insights and signals report discusses the Foresight on AI report recently released by Policy Horizons Canada, as well as reports from the Canadian Centre for Cyber Security (the **National Cyber Threat Assessment 2025-2026**) and the AI Safety Institute (the **International Scientific Report on the Safety of Advanced AI**, also known as the International AI Safety Report). It also shares some AI and cybersecurity related news from **Global Affairs Canada**, the **National Cybersecurity Consortium**, and the University of New Brunswick—which is set to receive **\$10 million dollars over the next five years in order to establish the Cyber Attribution Data Centre**.

A number of the reports and news items included provide a wide array of information, enthusiasms, concerns and admonishments with respect to the state of AI in Canada and globally; the intent of this report is to extract and synthesize the safety and cybersecurity implications across all items.

Foresight on AI report from the Government of

Canada

Policy Horizons Canada, the Government of Canada's centre for policy foresight, released its much-anticipated **Foresight on AI report**, in which it adopts the Organization for Economic Cooperation and Development (OECD) definition of AI as a

machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment. (4)

The report, which was released at the beginning of February 2025, presents 10 insights into the future “of and with” AI. Insights include AI's potential to power cyber threats, “break the internet as we currently know it” (5), stay biased forever (“due to conflicting perspectives on fairness,” 5), reshape how we relate to each other, and become even more prevalent in the lives of children, reshaping their “lives in the present and future” (6).

Cybersecurity

Each of the insights discussed in the foresight on AI report have wide-ranging implications and pick up on trends flagged by the Canadian Centre for Cyber Security in their **National Cyber Threat Assessment 2025-2026** report. Notably, the National Cyber Threat Assessment report points to the capacity of AI to lower barriers to engagement when it comes to malicious cyber activity (32). As INKE partner Aaron Mauro (Associate Professor, Digital Media and Chair, Department of Digital Humanities at Brock University) succinctly puts it:

Hackers can act faster, with less training, and across languages with ease. They can potentially develop new species of malware, which may be harder to detect or mitigate. They can even automate attacks and access in ways not easily predictable by human motivations, which again makes them hard to defend and anticipate.

Critical infrastructure, the report also notes, is particularly attractive as a target for

threat actors because “these entities are perceived as being more willing to pay large ransoms to prevent disruptions to critical operations” (25).

The **National Cybersecurity Consortium**—founded by five Canadian universities in 2020—has **announced** that 37 Canadian projects related to the advancement of Canadian cybersecurity will be receiving \$22.8 million dollars, bringing investment totals in Canadian cybersecurity to over \$60 million. Critical infrastructure features prominently across the **range of funded projects**.

Mélanie Joly, Former Canadian Minister of Foreign Affairs, announced in February that **Global Affairs Canada** will be providing \$1.8 million in funding to a project delivering targeted digital security support headed by **Access Now**. Digital safety is an increasing concern for many around the world, as intimidation, harassment, disinformation, technological attacks and surveillance threats continue to increase and diversify, fueled by AI (**Canadian Centre for Cyber Security 2024**; **Global Affairs Canada 2025**).

We can also expect to see more from the University of New Brunswick, who is to receive **\$10 million over the next five years in order to establish the Cyber Attribution Data Centre**. A key aim of the centre will be to “train and equip the next generation of artificial intelligence (AI) cyber attribution specialists,” and “establish AI-powered data services to report findings and intelligence to vital government stakeholders” (**McLaughlin 2024**).

Lastly, recently elected Prime Minister Mark Carney **announced** Canada’s first Minister of Artificial Intelligence and Digital Innovation, Evan Solomon. Also responsible for the Federal Economic Development Agency for Southern Ontario, Solomon is the Member of Parliament for Toronto Centre. It’s not clear yet what Solomon’s mandate will be, though a **source quoted by the CBC** has suggested the Liberal party’s platform (“**Canada Strong**”) offers a good idea (see also **Tunney 2025** for an overview).

International AI Safety Report

Early 2025 has also seen the release of the **International Scientific Report on the Safety of Advanced AI** (also known as the International AI Safety Report) was

released, offering the first consolidated, global assessment of the safety risks posed by frontier AI systems. The report brings together insights from over 100 expert representatives nominated by 30 countries, the OECD, the EU, and the UN, as well as other world-leading experts:

For the first time in history, this report . . . brought together expert representatives . . . to provide a shared scientific, evidence-based foundation for these vital discussions. We continue to disagree on several questions, minor and major, around general-purpose AI and its capabilities, risks, and risk mitigations. However, we consider this report essential for improving our collective understanding of general-purpose AI and its potential risks, and for moving closer towards consensus and effective risk mitigation, to ensure that humanity can enjoy the benefits of general-purpose AI safely. The stakes are high. We look forward to continuing this effort. (215)

The 298-page report aims to overview the current state of research on AI in order to facilitate discussions, policy and otherwise. It focuses on “general-purpose” AI, which is AI that can perform a wide range of tasks (26)—not to be confused with artificial general intelligence, which describes a “selective, human-like, broad, flexible kind of intelligence that can do all kinds of learning and perform all kinds of tasks” (Plumb 2022).

The authors of the report further distinguish between an “AI model” (a model that can be adapted to perform a variety of tasks) and an “AI system” (a combination of components that is designed to be useful to humans, and include one or more AI models). The model GPT-4, for instance, could be thought of as the engine that drives ChatGPT, an AI system (26-27). The report does not address narrow AI (a type of AI trained to perform a specific and narrow range of tasks), though it notes that this type can also pose significant harms as well.

There are three key sections in the report. The first section usefully summarizes the capabilities of general-purpose AI, including how it is developed and anticipated capabilities to come. It opens with a set of definitions and also explains the process of developing general-purpose AI. “There are many different types of general-purpose AI,” it reads, “but they are developed using common methods and principles” (32).

The second (and longest) section summarizes the risks associated with AI use and is

divided into four sub-sections. The first sub-section focused on risks from the malicious use of AI (such as harm to individuals through fake content, the manipulation of public opinion, cyber offences, and biological and chemical attacks). The second sub-section considers risks associated with malfunctions, including reliability issues, bias, and a loss of control. The third sub-section considers systemic risks, which include risks to the labour market, global divides in AI research and development, market concentration (and single points of failure), environmental risks, and risks to privacy as well as copyright infringement. The last sub-section focuses on the impact of “open-weight” models, those with publicly available “weights” – where weights refer to the parameters that represent how strongly nodes are connected in a network. Weights play a role in how a model responds to an input. Open-weight models *may* be open-source models (available for public download), but they aren’t always.

The report notes several different open release options. These are:

- **Fully open models** are open-source models “for which weights, full code, training data, and other documentation (e.g. about the model’s training process) are made publicly available, without restrictions on modification, use and sharing.” (151)
- **Fully closed models** are those which “weights and code are proprietary, for internal use only” (151)
- **Partially open models** are those which share “some combination of weights, code, and data under various licences or access controls, in an attempt to balance the benefits of openness against risk mitigation and proprietary concerns” (151)

The final section of the AI Safety report focuses on approaches to risk management, which the authors define as “The systematic process of identifying, evaluating, mitigating and monitoring risks” (158). The first sub-section provides an overview of risk management that manages to be brief and comprehensive, and is a good resource for those who may be new to risk management and its various approaches. This final section further describes the challenges for policymaking and risk management, and ends with some thoughts on the challenges of risk monitoring and mitigation. Unfortunately, as the authors note, currently there is no single or perfect safety measure available. However, they point out that “defence in depth,” which they define as “having multiple layers of protection and redundant safeguards” in

place, can increase “confidence in safety” (205).

The report ends with essentially five conclusions, as well as a list of acronyms and glossary. The core conclusions are:

- The first International AI Safety Report finds that the future of general-purpose AI is remarkably uncertain. (214)
- To reap the benefits of this transformative technology safely, researchers and policymakers need to identify the risks that come with it and take informed action to mitigate them. (214)
- There exist technical methods for addressing the risks of general-purpose AI, but they all have limitations. (214)
- AI does not happen to us; choices made by people determine its future. (214)
- For the first time in history, this report and the Interim Report (May 2024) brought together expert representatives nominated by 30 countries, the OECD, the EU, and the UN, as well as several other world-leading experts, to provide a shared scientific, evidence-based foundation for these vital discussions. (215)

Key Considerations and Questions for Further Discussion

Both the recent Policy Horizons Canada foresight report and the International AI Safety Report underscore the wide-ranging societal implications of AI—from reshaping digital environments and human relationships to entrenching algorithmic bias and enabling malicious cyber activity.

Meanwhile, a number of policy voices around the world are increasingly calling for AI safety knowledge and governance tools to be treated as global public goods (e.g., [Blomquist et al. 2025](#)). This framing offers both an opportunity and a challenge for open scholarship which has rigorously engaged in discussions relating to public goods and the commons (see, for instance, [Arbuckle et al. 2022](#); [Fitzpatrick 2024](#); [Fitzpatrick 2024a](#); Willinsky 2009).

Further, the asymmetries between AI-developing countries and those primarily affected by AI implementation mirror longstanding challenges with knowledge inequity ([Blomquist et al. 2025](#); [Ma 2024](#); [Pooley 2024](#)). While open scholarship has

aimed to respond to and bridge these challenges, general purpose artificial intelligence models and tools—along with policy and regulation efforts—introduce new complexities and risks.

One such example of the complexity open infrastructure and open scholarship journals are facing includes increased—at times debilitating—access by automated bots (sometimes referred to as crawlers) mining sites for training data (Confederation of Open Access Repositories (COAR) 2025; Decker 2025; Hinchliffe 2025). As the “The impact of AI bots and crawlers on open repositories: Results of a COAR survey” report indicates, open access repositories are being accessed day and night by automated bots, causing noticeable slowdowns and system crashes. James MacGregor, Director of Research Infrastructure and Development at the Canadian Research Knowledge Network, understands first-hand the risks of this kind of activity:

As caretakers of the technical infrastructure that serves the **Canadiana** and **Héritage** collections, collectively consisting of 65M images of Canadian documentary history, we are especially concerned about problematic access to large-scale datasets. One example occurred in 2023 when access to Internet Archive was briefly interrupted globally due to a poorly-behaving bot **attempting to bulk-ingest its OCR data**. (This service interruption pales in comparison and intent to the outright malicious cyberattack **suffered by Internet Archive in Fall 2024**.)

Preserving values of openness and accessibility while also navigating new technological complexities and risks, as well as persistent inequities at the intersection of AI safety, cybersecurity, and open scholarship will be a challenge—one open scholarship practitioners, advocates, and scholars are well-positioned to meet.

References

Arbuckle, Alyssa, Ray Siemens, Jon Bath, Constance Crompton, Laura Estill, Tanja Niemann, Jon Saklofske, and Lynne Siemens. 2022. "An Open Social Scholarship Path for the Humanities." *The Journal of Electronic Publishing* 25 (2). <https://doi.org/10.3998/jep.1973>.

Bengio, Yoshua, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, et al. 2025. "International AI Safety Report." DSIT 2025/001. The International Scientific Report on the Safety of Advanced AI. Department for Science, Innovation and Technology (UK) and AI Safety Institute (UK). <https://www.gov.uk/government/publications/international-ai-safety-report-2025>.

Blomquist, Kayla, Elisabeth Siegel, Kwan Yee Ng, Tom David, Brian Tse, Charles Martinet, Matt Sheehan, et al. 2025. "Examining AI Safety as a Global Public Good: Implications, Challenges, and Research Priorities." Oxford Martin AI Governance Initiative, Concordia AI, and Carnegie Endowment for International Peace. <https://carnegieendowment.org/research/2025/03/examining-ai-safety-as-a-global-public-good-implications-challenges-and-research-priorities?lang=en>.

Canada, Global Affairs. 2025. "Minister Joly Announces Support for Digital Security of Human Rights Defenders." News Releases. February 12, 2025. <https://www.canada.ca/en/global-affairs/news/2025/02/minister-joly-announces-support-for-digital-security-of-human-rights-defenders.html>.

Fitzpatrick, Kathleen. 2024. "Open Infrastructures and the Future of Knowledge Production, Part 1." *Platypus: The Blog of the Humanities Commons Team* (blog). January 5, 2024. <https://team.hcommons.org/2024/01/05/open-infrastructures-and-the-future-of-knowledge-production-part-1/>.

Fitzpatrick, Kathleen. 2024. "Open Infrastructures and the Future of Knowledge Production, Part 2." *Platypus: The Blog of the Humanities Commons Team* (blog). January 8, 2024. <https://team.hcommons.org/2024/01/08/open-infrastructures-and-the-future-of-knowledge-production-part-2/>.

Ma, Lai. 2024. "Generative AI for Academic Publishing? Some Thoughts About Epistemic Diversity and the Pursuit of Truth." *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 7 (1): 1–5. <https://doi.org/10.18357/kula.287>.

McLaughlin, Kathleen. 2024. "Government of Canada Announces Funding for Cyber Attribution Data Centre at UNB to Advance National Cybersecurity." *University of New Brunswick News* (blog). December 13, 2024. <https://blogs.unb.ca/newsroom/2024/12/federal-funding.php>.

National Cybersecurity Consortium. 2024. "NCC Distributes \$22.8M in Funding to Advance Canadian Digital Security." *News* (blog). October 11, 2024. <https://ncc-cnc.ca/news-2024-funding-announcement/>.

Plumb, Taryn. 2022. "AI Act: What Does General Purpose AI (GPAI) Even Mean?" *VentureBeat* (blog). September 15, 2022. <https://web.archive.org/web/20241114222307/https://venturebeat.com/ai/ai-act-what-does-general-purpose-ai-gpai-even-mean/>

Policy Horizons Canada. 2025. "Foresight on AI: Policy Considerations." Ottawa: Government of Canada. <https://horizons.service.canada.ca/en/2025/02/10/ai-policy-consideration/index.shtml>.

Pooley, Jefferson. 2024. "Large Language Publishing: The Scholarly Publishing Oligopoly's Bet on AI." *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 7 (1): 1–11. <https://doi.org/10.18357/kula.291>.

Prime Minister of Canada. 2025. "Prime Minister Carney Announces New Ministry." News Releases. May 13, 2025. <https://www.pm.gc.ca/en/news/news-releases/2025/05/13/prime-minister-carney-announces-new-ministry>.

Tunney, Catherine. 2025. "Canada Now Has a Minister of Artificial Intelligence. What Will He Do?" *CBC News*, May 17, 2025, sec. Politics. <https://www.cbc.ca/news/politics/artificial-intelligence-evan-solomon-1.7536218>.

Willinsky, John. 2006. *The Access Principle: The Case for Open Access to Research and Scholarship*. Digital Libraries and Electronic Publishing. Cambridge, Mass.: MIT Press.

Search



Archives

Select Year

Categories

- Community News
- English
- French
- Insights and Signals Reports
- Observations
- Observations and Responses
- Policies
- Responses

Tags

- AI bots / Robots d'indexation IA
- AI governance / Gouvernance de l'IA
- AI safety / Sécurité de l'IA
- Berlin Declaration / Déclaration de Berlin
- Bethesda Statement / Déclaration de Bethesda
- biodiversity / biodiversité
- Budapest Statement / Déclaration de Budapest
- Canada
- Canadiana.org
- Canadian government/le gouvernement du Canada
- CAPOS
- CARL / ABRC
- collaboration
- community engagement / engagement communautaire
- Compute Canada / calcul Canada
- copyright / droits d'auteurs
- credibility / crédibilité
- CRKN / RCDR
- Cybersecurity / Cybersécurité

data management / gestion des données

diamond open access / le libre accès diamant digital scholarship / version numérique

en français / French English / en anglais

events and gatherings / événements et rassemblements

Federation for the HSS / Fédération des sciences humaines

funding agencies / organismes de financement

generative artificial intelligence / l'intelligence artificielle générative

identity management / gestion de l'identité implementation / mise en oeuvre INKE

international policy / politique internationale

licensing agreements / accords de licence Naylor Report / le rapport Naylor

open access / libre accès open data / données ouvertes

open education / éducation ouverte open government / gouvernement ouvert

open infrastructure / infrastructure ouverte open science / science ouverte

open social scholarship / approches sociales des savoirs ouverts

open source software / les logiciels libres ORCID peer review / critique des pairs

Perpetual Access / Accès perpétuel PKP Plan S

Plan S update / mise à jour du Plan S policy / politique

policy guide / guide des politiques promotion et titularisation publishing / édition

RDC / DRC RDM RECODE recommendations / recommandations

reports / les rapports repositories / les dépôts

research creation / recherche-crédation

research evaluation / l'évaluation de la recherche

research libraries / les bibliothèques de recherche

research output / les résultats de la recherche

research security / sécurité de la recherche RPT / révision

scholarly communication / la communication savante SFU Library / Bibliothèque

social media / les médias sociaux Tri-Agency / des trois organismes UK



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

