OSLER

ISED's Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems

OCTOBER 2, 2023 6 MIN READ

Related Expertise

- Artificial Intelligence
- Intellectual Property
- Privacy and Data Management
- Risk Management and Crisis Response
- Technology

Authors: Simon Hodgett, Sam Ip, Alannah Safnuk

On September 27, 2023, the Honourable François-Philippe Champagne, Minister of Innovation, Science and Industry, announced Canada's Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems (the Code of Conduct).

Recognizing the proliferation of innovative AI systems capable of generating content – such as ChatGPT, DALL·E 2, and Midjourney — the Code of Conduct is a set of voluntary commitments intended to commit developers and managers of advanced generative systems to take steps to identify and mitigate related risks. Innovation, Science and Economic Development Canada (ISED) has stated in the news release accompanying publication of the Code that it is intended to act as a "critical bridge between now" and when the proposed Artificial Intelligence and Data Act (AIDA), which was introduced as part of Bill C-27 in June 2022, comes into force.

The launch of the Code of Conduct follows a brief period during which an earlier draft was published. Notably, Osler hosted an AccessPrivacy workshop on August 30 designed to help organizations understand the scope, meaning and impact of the Code of Conduct on the earlier draft of the Code, which can be found here. A copy of this session was submitted to the Minister.

Key Features

The Code of Conduct is intended to apply to advanced generative systems, though the Code acknowledges that many of the measures are broadly applicable, including on a range of high-impact AI systems.

Unlike the earlier draft, the final version of the Code of Conduct distinguishes between measures applicable to all advanced generative AI systems and those applicable to advanced generative AI systems available for public use, which may give rise to a greater risk of potentially harmful or inappropriate use. For that reason, the Code of Conduct suggests that additional measures should apply in such instances. In addition, the final version of the Code of Conduct applies to specific actors, referred to as developers and managers, recognizing

that such participants in the AI ecosystem have different responsibilities.

Moreover, the final version of the Code of Conduct provides greater detail with respect to what measures are necessary for achieving the following six outcomes and who within a firm is responsible for ensuring those measures are undertaken, which are summarized as follows.

1. Accountability

The Code of Conduct requires firms to implement a clear risk management framework that is proportionate to the scale and impact of their activities. The Code of Conduct commits developers and managers to, among other measures, implement comprehensive risk management frameworks, which includes policies, procedures, and training to ensure staff understand their duties and the organization's risk management practices. Firms also commit to sharing information and best practices on risk management with other firms playing complementary roles in the ecosystem. In addition, developers at firms whose AI systems are available for public use commit to employing multiple lines of defence, including conducting third-party audits, to secure the safety of their AI systems prior to release.

2. Safety

The Code of Conduct highlights the importance of risk assessments and mitigations in support of safe operation of AI systems prior to deployment. Measures identified in the Code of Conduct to ensure the safety of AI systems include the performance of comprehensive risk assessments of reasonably foreseeable potential adverse impacts, a commitment by developers of AI systems to implement proportionate measures to mitigate risks of harm and for developers to make available guidance on appropriate system usage to downstream developers and managers.

3. Fairness and equity

The Code of Conduct recognizes AI systems have the potential to adversely impact fairness and equity, such as by perpetuating biases, and encourages assessment and mitigation at different phases in the development and deployment of systems. To this effect, the Code of Conduct commits developers to assess and curate datasets used for training to manage data quality and potential biases. Developers also commit to implement diverse testing methods and measures to assess and mitigate risk of biased output prior to the release of the AI system.

4. Transparency

The Code of Conduct acknowledges that individuals require sufficient information to make informed decisions and evaluate how risks are being addressed. Where systems are available for public use, the Code of Conduct commits developers to publish certain information regarding the AI system, such as: (1) information on the AI system's capabilities and limitations; and (2) a description of the types of training data used to develop the AI system. Developers of AI systems available for public use also commit to providing a reliable and freely available method to detect content the system generates. Finally, managers of both AI systems for public and private use commit to ensure AI systems that could be mistaken for humans are clearly and prominently identified as AI systems.

5. Human oversight and monitoring

The Code of Conduct emphasizes the importance of human oversight and monitoring. Managers commit to monitoring the operation of the AI system for harmful use or impact after the AI systems are made available. Managers also commit to inform the developer and/or implement usage controls to mitigate harm, as needed. For developers of AI systems, the Code of Conduct commits them to maintain a database of reported incidents that occur after deployment, and provide updates as needed to ensure effective mitigation measures.

6. Validity and robustness

The Code of Conduct underscores the importance of systems to operate effectively and, as intended, to be secure against cyber-attacks. Developers across all firms must ensure that AI systems operate effectively and are secured against cyber-attacks. Before deployment, the Code of Conduct commits developers to use a wide variety of testing methods across a spectrum of tasks and contexts to measure performance and ensure robustness. Developers also commit to employ adversarial testing (i.e., red teaming) to identify vulnerabilities, and commit to perform an assessment of cyber-security risk and implement proportionate measures to mitigate risks. In order to assess the validity and robustness of the AI systems, developers commit to perform benchmarking to measure the AI system's performance against recognized standards.

Lastly, signatories also make a general commitment to supporting the ongoing development of a robust, responsible AI ecosystem in Canada. This includes contributing to the development and application of standards, sharing information and best practices with other members of the AI ecosystem, collaborating with researchers working to advance responsible AI, and collaborating with other actors, including governments, to support public awareness and education on AI.

Next steps

As Canadian organizations consider the adoption of the Code of Conduct, we anticipate there will be significant discussions, including regarding the scope of application, delineation of actors, the meaning of specific definitions (e.g., developers, managers, high-impact systems, etc.), as well as guidance on the practical implementation of some of the prescribed measures, and also the interplay with forthcoming AIDA. ISED has indicated the Code will incorporate feedback from various participants in the AI ecosystem and would publish a summary of feedback in the coming days, which may illuminate some of the rationale behind some of the prescribed measures.

While the Code is voluntary in nature, we anticipate it will be widely referenced across various Canadian industries in the development and management of AI systems.