Campaign for AI Safety

Submission: Development Of Canadian Code Of Practice For Generative Artificial Intelligence Systems

This submission is made by the Campaign for AI Safety in response to the Canadian Guardrails for Generative AI – Code of Practice by Innovation, Science and Economic Development Canada. We trust this submission is of assistance.

The Campaign for AI Safety is a not-for-profit association established in Australia with members around the world. We are concerned about the dangers AI poses to people and advocate for a stop to the advancement of certain AI capabilities. We also advocate for regulation that promotes and mandates safe and ethical AI. We are not affiliated with any political group. Please visit campaignforaisafety.org for more information.

We agree with the proposed elements of safety, fairness and equity, transparency, human oversight and monitoring, validity and robustness and accountability in the voluntary code of practice for developers and deployers of generative AI technology.

We propose specific measures in the safety, transparency and accountability elements that developers and deployers could take to ensure this technology evolves safely. These remedying actions are proportional to the potentially large-scale societal impacts if AI were to go rogue or be misused and within the means of deeply resourced AI developers.

1. Safety

We consider safety to be the most crucial element to the future of AI because there is currently little private incentive for AI developers to prioritise safety testing, accuracy, and quality and to consider third-party impacts in the race[1] to develop more powerful and risky AI systems. For every AI safety research paper published, there are 50 AI general capabilities papers[2] which suggests there is a lack of knowledge about how to improve AI safety or minimise risks. This is quite concerning as there is increasing evidence of emerging and potentially dangerous capabilities in generative AI systems such as persuasion and manipulation, situational awareness and long-horizon planning[3].

In our view, the safety element needs to be much more than identifying risks, preventing misuse, and informing end users of the capabilities and limitations of the system. Firstly, it should prevent dangerous technology from being created, as once it is developed it can be hard to stop its proliferation and misuse. Secondly, it should create a robust safety culture at the organisational level, whereby safety is a key consideration in every member's work and not an impediment.

The safety element should be:

- Encourage developers to **not develop generative AI systems that are more powerful than what currently exists** (e.g. not more powerful than GPT-3 or GPT-4).
- **Dedicate a substantial portion of its research** funding (say, at least 50%) to the advancement of alignment, reliability, and explainability, and making this available, to build a knowledge base of AI safety accessible to all.
- At least half of its research staff should be employed in the safety area and not work directly towards advancing capabilities, modality, or model sizes
- **Establish an internal safety committee** that:
    - has legal veto power on the deployment of unsafe systems (similar to internal risk management committees at banks and other financial institutions);

---

[1] "These two AI models claim to be better than ChatGPT. Here's what we know", Sabrina Ortiz, Associate Editor on ZDNET (27 June 2023); "Google DeepMind CEO Demis Hassabis Says Its Next Algorithm Will Eclipse ChatGPT", Will Knight, WIRED (26 June 2023).
[2] "Emerging Technology Observatory Research Almanac", Zachary Arnold, Jennifer Melot, Dewey Murdick and Brian Love, Center for Security and Emerging Technology. AI Safety (19 May 2023).
[3] "Model evaluation for extreme risks", Toby Shevlane, et al. (24 May 2023).

- consist of public officers with remuneration not tied to the entity's financial performance; and
- company directors are not able to overrule the directions of the safety committee.

2. Accountability

The code proposes performing internal and external audits before and after the system's installation and operation. We suggest **disclosing the results of external audits after each API update release**. Evaluation and audits must continue after deployment as it is not always possible to fully anticipate how the system will work post-deployment and users may use the system in new ways which give rise to unanticipated or new risks. Continuous auditing is also needed because developers may make updates to the system post-deployment which may increase risk.

Public algorithmic audits can incentivise developers to prioritise improving their technologies. Researchers found that all the companies participating in Gender Shades, a public audit of gender and skin type performance disparities in commercial facial analysis models, made significant improvements in 7 months by reducing classification bias in subsequent commercial APIs[4].

3. Transparency

We believe individuals and businesses should be informed, to the extent possible, when AI-enabled tools are used and to understand the methods and potential outcomes.

Currently, businesses that purchase AI systems do not have adequate information about how they work, preventing them from truly evaluating the costs and benefits of their consumption and the suitability of these systems. This could lead to unintended consumer harm (e.g. a charity organisation using a chatbot that provides harmful advice to people with eating disorders[5]). At the same time, individuals or end users do not always know they are interacting with AI and have a right to know so that they can challenge outcomes and seek remedies to address any harm caused.

The code proposes to encourage developers and deployers to "meaningfully explain" the processes used to develop the AI system. We suggest strengthening this element by **publishing model cards and datasheets of datasets for high-risk generative AI systems** (we define "high risk" as those systems trained on large language models that are as powerful as GPT-3 or greater).

Model cards are a form of documentation and the principle is similar to safety data sheets or material safety data sheets required by a range of industries under Canada's occupational health and safety laws.

The model cards could detail:

---

[4] "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products", Inioluwa Deborah Raji and Joy Buolamwini, Conference on Artificial Intelligence, Ethics, and Society, MIT Media Lab (2019).
[5] "US eating disorder helpline takes down AI chatbot over harmful advice", Lauren Aratani, The Guardian (1 Jun 2023).

- performance characteristics
- training data (comprehensive references, not vague descriptions)
- the context in which models are intended to be used
- performance evaluation procedures.

We find the set of sections proposed in detail for model cards by Margaret Mitchell et al.[6] to be a useful guide for presenting this information.

Disclosing training datasets, model characteristics and the full results of evaluations (including both positive and negative test results) will help build public trust and confidence in AI and help businesses evaluate the suitability of these systems to their context.

We support the code's encouragement to disclose AI-generated content through means such as watermarking and to more generally make users aware they are interacting with AI systems or consuming AI-generated content.

An **exception should be made for computer-generated imagery in feature films and clearly recognisable animation**. We believe that the obligation should primarily fall on the publishers of content (including online platforms and their users). For audiovisual content, AI companies can be encouraged to embed watermarks and provide means to detect or check AI-generated content.

Watermarking and other technologies to identify AI-generated content have been found to work poorly[7] and are likely to be circumventable[8]. They **should not be relied upon as an enforcement tool** if the Government of Canada is considering mandating the proposed code or enforcement of AI regulation.

Lastly, we believe in transparency but some types of transparency can lead to catastrophic consequences. While it is not currently part of the proposal, we caution against encouraging or requiring developers to make too much information about AI systems available, specifically:

- open-sourcing of powerful, high-risk generative AI models which would allow malevolent actors to access and use for harmful purposes
- publishing "know-how" of the methods of training of very powerful models which can lead to their undesirable proliferation and consequent misuse.

---

[6] "Model Cards for Model Reporting", Margaret Mitchell, et al. (14 Jan 2019): Figure 1, Summary of model card sections and suggested prompts for each (Page 3).
[7] "We pitted ChatGPT against tools for detecting AI-written text, and the results are troubling", Armin Alimardani, Emma A. Jane (20 February 2023).
[8] "AI watermark remover tool lets users remove watermarks with a single click", Australian Photography (30 January 2023).