AIGS GSIA

4TH REVIEW OF THE

DIRECTIVE ON AUTOMATED DECISION MAKING

AIGS is pleased to present its recommendations in response to the public consultation on the Directive on Automated Decision-Making

January
2025

Prepared By:
Kathrin Gardhouse
Suchet Mittal
Yohan Matthews
Connor Fransoo
Julian Guidote

www.aigs.ca

AIGS GSIA

AI Governance & Safety Canada
Gouvernance et sécurité de l'IA Canada

January 8th, 2025

Jonathan Macdonald
Director, Responsible Data and AI
Office of the Chief Information Officer, Treasury Board of Canada Secretariat
Jonathan.Macdonald@tbs-sct.gc.ca

Dear Mr. Mcdonald,

We commend the Treasury Board of Canada Secretariat for its thoughtful revisions of the Policy on Service and Digital and the Directive on Automated Decision-Making and we appreciate the opportunity to provide feedback on the proposed updates.

AI Governance & Safety Canada is a nonpartisan not-for-profit and a community of people across the country, working to ensure that advanced AI is safe and beneficial for all. We provided input to ISED on the Voluntary Code of Practice for Generative AI, participated in TBS's AI Strategy consultations, delivered extensive recommendations for Bill C-27, and delivered our flagship white paper *Governing AI: A Plan for Canada*.

Our submission addresses the gaps in the current revisions regarding the readiness to manage the risks associated with advanced AI systems. These systems, expected to be capable of self-improvement, strategic planning, and autonomous decision-making, present challenges that necessitate agile, proactive oversight frameworks.

We propose four critical amendments to enhance the DADM and Policy:

1. Continuous Monitoring and Real-Time Revisions of AIAs
2. Enhanced Requirements for Human Oversight
3. Mandatory Safety Cases for Deployment
4. Scenario Planning and Stress Testing

In addition, we strongly support expanding the directive's scope to include domains such as national security, law enforcement, and policing, with safeguards to protect sensitive information and individual rights.

We remain available for any assistance that you require.

Sincerely,

Kathrin Gardhouse
Policy Lead
AI Governance & Safety Canada
contact@aigs.ca

Overview

Advanced AI systems pose serious risks to society without proper safeguards. An AI emergency response system might overlook vulnerable communities during disasters, while AI-driven policy decisions could misinterpret economic trends and damage public services. These scenarios highlight why robust oversight is crucial. Our recommendations outline practical steps to ensure AI systems serve everyone's interests while preventing unintended harm.

With the exception of the risk to democracy, the currently proposed updates to the Policy on Service and Digital (the Policy) and the Directive on Automated Decision-Making (DADM) do not fully address the transformative risks of advanced AI, which leading experts expect could arise within the next 1-5 years. We refer to our white paper for further details regarding the risk landscape and existing uncertainties.

We define advanced AI as a system that leverages sophisticated machine learning techniques and architectures to perform complex tasks, such as multi-step reasoning and strategic planning, at or above human-level proficiency and across diverse environments with minimal intervention. These systems may include capabilities such as self-improvement or agentic behaviour, allowing them to perceive and respond to their environment autonomously (e.g., supply-chain or emergency response optimizer, smart city planner, policy maker).

Advanced AI will be difficult to control. Such systems could cause social disruptions through rapid changes in the labour force, compromise national security, and exhibit harmful emergent behaviors, leading to the amplification of societal inequities and catastrophic accidents. If deployed in complex, high-stakes contexts like resource allocation, policy design, and national security, the likelihood for such risks to manifest and their severity increases significantly. The pace of development resulting from the race between AI labs and nations to be the first to develop and deploy cutting edge models is accelerating. Consequently, effective governance of advanced AI must be agile and future-focused to react to unforeseen developments.

Recommendations

We maintain that the DADM needs to be ready to address the governance challenges that the use of advanced AI could give rise to. Here are four key amendments that should be introduced to the DADM or the Policy as critical first steps to better address emerging risks:

1. **Continuous Monitoring and Real-Time Revisions of AIAs:**
   AI systems powered by neural architectures are no longer static; they evolve post-deployment, adapt to new data, and may exhibit emergent behaviors. The AIA should shift from a one-time risk assessment with changes required when the scope and functionality changes, to a dynamic process requiring continuous monitoring of deployed AI systems. Departments must establish mechanisms for ongoing evaluation of system outputs, behaviors, and cumulative societal impacts. This could include mandatory auditing triggers—such as shifts in performance or behaviours outside expected parameters—and real-time updates to mitigation strategies as risks emerge. Section 6.1.1. (Slide 8) is a step in the right direction, however, we recommend moving away from requiring review on a “scheduled basis” in favor of continuous monitoring with scheduled reporting/approving.

2. **Enhanced Requirements for Human Control and Oversight:**
   AI systems integrated into decision-making processes must include enforceable requirements for human-in-the-loop oversight, ensuring that human operators can meaningfully intervene, override decisions, and disable systems. The human oversight requirements currently in place for Level III and IV systems seem insufficient in the context of advanced AI. For systems with advanced autonomy or adaptive capabilities, additional safeguards must be implemented, such as:
   - Real-time monitoring tools to detect unexpected changes in behavior.
   - Strict accountability mechanisms to ensure responsible ownership and control as well as clarity around how harmed individuals can obtain redress.

3. **Mandatory Safety Cases for Development and Deployment of Advanced AI:**
   The Policy should mandate Safety Cases for the use of advanced AI systems with specific capabilities that are difficult to govern. This should include research or experimentation, which are currently excluded as per Section 5.2 of the proposed amendments. An AI Safety Case solution seems superior to a ban, as the WEF's AI Agents White Paper points out that "the application of AI agents can play a crucial role in addressing the shortfall of skills in various industries, filling the gaps in areas where human expertise is lacking or in high demand." These advantages may make the implementation of advanced AI in (high-risk) public sector decision-making unavoidable.

   Safety Cases are a fast-advancing area that addresses the unpredictable pace of capability gains in frontier AI, where each new system's unique capabilities demand tailored safety measures. They can be defined as a structured argument supported by evidence that a system is safe enough to be deployed in a given operational context. This approach has been used in many other safety-critical industries like aviation, nuclear energy, and autonomous vehicles. For instance, in nuclear energy, Safety Cases are required by the Canadian Nuclear Safety Commission to ensure safe operations of facilities. In 2011, a component of the Safety Case for the Darlington Nuclear Generating Station led to significant safety enhancements, including the installation of an emergency power generator to maintain critical systems during power loss and a containment filtered venting system to reduce the impact of radioactive releases in severe accidents. These improvements highlight the effectiveness of Safety Cases in identifying vulnerabilities and implementing targeted mitigation measures in high-risk industries.

   There are also examples of work on AI Safety Cases relating to cyber risk and scheming and further details on the UK AISI's work in this regard.

   An AI Safety Case has four key components:
   - **Scope** – The conditions under which an AI system is considered safe, detailing its specifications (such as architecture, training processes, and safeguards) and intended deployment context (e.g., API access or release of model weights). This could include which system or context changes, like post-deployment fine-tuning or expanded access, fall within the scope or will require updates to the case.
   - **Objectives** – The requirements for deployment, often formatted as risk thresholds (e.g., $\geq 10^{-7}$ annual probability of events causing ≥1,000 fatalities) with categories such as "unacceptable," "as low as reasonably practicable," and

"acceptable" risks. Comparative objectives, like ensuring a system is at least as safe as its human counterpart, are also common.
    - **Arguments** – A structured argument composed of a collection of verified subclaims (as premises) which serve as evidence of having met the objective(s) (serving as a conclusion). They account for the relevance of the evidence using a logical framework to demonstrate the sufficiency of the safety measures.
    - **Evidence** – Supporting materials such as capability evaluation results (including stress testing, edge case handling, and adversarial scenarios), expert judgements, formal proofs, and risk analyses. The viability of empirical evidence used to test frontier models remains unclear, hence the required evidence will have to evolve accordingly.

Safety Cases are prepared by an AI developer when deciding whether to deploy a system or use it in a high-risk scenario, reviewed by an independent third party, and reviewed and approved by decision-makers (e.g., the Assistant Deputy Minister).

The capabilities that should require mandatory Safety Cases, inspired by the current risk taxonomy provided in the draft Codes of Practice under the EU AI Act, should include:

- Cyber-offensive capabilities, Chemical, Biological, Radiological and Nuclear (CBRN) capabilities, and other weapon acquisition or proliferation capabilities
- Self-replication, self-improvement, and ability to train and develop itself or other models
- Long-Horizon Planning, Forecasting, and Strategizing
- Self-reasoning (a model's ability to reason about and modify the environment – including its own implementation – including the ability to self-modify)
- Automated AI research and development

Safety Cases complement AIAs. While AIAs result in a risk score onto which it maps mandatory mitigation measures, a Safety Case quantifies the risk (in its objectives) of not achieving a safety goal (e.g., in terms of fatalities or material damages) and thus provides a tangible basis for a decision to deploy or not to deploy, permitting confidence that adverse outcomes have been appropriately analyzed and mitigated.

4. **Mandatory Scenario Planning and Stress Testing:**
Scenario planning and stress testing should be mandatory for advanced AI systems to address the heightened stakes and ensure responsible governance. These efforts must focus on identifying and mitigating foreseeable and unforeseen risks. Elements include:
    - **Exploring catastrophic failure modes** and worst-case scenarios, such as system misuse, emergent behaviors, or societal disruptions.
    - **Evaluating societal impacts**, including exacerbated inequalities, eroded public trust, and long-term harms to vulnerable groups or democratic processes.
    - **Guiding efforts through multi-stakeholder reviews**, incorporating input from the public, domain experts, and ethicists to capture diverse perspectives.
    - **Connecting every identified risk to a clear response strategy**, such as fail-safes, shutdown mechanisms, and recovery protocols.

An acceptable level of effort should be explicitly determined based on the system's risk profile. High-risk systems should require proportional resource allocation, with at least 20–30% of the project budget or time dedicated to scenario planning and stress testing. By comparison, safety-critical software in industries such as aviation, where compliance with standards such as DO-178C is required, often need an additional 25–40 percent of development budgets for testing and validation.

Mandatory Scenario Planning and Stress Testing is relevant both in instances where no Safety Case is required as well as in addition to a Safety Case. This is because the Safety Case will be provided by the developer, likely with a broad range of possible use cases in mind. The scenario planning and stress testing needs to be conducted by the decision maker subject to the DADM and be tailored to the specific use case.

Scope Expansion

Finally, we wish to express our support for the proposed elimination of the exclusion of several public institutions from the scope of the DADM. We also strongly encourage the inclusion of national security, law enforcement, and policing within its scope while taking precautions to prevent the inappropriate disclosure of sensitive information. Expanding the directive's scope to cover these areas is especially critical given the significant societal impact of these domains. Automated systems in these areas have the potential to profoundly influence individual rights and freedoms, including decisions on surveillance, criminal investigations, or immigration enforcement. Without the safeguards mandated by the DADM, these applications could disproportionately affect marginalized groups, perpetuate bias, or lead to severe legal and ethical consequences.