# Canadian Forum for Digital Infrastructure Resilience (CFDIR)

## Artificial Intelligence / Machine Learning (AI/ML) Working Group (WG)

# OPERATIONALIZING THE GEN-AI VOLUNTARY CODE OF CONDUCT FOR CANADIAN CRITICAL INFRASTRUCTURE ENTITIES

Version v1.0 – 2024 September 25

AUTHORED BY:
**AI/ML WG**
**of the Canadian Forum for Digital Infrastructure Resilience (CFDIR)**

**TLP:CLEAR**

# CONTENTS

# FOREWORD

The contents of this document were developed during the course of Canadian Forum for Digital Infrastructure (CFDIR) Artificial Intelligence / Machine Learning (AI/ML) Working Group meetings, from 2023 Sept – 2024 July.

# ACKNOWLEDGEMENTS

The information and recommendations contained herein were informed by the active participation and engagement of subject matter experts from the following organizations (listed alphabetically):

## CFDIR MEMBERS:

Industry :

- BlackBerry
- IBM
- Thales Canada

Government :

- Canadian Centre for Cyber Security (CCCS)
- Innovation, Science, and Economic Development (ISED)

## REVISION HISTORY

*The following table describes the dates of the major changes to this document.*

| Authors | Date / Version | Notes |
|---|---|---|
| CFDIR AI/ML WG | Version 1.0<br><br>2024 September 25 | Version 1.0 |
| | | |

# 1. EXECUTIVE SUMMARY

This document is intended as a practical guide to help critical infrastructure (CI) sector operators and owners identify, mitigate, and manage risks associated with the deployment and use of Artificial Intelligence systems, tools and applications in their environments and systems. The document proposes key considerations for AI risk governance by CI sectors. The guidance is aligned *with Canada's Voluntary Code of Conduct on Generative AI* and *NIST's AI Risk Management Framework*, as well as other risk mitigation best practices outlined in Section 5.

The primary audience for this document are CI operators and owners, and more specifically individuals and teams responsible for DevSecOps, cybersecurity, privacy, compliance and legal teams in CI entities as well as users of critical infrastructure who need to know that the safety and security of CI is a priority. The document is meant to provide these individuals with key considerations to protect their systems against implementation risks of AI in their IT and OT infrastructure.

Developers of advanced AI systems are also encouraged to consider this guidance so they can embed security, accountability, safety, transparency, fairness and equity into their models and systems from the start. This document is intended to be updated as required. CFDIR's AI working group will review the content and usefulness of the guidance on a regular basis and determine if an update is appropriate.

# 2. INTRODUCTION AND PURPOSE

These guidelines are intended to help CI owners and operators identify, manage and address risks to the safety and security of the public.

Like many other guidelines and standards that intend to enhance public security and safety alongside the commercial use of advanced technologies, their interpretation for a specific application, and in particular for critical infrastructure, is needed to allow for operational conformity. Given the rapidly evolving technological landscape and fluid regulatory environment, these guidelines are intentionally broad and non-prescriptive. The guidelines are neither a checklist nor a set of steps that need to be followed in their entirety. Rather, organizations can apply as many – or as few – of the suggestions contained in these guidelines as apply to their industry and depending on their risk tolerance.

AI technologies have significant potential to drive economic growth, transform society, and help tackle some of the world's most pressing challenges including climate change, inequality, and disease. At the same time, these technologies also pose risks that can be high impact, particularly in critical infrastructure. AI systems are complex and frequently opaque. Identifying and managing risks associated with advanced AI systems is challenging because AI systems are constantly learning and evolving their functionality over time. AI risk management can help drive the safe, secure, and trustworthy use of AI in critical infrastructure by prompting organizations to think more critically about the potential negative impacts in advance of deployment. A core premise of this document is that risks must be managed to maximize the positive impact of AI systems.

The first step of interpreting is understanding the zones of tension between what beneficial outcomes AI can deliver and its potential inappropriate use for malicious or fraudulent purpose with possible unsafe or insecure impact. AI, and its encompassing technologies (deep learning, machine learning, operational research, data analytics, AI agents, Large Language Models etc.) deliver advanced and enhanced functionalities, new levels of performance and optimization to designers, operators, supply chain vendors, clients and users of critical infrastructure. These beneficial outcomes are not realized by default, they require acceptance, adoption, and transparency at multiple levels and the open exchange of information regarding risk and benefit between the full range of critical infrastructure stakeholders.

For example, the designer of an AI system will need to demonstrate to the approving authority within the CI entity that above and beyond functionality and performance, the security and safety of the system meets the needs of the critical infrastructure provider, operator and user. At the other end of the spectrum, it needs to be recognized that users of critical infrastructure often take the safety and security of CI for granted. In effect, they place their trust in the operator for ensuring this safety and security. Most important for the user is that infrastructure does what it is designed to do consistently. For the user, operational transparency is important. This means being transparent about when the performance of CI services deviates or degrades from the norm, why that performance was impacted, and communication about when the service will be restored to its expected state.

There is also a risk for AI to be used inappropriately in critical infrastructure. This includes fraudulent acts as well as malicious or criminal intent to disrupt services or hold operators, owners, and in some cases, users for ransom. Perpetrators are aware of the criticality of these services and maximize their efforts to extract ransom by causing either maximum or strategic disruption to those services. As AI becomes integrated into these systems to improve effectiveness and efficiency, the potential attack vectors that could impact these systems also increase. And herein lies the conundrum. AI systems have the potential to help CI operators serve their customers more effectively, inclusively and efficiently and reduce their climate impact. At the same time, these systems introduce additional risks, both known and unknown.

*Canada's Voluntary Code of Conduct on Generative AI* aims to help entities manage this conundrum by mitigating risks throughout the entire life cycle of the system, including the software and hardware that underpins it. The more entities that sign on to the code, adhere to its recommendations, and create a community of practice that values safety, security, transparency and accountability as much as innovation and AI adoption, the better the public will be served.

At the core of our effort is the goal of helping CI entities better understand where they should be vigilant and where they can best limit the risks and maximize the benefits of AI in CI systems to the extent possible. This document puts forward suggestions for mitigating and managing unintended outcomes and vulnerabilities, without being prescriptive. Each company or organization should choose to develop their own internal mitigation and management plan to achieve their commitment to align with *Canada's Voluntary Code of Conduct on Generative AI*. As is noted in *NIST's AI Risk Management Framework*, policies and resources for AI risk management should be prioritized based on the assessed level or risk and potential impact of an AI system in that environment.

# 3. AI RISK GOVERNANCE FRAMEWORK FOR CI SECTORS

## 3.1 OVERVIEW

AI will be part of CI organizations infrastructure whether as a planned capability or as part of the IT/OT products they use. In addition to their own specific requirements most organizations will be subject to legal and sector-specific regulation on AI. They will have to ensure related risks are managed effectively. The US' *NIST AI Risk Management Framework* (RMF) provides a framework for this. The framework identifies four key areas:

- **Govern** – establish the organizational elements to manage AI risk including policy, procedures, organizational values and principles, and culture throughout AI system's lifecycle;
- **Map** – establish the context to frame risks related to an AI systems. This includes understanding and managing the organization's specific obligations in law and sector-specific regulation as well as the risks and opportunities specific to the organization and its context;
- **Measure** – employ quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts on organization's operations and data;
- **Manage** – continuously monitor AI risks and take action to manage.

This framework can be used by CI organizations to understand and manage AI risk in their organization. The details on implementation of the framework in CI organizations borrows heavily from the US Department of Homeland Security's *Mitigating Artificial Intelligence (AI) Risk: Safety and Security Guidelines for Critical Infrastructure Owners and Operators* publication.

## 3.2 GOVERN

Governance aims to establish the organizational elements to manage AI risk including policy, procedures, organizational values and principles, and culture throughout AI system's lifecycle. This includes following an overall "secure by design" approach for AI systems. Leadership is required to take ownership and prioritize the safety and security outcomes of AI system implementation.

Actions organizations should consider include:

- Establishing a management framework for AI systems – This includes policy and procedures for implementing and operating AI systems as well as terms of reference to establish accountability;
- Establishing detailed plans for cybersecurity risk management – This includes preparing incident procedures, attacks against organizational systems as well as AI system failure;
- Establishing secure by design practices – This includes establishing secure design practice across the AI system lifecycle;
- Establish governance of enterprise data used with AI – AI systems require use of organizational information to be useful. The information and its use in AI systems needs to be tracked and carefully managed;
- Establish roles and responsibilities with AI vendors – Roles and responsibilities within the organization as well as with the vendor need to be established to ensure secure and reliable operation of systems;
- Prepare your workforce – AI is being introduced into many products and is rapidly becoming part of business. Organizational and operational staff need to be aware of the risks involved with system and the data they use through training and awareness campaigns;
- Assess the trade-offs of AI deployment models – Understand that not all models present the same features and risks. These must be weighted. Consider the provenance and testing done on the models for safety.

- Establish transparency in AI system use – Ensure output of AI systems is clearly documented and ensure users of the output are aware of the AI use as well as considerations they should be aware of in using the AI output; and
- Collaborate with government and industry groups – Share your experience on implementing and managing AI systems. Be open to the experience of others.

These actions should be integrated into existing governance structures within your organization. They should be part of the organization's regular course of business.

## 3.3 MAP

Organizations need to understand their tolerance for risk. This tolerance includes addressing the organization's responsibilities related to public safety, laws, and specific sector regulation. It can also include organization specific risks in areas such as finance and reputation. Actions organizations should consider in this area include:

- Tracking all existing and proposed AI use cases – Organizations must keep aware of AI implementation as part of their systems. This includes to the extent possible understanding "shadow-IT" implementations;
- Document related safety and trust impacts – Organizations must assess and document risks of systems including expected benefits. Trust considerations include safety, security and resilience, accountability and transparency, explainability and interpretability, and privacy protection;
- Conduct AI impact assessments as part of any AI system deployment – Organizations need to understand the impact of AI in the overall system. This includes procured systems;
- Identify AI systems requiring human supervision – AI implementations that make decisions should generally be considered for human supervision to address trust considerations;
- Manage AI vendor supply chains for AI related risks – AI systems are made up of multiple components and have dependencies. Organizations need to understand responsibility for AI security and safety of components across the system lifecycle;
- Keep informed of the evolving AI risk space – The AI landscape is changing very fast. Organizations need to stay aware.

As with governance, these actions should be integrated into existing compliance and risk management structure within your organization.

## 3.4 MEASURE

Organizations need to establish systems to assess, analyze and track AI risks. They need to be able to employ quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts. Actions organizations should consider in this area include:

- Define metrics and approaches – Organizations need to understand the risk and performance characteristics of their systems and identify how to measure them for the purposes of managing the systems;
- Continuous testing of AI systems – Organizations should continually test their AI systems for safety and performance
- Assess performance of risk controls – Organizations should regularly assess the implementation and effectiveness of controls they have established to manage their AI implementations;
- Establish practices to prevent disclosure of sensitive information – As part of training, models can incorporate specifics of the data that can potentially be extracted. Publicly available chat agents can retain prompt information. Organizations need to carefully consider the information used in developing systems or interacting with external AI services to ensure sensitive data is not accidentally disclosed;
- Measure AI system performance and output – Organizations need to establish metrics for AI system performance and quality and need to continually assess the system against these metrics;
- Test and evaluate – In addition to passive monitoring organizations need to actively test and evaluate their systems. This includes red-teaming; and
- Establish reporting mechanisms – Organizations should ensure that the metrics they collect and the analysis done is reported to risk owners and stakeholders.

These actions should be integrated into existing risk assessment and reporting structures within your organization. These reports should be fed back into the design and development processes for future CI systems.

## 3.5 MANAGE

Take action on risks and allocate risk resources to mapped and measured risks regularly in accordance with governance functions. Actions organizations should consider in this area include:

- Prioritize AI safety and security risks – Organizations need to make AI safety and security a priority aspect of their implementations;
- Follow cybersecurity best practice – Organizations should understand and follow best practices for cybersecurity as a foundational aspect of managing AI systems;
- Implement new or strengthened mitigation strategies – AI systems and associated risks is a rapidly evolving domain. Organizations need to keep aware of new risks and means of mitigation relevant for their systems and implement them as appropriate;
- Implement tools such as watermarks, content labels and authentication techniques – As part of management and transparency, organizations need to ensure content produced by AI is appropriately labelled;
- Apply appropriate security controls – Organizations need to ensure appropriate security controls are deployed with their systems. Controls and their implementation need to be regularly reviewed;
- Apply mitigations prior to deployment – Organizations need to ensure that appropriate controls and means of mitigation are applied before systems are deployed;
- Monitor AI systems' inputs and outputs – Organizations should ensure their AI systems are continuously monitored and details captured in logs and reviewed for issues;
- Build AI systems with resilience in mind – Organizations need to ensure built or procured systems are resilient; and
- Regularly review risks that have poor measurements – Organizations should pay particular attention to risk categories that they find hard to quantify.
- Respond to incident in accordance with incident management plans – Organizations must prepare incident management plans and playbooks to respond to potential issues.

It must also have the ability to action those plans and playbooks in response to AI incidents.

These actions should be integrated into existing risk management structures within your organization.

# 4. CONCLUSION

This document is not intended to be sector specific other than critical infrastructure, but even then, with all the trades and business domains involved, it must remain somewhat generic. With the help of suggestions provided above, it will be up to each company and organization to increase their awareness and potentially devise a conformity plan to *Canada's Voluntary Code of Conduct on Generative AI*. The specifics and suggestions in Section 3 are mere suggestions that could be considered as organizations develop their plans as per their context requires. Ultimately, the goal is to provide the CI sector with guidance for a roadmap to enhance the safety and security of their systems while leveraging the benefits that AI can bring.

# 5. REFERENCE DOCUMENTS

## 5.1 RISK MITIGATION BEST PRACTICES

i. Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems

ii. NIST AI Risk Management Framework

iii. NIST AI 600-1

iv. DHS Safety and Security Guidelines for Critical Infrastructure Owners and Operators

v. ICT supply chain risk management

vi. Secure-by-Design

vii. Energy Security Technical Advisory Committee (E-STAC) | Canadian Gas Association (cga.ca)

# ANNEX A: GLOSSARY

| | |
|---|---|
| AI | Artificial Intelligence |
| CFDIR | Canadian Forum for Digital Infrastructure Resilience |
| CI | Critical Infrastructure |
| GenAI | Generative AI |
| IT | Information Technology |
| ML | Machine Learning |
| OT | Operational Technology |
| RMF | Risk Management Framework |
| | |

This document is planned to be updated periodically, to reflect feedback and comments from implementing the information described herein.

**TLP:CLEAR**

Prepared by the AI/ML Working Group of the Canadian Forum for Digital Infrastructure Resilience (CFDIR)