New NIST framework provides organizations guidance on AI governance and risk management

March 14, 2023

On Jan. 26, 2023, the National Institute of Standards and Technology (NIST) published the Artificial Intelligence Risk Management Framework (AI RMF). Recognizing the unique complexity and deep impact of artificial intelligence (AI) tools and systems, this framework provides guidance regarding the management of risks associated with their design, development, deployment and use, such as biased decision making, and proposes best practices.

The AI RMF offers useful guidance on AI governance and risk management at a time when Canadian organizations are preparing themselves for new legislation governing automated decision-making tools (coming in Québec as of September 2023). It also comes at a time when Bill C-27 – and its proposed *Consumer Privacy Protection Act* (CPPA) and *Artificial Intelligence and Data Act* (AIDA) – has resumed debate in the House of Commons this session.

Even in the absence of specific AI laws and regulation, the AI RMF can help organizations implementing a governance structure and controls to foster responsible design, development, deployment and use of AI systems, and mitigate associated compliance risks (which, in some case, arise from existing legislation, such as privacy, employment and human rights law).

What is NIST's AI RMF?

NIST is part of the U.S. Department of Commerce, and its mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards and technology in ways that enhance economic security and improve quality of life.

The AI RMF is a voluntary framework that provides organizations with tools to evaluate, classify, communicate and manage the risks presented by AI systems. It also discusses seven characteristics of trustworthy AI systems.

Along with the "Core" AI RMF, NIST released several accompanying documents and an explainer video, including:

- A playbook setting out processes that organizations can customize to achieve the four outcomes of the AI RMF: govern, map, measure, and manage risks associated with AI systems.
- A roadmap, which lists activities that NIST may undertake to advance the AI RMF, both on its own and in collaboration with private and public sector stakeholders. For example, given that using the AI RMF requires organizations to define their own risk tolerance, NIST intends to provide guidance on methods for developing reasonable risk tolerances.
- Crosswalks, which describe how the AI RMF interacts with other prior AI guidance. The two draft Crosswalks already published relate to (i) the ISO/IEC 23894:2023 standard on AI and (ii) an illustration of how NIST AI RMF trustworthiness characteristics relate to the OECD Recommendation on AI, Proposed EU AI Act, U.S. Executive Order 13960 (*Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*), and the White House's Blueprint for an AI Bill of Rights.

How does the AI RMF define AI systems?

The AI RMF proposes a definition of "AI system" adapted from the OECD Recommendation on AI and the ISO/IEC 23894:2023 standard on AI, which shares certain similarities with AIDA's definition:

NIST AI RMF	AIDA (s. 2)
An engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.	*A technological system that, autonomously or partly autonomously, processes data related to human activities through the use of a genetic algorithm, a neural network, machine learning or another technique in order to generate content or make decisions, recommendations or predictions.*

In both the AI RMF and AIDA's case, the intent appears to be to capture a broad spectrum of systems. While the two definitions both refer to the same types of output, namely decisions, recommendations and predictions, AIDA also mentions the generation of *content*. Unlike AIDA, the AI RMF does not refer to specific types of technology, presumably in an attempt to make its definition technology neutral. Furthermore, both definitions recognize that AI systems may have varying levels of autonomy. This aligns with the CPPA's definition of *automated decision systems*, but contrasts with Québec's Bill 64, which regulates the use of personal information to render a decision based *exclusively* on automated processing (s. 12.1).

Framing risk: evaluating, classifying, communicating and managing risks posed by AI systems

The AI RMF offers guidance for AI risk management to minimize the negative impact of AI systems. Unlike proposed AI laws in Canada and the EU, the AI RMF does not

propose a tiering of AI systems based on their risk or potential impact, nor does it set defined criteria to measure risk. In contrast, AIDA requires covered organizations to assess whether the AI system is a "high-impact system" according to criteria to be prescribed by regulation (s. 7). The determination that a system is "high-impact" would result in risk mitigation, transparency and explainability obligations, as well as a requirement to notify the minister if the use of the system results or is likely to result in material harm (ss. 8, 9, 11 and 12).

The AI RMF provides guidance for organizations to understand and address AI risks, impacts and harms. Examples of harms include:

- Harm to people, which includes individual harm (e.g., harm to a person's civil liberties), harm to a group (such as discrimination), and societal harm (such as harm to educational access).
- Harm to an organization's business operations or reputation, or harm arising from security breaches or monetary loss.
- Harm to an ecosystem, such as harm to the global financial system or to natural resources.

Interestingly, these examples are broader than the definition of harm under AIDA, which focuses on harms to an individual, namely physical or psychological harm to an individual, damage to an individual's property, economic loss to an individual (s. 5(1)).

In framing AI risks, the AI RMF notes that organizations will need to address certain challenges, in particular the following ones:

- *Risk measurement*: The AI RMF begins by proposing that the risks posed by AI systems can be measured using a classic matrix, which describes risk as a function of (i) the negative impact, or magnitude of harm, that would arise if the circumstance or event occurs and (ii) the likelihood of occurrence. Not all risks may be foreseeable when an AI system is conceived or implemented, and, if they are, they may be difficult to measure. Further, risks may need to be evaluated differently depending on the part of the AI lifecycle at which they emerge. Organizations should take a flexible approach to risk measurement, so that they may respond appropriately when new risks emerge.
- *Risk tolerance*: The AI RMF does not prescribe risk tolerance for AI systems. However, it articulates some points that organizations may consider when determining their tolerance for these systems. An organization's tolerance for risk associated with its use of an AI system will be different depending on the system's purpose, as well as the policies and norms regarding AI systems established by various interested parties (e.g. the AI system owner, users, as well as government and non-government policy makers). Risk tolerances will also likely change over time and during the AI system's lifecycle.
- *Risk prioritization*: Recognizing that organizations are unlikely to eliminate risk, the AI RMF offers some considerations for organizations when determining which risks are most salient. Organizations should create a strong risk management culture and efficient risk triaging protocols, both of which will allow them to dedicate resources to managing the most important risks first. Further, organizations should consider what factors in its circumstances would most appropriately increase risk prioritization. For example, an AI system that interacts with personal information or manages large datasets should be prioritized.

Finally, once an organization has prioritized and/or minimized the risk associated with an AI system, it must assess and be satisfied that the residual risk to be borne by end users is tolerable.

- *Organizational integration and management of risk*: Organizations should avoid considering the risks that AI systems pose in isolation, and instead treat them in an integrated manner and incorporate them into their overall risk management strategy. While some risks, like confidentiality and cybersecurity, are common to other resources like software and data management processes, others are unique to AI systems and may require a substantial amount of effort to situate within an organization's risk management practices.

Characteristics of trustworthy AI systems

The AI RMF framework sets out seven characteristics of trustworthy AI systems. Each of these characteristics will not apply to the same extent to each AI system. Further, AI systems do not exist in a vacuum, and thus each characteristic can vary in its importance depending on an AI system's use as well as the context of such use. Each characteristic's importance will also depend on the data it relies on, its output, the creators' decisions, as well as the extent to which humans interact and oversee the system.

The AI RMF proposes that trustworthy AI systems are:

1) *Accountable and transparent*: Users should have access to the appropriate level of information regarding the AI system. This level will depend on how and in what context individuals use the AI system, the risk associated with it, and the system's AI lifecycle stage, among other factors.

These characteristics are broadly in line with AIDA's transparency requirements for high-impact systems. More specifically, AIDA requires those who make high-impact AI systems available for use and who manage the operation of such systems to publish a plain-language a description of the system on a publicly available website. This description must include information set out by AIDA and to be prescribed by regulation (s. 11).

Bill 64 and Bill C-27 also include transparency requirements regarding automated decision-making. For example, Bill 64 gives an individual the right to be informed of an organization's exclusive reliance on an automated decision-making (ADM) tool, if that ADM tool is making decisions about the individual based on their personal information (s. 12.1). The individual also has the right to "submit observations" to someone within the organization who can review the decision. Further, upon request, the organization must inform the individual of 1) the personal information used to render the decision; 2) the reasons and the principal factors and parameters that led to the decision; and 3) the right of the person concerned to have the personal information used to render the decision corrected.

2) *Valid and reliable*: The AI system should perform its task both properly and consistently over time and in the range of circumstances in which it was intended to operate. This characteristic also underpins the five other characteristics listed below.

AIDA does not explicitly impose obligations in this regard.

3) *Safe*: Organizations should prioritize safety at each stage of the AI lifecycle and should not design or use AI systems that "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered". Beyond the responsible design, development, and deployment of AI systems, organizations should provide clear information on how to interact with the system safely and ensure they align with industry safety standards.

This characteristic is comparable with AIDA's principles regarding risk of harm mitigation for high-impact systems.

4) *Secure and resilient*: In today's infinitely connected environment, cybersecurity should be top-of-mind when designing and deploying AI systems. AI systems should be able to withstand adverse events and unexpected changes, and, in cases where they cannot, should be designed to fail safely.

While AIDA does not directly address these characteristics, AI systems that are not secure and resilient create a risk of harm and biased output, which are key concerns under AIDA.

5) *Explainable and interpretable*: All who interact with an AI system should be able to understand its purpose and impact.

These principles go hand-in-hand with the accountability and transparency principles and are aligned with explainability requirements under AIDA, Bill 64 and the CPPA.

6) *Privacy-enhanced*: Privacy should be central to the development of AI systems. Privacy-enhancing technologies, de-identification, and data aggregation are all useful tools for the development of privacy-enhanced AI systems. That said, in situations where data is sparse or otherwise incomplete, these strategies can reduce the accuracy of AI systems.

In Canada, privacy requirements would be addressed by Canadian privacy laws rather than by AIDA.

7) *Fair with harmful bias managed*: Since AI systems are the product of the humans that create them and the data that they used, efforts should be made to reduce and manage bias when designing and using these systems. These characteristics are similar to AIDA's principles regarding risk of biased output for high-impact systems.

Key takeaways

Even in the absence of AI legislation in Canada, any Canadian organization designing, developing, deploying, or using AI systems should be developing its risk management framework to mitigate the various categories of risks (*e.g.*, legal, ethical, business, reputational, etc.) that may arise from these activities. This may concern a broad range

of organizations that are not necessarily technology companies at their core, given the prevalence of IT tools that include an AI component.

For these organizations, the NIST's AI RMF may serve as a useful tool to develop a governance framework around AI. Organizations that that have an AI risk mitigation in place will be in a better position to respond to any upcoming AI legislation and privacy legislation governing certain AI systems.

If you have any questions about the new NIST framework, and how the framework can provide your organization guidance on AI governance and risk management, please reach out to any of the authors or key contacts listed below.

By

Marc Vani

Expertise

Cybersecurity, Privacy & Data Protection, Technology, Artificial Intelligence Law

---

BLG | Canada's Law Firm

As the largest, truly full-service Canadian law firm, Borden Ladner Gervais LLP (BLG) delivers practical legal advice for domestic and international clients across more practices and industries than any Canadian firm. With over 725 lawyers, intellectual property agents and other professionals, BLG serves the legal needs of businesses and institutions across Canada and beyond – from M&A and capital markets, to disputes, financing, and trademark & patent registration.

blg.com

BLG Offices

Calgary

Centennial Place, East Tower
520 3rd Avenue S.W.
Calgary, AB, Canada
T2P 0R3

T 403.232.9500
F 403.266.1395

Ottawa

World Exchange Plaza
100 Queen Street
Ottawa, ON, Canada
K1P 1J9

T 613.237.5160
F 613.230.8842

Vancouver

1200 Waterfront Centre
200 Burrard Street
Vancouver, BC, Canada
V7X 1T2

T 604.687.5744
F 604.687.1415

Montréal

1000 De La Gauchetière Street West
Suite 900
Montréal, QC, Canada
H3B 5H4

T 514.954.2555
F 514.879.9015

Toronto

Bay Adelaide Centre, East Tower
22 Adelaide Street West
Toronto, ON, Canada
M5H 4E3

T 416.367.6000
F 416.367.6749