April 2022

Lessons from the NIST AI RMF for the EU AI Act

Input for the US-EU TTC

Contact
carlos@futureoflife.org

Foreword

The Future of Life Institute (FLI) works to promote the benefits of technology and reduce their associated risks. FLI has become one of the world's leading voices on the governance of artificial intelligence (AI) and created one of the earliest and most influential sets of governance principles, the Asilomar AI Principles. FLI maintains a large network among the world's top AI researchers in academia, civil society, and private industry.

Lessons from the NIST AI RMF for the EU AI Act

The forthcoming Artificial Intelligence Act is set to place the EU at the global vanguard of regulating this emerging technology. However, models for the governance and mitigation of AI risk from outside the region can still offer important lessons for EU decision-makers to learn and issues to consider before the Act is passed. This is certainly true with Article 9 of the EU AI Act, which requires developers to establish, implement, document, and maintain risk management systems for high-risk AI systems. This document outlines several key ideas for EU decision-makers to consider from the AI Risk Management Framework (AI RMF), a soft law tool, still under development, which is aimed at assessing the capabilities and effects of AI systems, and managing identified risks.[1]

Risk management for all systems and inclusive of stakeholders

Unlike Article 9 of the EU AI Act, the AI RMF by NIST is designed to facilitate the management of risks from any type of AI system, not just those classified as high-risk. To do this, the AI RMF provides interested parties with examples of potential sources of harm directed at **people** (divided into effects that are felt at the individual, group/community, and societal level), **organizations** (harm that can impact technical systems and business operations), and **systems** (harm to an organized assembly of interconnected and interdependent elements and resources). The EU AI Act is not currently designed to do this, but it could benefit from encouraging all developers and deployers of AI systems to perform similar risk assessments, so that potential harms at all levels of society are identified and documented.

Similarly, the AI RMF acknowledges and includes the diversity of stakeholders involved in the analysis of an AI system's impact. In a best-case scenario, NIST believes that representatives from a number of stakeholder groups should be involved in the analysis of AI risks. The EU AI Act could likewise empower entities that produce high and low impact AI systems to engage directly with these stakeholders to uncover direct and indirect harms. These stakeholder groups would/might include:

- **AI system stakeholders:** Those who have the most control and responsibility over the design, development, deployment, and acquisition of AI systems, and the implementation of AI risk management practices.
- **Operators and evaluators:** Individuals who provide monitoring and formal/informal testing, to evaluate and verify system performance, relative to both technical and socio-technical requirements.
- **External groups:** People who provide formal and/or quasi-formal norms or guidance for specifying and addressing AI risks. These include advocacy groups, civil society, and standards organizations.
- **General public:** Individuals most likely to experience the direct impacts of AI technologies, positive and negative.

AI RMF Core as a complement to paragraph 2 of Article 9

The EU and US documents emphasize different levels of autonomy in the generation of risk management frameworks. The second paragraph of Article 9, on the one hand, requires the creation of

1 The NIST AI RMF will officially be published on January of 2023. Therefore, the information herein is based on a first draft available on this site.

a process that identifies, analyzes, estimates, and evaluates pre- and post-deployment risks followed by the adoption of risk management measures. As long as entities comply with these conditions, they are free to generate their own risk management systems.

The implementation of the AI RMF, on the other hand, relies on the alignment of incentives. In other words, an entity's self-interest plays an important part in the adoption of this framework. For instance, entities can gain reputational good-will from customers, and protection in judicial proceedings by using best practices that serve as a mitigating factor in sentencing; they can also synchronize the practices of supply chain providers, etc. Furthermore, instead of a blank slate, NIST created a structure to guide entities into managing the risks of their own AI systems. EU stakeholders could adopt elements of this structure to minimize heterogeneity between risk management practices, and to fulfill the requirements of Article 9. This could catalyze both direct and indirect benefits - for instance, by decreasing the transaction costs of operating within the US and EU.

The NIST AI RMF guidelines are made up of four functions: map, measure, manage, and govern. Each is meant to organize a particular area of high-level risk management :

- **Map:** Context is recognized, and risks related to that context are identified;
- **Measure:** Identified risks are assessed, analyzed, or tracked;
- **Manage:** Risks are prioritized and acted upon based on a projected impact;
- **Govern:** A culture of risk management is cultivated and present.

These functions are further divided into categories and sub-categories, in order to specify desired actions and outcomes. For example, the AI RMF requires AI system stakeholders to take the following steps under the measure heading:

Measure Function

Category	Sub-category
Appropriate methods and metrics are identified and applied	Elicited system requirements are analyzed.
	Approaches and metrics for quantitative or qualitative measurement of the enumerated risks...are identified and selected for implementation.
	The appropriateness of metrics and effectiveness of existing controls is regularly assessed and updated.
Systems are evaluated	Accuracy, reliability, robustness... are measured, qualitatively or quantitatively.
	Mechanisms for tracking identified risks over time are in place....
Feedback from appropriate experts and stakeholders is gathered and assessed	Subject matter experts assist in measuring and validating whether the system is performing consistently with their intended use and as expected in the specific deployment setting.
	Measurable performance improvements (e.g. participatory methods) are identified, based on consultations.

A common vocabulary

All relevant parties would benefit from reaching a consensus on the definitions of key terms related to AI system management. While the EU AI Act and the AI RMF are under development, decision-makers for both initiatives should seize the opportunity to develop a shared understanding of core AI ideas, principles, and concepts, and codify these into a common transatlantic vocabulary. As seen below, the first draft of the AI RMF has begun this process, by identifying where both initiatives are in agreement, and where they diverge. But further work is needed to reach a broader consensus (see page 8 of this link).

	AI RMF	EU AI ACT
Shared	Accountability Safety Privacy Transparency Fairness	
Divergent	Accuracy Explainability Interpretability Managing bias Reliability Resilience or ML security Robustness	Data governance Diversity Environmental and social well-being Human agency and oversight Technical robustness Non-discrimination