Request for Information:
Artificial Intelligence Risk Management Framework

April 29, 2022

Overview

Google welcomes the opportunity to provide comments in response to the National Institute of Standards and Technology's (NIST) Request for Information (RFI) on the Artificial Intelligence Risk Management Framework (AI RMF or Framework).

Google supports NIST's effort to create a risk management framework that is voluntary, flexible, and can adapt as the AI ecosystem continues to evolve, while providing clear guidance to developers and deployers of AI systems on how to manage risk. The current draft of the framework provides a useful starting point for organizations seeking to better understand how to incorporate AI risk management into their existing structures and governance processes. The four functions of the "AI RMF Core" provide a helpful high-level overview of how to structure AI governance within organizations, and are aligned with Google's and other stakeholders' own approaches.

As the framework acknowledges, risks are deeply dependent on the specific application and context of use, and vary widely across the broad spectrum of AI technologies and applications. As the AI ecosystem continues to mature, new techniques are developed and applied to new applications, the framework must continue to evolve as well.

Furthermore, as the framework notes, many of the risks and impacts the framework is intended to address are not well understood and cannot be easily measured. There is also little common understanding of how risk thresholds should be established by organizations developing and deploying AI. Google supports NIST's plan to release a companion document citing responsible practices, and NIST's ongoing efforts to develop benchmarks and metrics for AI systems, which will complement the high-level framework and enable consistent responsible practices.

Suggestions to further strengthen the framework

The draft framework could be further strengthened by aligning its taxonomy with widely used definitions in the AI industry, clarifying the roles of different stakeholders in the AI value chain, expanding on the difference between fairness and unfair bias, acknowledging the trade-offs often required between different aspects of trustworthiness, and clarifying how the AI RMF aligns with other frameworks and standards, including the NIST Cybersecurity and Privacy Frameworks and international standards. Specifically:

> *Align the AI RMF taxonomy with widely used definitions in industry and existing standards*

The draft framework includes a number of terms and provisions that would benefit from further clarification to ensure that stakeholders in the AI ecosystem can understand risk management as applied across different use cases and industries. The way some terms are defined in this draft do not align with how they are used in industry. For example, the term "accuracy" may be misleading with regard to assessing AI systems as it has a specific technical meaning. For purposes of risk mitigation, it may be more effective to replace "accuracy" with another term, such as "correctness" or "usefulness," to avoid confusion. Similarly, "interpretability" is widely used in industry to describe technical methods of understanding how models operate and connect inputs to outputs.

More generally, the three-class taxonomy of risks and characteristics in the framework (dividing risks and attributes into technical, socio-technical and guiding principles) is somewhat confusing, as all these attributes have both technical and human components. Assigning a single designation to a given term may not accurately capture this. As noted above, "interpretability," labeled as "sociotechnical," is widely used in industry to describe technical attributes of AI systems, while "security" often depends on deployment decisions and user behavior, as well as aspects of model design, suggesting it may have attributes of both technical and human components.

This also extends to descriptions of key actors in the AI value chain, which should be explicitly defined in the framework. We recommend aligning with definitions from existing international standards. For example, ISO 31000 defines stakeholder as a "person or organization that can affect, be affected by, or perceive themselves to be affected by a decision or activity".

Clarify the roles of different stakeholders in the AI value chain

Figure 1, which outlines key stakeholder groups, appears to be silent or ambiguous on inter-dependencies of stakeholders. Each stakeholder will be responsible for managing risks associated with potential application of systems throughout the AI system lifecycle. Risks associated with general-purpose systems are particularly unique in that a system may be made available and further modified in a way that could create a different risk classification.

The deployer may further depend on product documentation from the developer of the system. This is key in the context of Risk Measurement (section 4.2). For example, the provider will often not have access to the operational data necessary for post-market monitoring if the AI system has been put into operation by another entity. To address this, we recommend ensuring that responsibilities, risks, and interdependencies between key stakeholders are clarified in the framework.

The framework would also benefit from greater clarity around how certain characteristics of AI systems impact different stakeholders. For example, paragraph 5.2.1 on page 11 states "Explainability refers to the user's perception of how the model works." However, explainability is not just important for users, but also for deployers, and even different stakeholders within developer organizations - each requiring a different form of explainability. Further, the term "stakeholder" better articulates the intended audience which could include providers, deployers, etc.

Expand on fairness guidance and the distinction between fairness and unfair bias

Paragraph 5.3.1 on page 13 describes the complexity around concepts of fairness, but provides little practical guidance in terms of how organizations should assess and mitigate potential fairness issues. In particular, it states ""Fairness is increasingly related to the existence of a harmful system, i.e., even if demographic parity and other fairness measures are satisfied, sometimes the harm of a system is in its existence." First, this claim is misleading because harms from AI systems derive from how they are designed, deployed and used, rather than from the mere "existence" of the system itself. Facial recognition, for example, can be used for a wide range of applications, from unlocking your phone to tracking criminal suspects, with very different risks and potential harms associated with these different applications.

Second, this statement distinguishes between the elimination of unfair bias and the fairness of the system, and further states that "absence of harmful bias is a necessary condition for fairness," seemingly implying that it is necessary but not sufficient, but the framework does not expand on what other criteria stakeholders should use to evaluate fairness in AI systems, or provide guidance on how to do so. As a starting point, international standards bodies such as ISO are in the final phases of publishing multiple standards around ethics, fairness, and bias (i.e. ISO 24027 Bias in AI - published and ISO 24368 AI Ethical and Societal Concerns - est. publication 2022). The AI RMF should provide more clear guidance on how to evaluate fairness in AI systems, including by referencing international standards, including those referenced above, to provide a set of best practices around AI fairness and bias for organizations.

Acknowledge trade-offs between different aspects of trustworthiness

Many of the attributes of trustworthy and responsible AI outlined in the framework can be in tension when it comes to designing real world products. For example, simpler techniques like static decision trees or statistical models are more easily interpreted and explained than systems that utilize deep neural networks (DNNs), but are often less accurate, and can be less resilient. Furthermore, as referenced in NIST's Special Publication "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," these simpler and more explainable models can actually exacerbate unfair bias "because restrictive assumptions on the training data often do not hold with nuanced demographics."

In light of the trade-offs that often arise between different attributes of trustworthy and responsible AI systems, it is important that the framework provide the flexibility for organizations to determine how best to balance competing equities and attributes for specific products and applications. NIST could also incorporate ISO standards addressing these points as they are released, such as the AI and privacy that will be released in the coming months. In addition, NIST could provide guidance for organizations on how to identify and balance tradeoffs among different attributes of the framework, leveraging expertise from industry, academia and civil society.

Clarify how the AI RMF aligns with and can be integrated with other standards and frameworks

As the AI RMF notes, "this framework aims to fill the gaps related specifically to AI" in regards to existing frameworks and standards for privacy, cybersecurity, safety, and infrastructure. It would be beneficial to add references to existing NIST frameworks,

including the Cybersecurity and Privacy frameworks (including key terms), and highlight how these documents work together to promote best practices. Specifically, there are numerous references to privacy (section 5.2.3) throughout this draft in which the NIST privacy framework could be cited.

It would also be beneficial to explicitly align the framework with recognized ISO standards on risk management (i.e. ISO 23894 and ISO 31000) to promote cohesion between ISO and NIST frameworks. Given the cross-border nature of the digital economy, AI regulatory frameworks and technical standards should ideally operate across nations and regions. Increased global alignment on AI regulation, including in the context of trade, will help to facilitate the understanding, adoption, use, and interoperability of AI technologies across different jurisdictions. Internationally recognized voluntary consensus standards such as ISO 31000 Risk Management (published) and ISO 23894 AI Risk Management (in the process of being published, target date of 2023-03-05) should be used as a guide when developing this draft. We also recommend that NIST consider initiatives such as MLPerf, which is developing benchmarks to assist evaluations of training and inference performance for hardware, software, and services.

Conclusion

The AI RMF represents an important step forward in advancing consistent responsible practices across the AI ecosystem, and providing clear guidance to stakeholders on how to understand and manage risk in AI systems, including integrating AI governance into existing governance structures and processes. Since Google released our AI Principles in 2018, we have developed and refined our own governance structures and processes, and built our library of responsible AI practices. This is an iterative process, and continues to evolve as our understanding of the benefits and risks of AI systems matures, and as our products utilize AI in innovative new ways. Google welcomes the opportunity to share insight based on our experience, and to learn from and engage with other participants. We look forward to continuing to work with NIST and our fellow stakeholders on these important matters.