MITRE

Request for Information on Artificial Intelligence Risk Management Framework

Response to Agency/ID:

National Institute of Standards and Technology: Docket No. 210726-0151

September 14, 2021

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

Approved for public release. Distribution unlimited 21-2807

© 2021 The MITRE Corporation.
All rights reserved.

Table of Contents

Introduction

The MITRE Corporation is pleased to respond to the Request for Information (RFI) on "Artificial Intelligence Risk Management Framework" released on behalf of the National Institute of Standards and Technology (NIST) (Docket No. 210726-0151).

As a not-for-profit organization, the MITRE Corporation works in the public interest to tackle difficult problems that challenge the safety, stability, security, and well-being of our nation through the operation of multiple federally funded research and development centers and labs and through participation in public-private partnerships. Working across federal, state, and local governments — as well as industry and academia — gives MITRE a unique vantage point. MITRE works to discover new possibilities, create unexpected opportunities, and lead by pioneering research for the public good to bring innovative ideas into existence in areas such as artificial intelligence (AI), intuitive data science, quantum information science, health informatics, policy and economic expertise, trustworthy autonomy, cyber threat sharing, and cyber resilience.

MITRE has a long history of partnering with federal agencies to apply the best elements of AI and machine learning (ML) while developing and supporting ethical guardrails to protect people and their personal data. Our team is committed to anticipating and solving future needs that are vital to the success and safety of the public and the country.

MITRE values the opportunity to contribute to this important discussion. We are also eager to engage further with NIST and the community it is leading. In the following pages, we have focused on answering two specific topics from the RFI:

- **T2:** How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: Accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI.[1]

- **T7:** AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST should consider to ensure that the AI RMF aligns with and supports other efforts.

[1] Note that this response is also relevant to topics T3, T4, and T5.

Response to Topic 2

T2: *How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: Accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI.*

A comprehensive AI risk management framework (RMF) should also address the unique ways that AI systems can be exploited by malicious attackers.

Much of the current work addressing AI risk management focuses on the important and challenging issues surrounding the development of AI systems with key characteristics — including high accuracy, explainability and interpretability, reliability, privacy, robustness, and resilience — while minimizing harmful biases or outcomes from misuse of the AI. In addition to these challenges, a comprehensive AI risk management framework should also address the unique ways that AI systems can be exploited by malicious attackers.

AI and ML systems represent a unique and rapidly expanding attack surface with associated risks not addressed by traditional cyber security controls or risk management frameworks. These attacks are distinct from traditional cyber-attacks and can be a highly effective means of exfiltrating sensitive consumer data, stealing intellectual property, or otherwise subverting the AI system for malicious purposes, even when the AI models meet traditional assurance standards and are effectively secured from cyber-attacks.

These attacks range from targeted attacks that subvert AI-enabled facial recognitions systems, affording malicious access to financial data and user accounts, to intentionally poisoning training data being used to develop algorithms for trading on the stock market. Unlike cybersecurity, organizations seeking to secure their AI and ML systems do not have large amounts of historical threat intelligence in which to ground risk assessments that would help focus limited resources; yet, AI service developers and providers should not wait for a major crisis before they begin addressing these threats.

In order to address this problem, MITRE, utilizing input from Microsoft and a broad coalition of private sector companies, developed the Adversarial Threat Landscape for AI Systems (MITRE ATLAS).[2] ATLAS was developed by synthesizing real world case studies, voluntarily submitted by a broad range of industry partners, which detail real attacks conducted against AI systems. The ATLAS team used them to build a robust

[2] https://atlas.mitre.org/

and common taxonomy of attack tactics and techniques that map to a broad range of contexts in order to empower security analysts across industry and within the government to detect, respond, and remediate threats against ML systems.

Since its release in the Fall of 2020, ATLAS has rapidly impacted AI security across multiple industry verticals, with security teams from companies as diverse as Microsoft, Bosch, Ant Financial Group, and Airbus testing their own AI systems using the ATLAS model. After using the model, these companies contributed the results of their tests as case studies to further improve the ATLAS knowledge base. Additionally, this collaboration resulted in Microsoft's release of a powerful open-source tool set called "CounterFit"[3] based on ATLAS,[4] which is a significant step forward towards mutual security practices. CounterFit gives companies who cannot afford dedicated AI security practitioners a robust ability to evaluate their own AI enabled systems against known AI threats.

The ATLAS program, created and led by MITRE, highlights some of the intentionally focused collaboration taking place across the AI industry to rapidly understand, detect, and mitigate these threats, and could be beneficial in informing NIST's AI RMF development in the area of AI security. Considering the unique ways that AI systems can be exploited by malicious attackers is also responsive to T3 as an important set of principles. ATLAS further serves as a prime example for T4 regarding how these types of AI risks are being incorporated in organizations across the AI community at large. The work of ATLAS also applies to T5, providing frameworks, methodologies, tools, guidelines, and best practices to identify, assess, prioritize, mitigate, and communicate AI security and adversarial vulnerability risks. Finally, ATLAS should be considered by NIST under T7, based on the criteria below.

Response to Topic 7

***T7:** AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST should consider to ensure that the AI RMF aligns with and supports other efforts.*

A model for assessing an organization's level of AI maturity should be considered in an overall AI RMF.

An organization with a high level of AI maturity is less likely to deploy an AI solution that will present a risk to its mission. Such an organization can more easily demonstrate

[3] https://www.microsoft.com/security/blog/2021/05/03/ai-security-risk-assessment-using-counterfit/

[4] At the time when CounterFit was released the ATLAS framework was still being called the "Adversarial Machine Learning Threat Matrix" ATLAS is a rebranding of that matrix.

compliance with applicable standards, best practices, and regulations – leading to more stable and reliable AI-enabled solutions. MITRE recommends that the AI RMF take into account the AI maturity level of any AI-deploying organization.

There are numerous approaches for measuring an institution's AI maturity, including the Government Services Agency's AI Capability Maturity Model[5] and MITRE's Organizational AI Maturity Model (AI MM).

The MITRE Organizational AI MM and its associated AI Assessment Tool (AI AT) serve to assess and guide effective readiness, adoption, and use of AI/ML across an organization. The AI MM defines the dimensions and levels of AI maturity and provides the foundation for an assessment using the AI AT.

The AI MM identifies five broad types of obstacles, known in the model as "pillars," which can impede the adoption and effective use of AI:

- Strategy and Budget: Does the organization have an implementation plan, partnerships with other agencies, and governance processes?
- Organization: Does the organization have a risk-tolerant culture that supports innovation, defined roles for AI development, and a plan for recruiting, training, and retaining AI talent?
- Technology Enablers: Does the organization have an approach for identifying and using new AI innovations, a method for verifying and validating proposed solutions, and the computing power needed to develop, deploy, and maintain these solutions?
- Data: Does the organization have data governance processes and audit capabilities in place to monitor compliance with AI/ML standards, for sharing, and for appropriate storage, retention, and access control?
- Performance and Application: Does the organization have an approach for integrating AI into business workflows, have monitoring processes in place to measure how well they support strategic outcomes, and have a due diligence process that promotes calibrated user trust and protects against unintended consequences in AI solutions?

Operationalizing AI and ML is not easy. Although many enterprises initiate AI and ML projects, the results often fall short of expectations, highlighting the need for organizations to prepare for AI and ML initiatives.

> *"…nearly 8 of 10 organizations engaged in AI and machine learning said that projects have stalled, according to a Dimensional Research report. The majority*

[5] https://coe.gsa.gov/2020/10/28/ai-update-2.html

> *(96%) of these organizations said they have run into problems with data quality, data labeling necessary to train AI, and building model confidence."*[6]

The combined MITRE Organizational AI Maturity Model and Assessment Tool can help organizations see more than one aspect of AI adoption, providing a key element to an AI RMF enabling assessment of an organization's AI-readiness, and thereby promoting a systematic path to success.

MITRE Contact Information

Please contact us to share your feedback or to learn more about MITRE's work on ATLAS and the Organizational AI Maturity Model.

Corresponding Contact:
Eric Bloedorn, Chief Scientist for AI Adoption

[6] 96% of organizations run into problems with AI and machine learning projects. Macy Bayern. May 24, 2019.