AI Village (AIV) Response to the NIST RFI on Artificial Intelligence Risk Management Framework (AI RMF)

Adelin Travers, University of Toronto & Vector Institute for AI
Anita Nikolich, University of Illinois at Urbana Champaign
Abhishek Gupta, Founder and Principal Researcher, Montreal AI Ethics Institute
Stella Biderman, EleutherAI
Brian Pendleton, Marymount University
Erick Galinkin, Rapid7
Brian Martin, AbbVie
John Irwin, Nvidia's AI Red Team
Anusha Ghosh, University of Illinois at Urbana Champaign

Preliminary Definitions

Artificial Intelligence (AI): A set of automated techniques encompassing formal logic, inductive reasoning,and statistical (Machine Learning) based automated categorization, decision making, forecasting and regression analysis.

AI System: a system that includes at least one AI component. Importantly, our definition includes both “end-to-end AI systems” where the AI *is* the entirety of the system as well as systems that combine AIs with non-AI components, such as human reasoning, traditional algorithms, interactions with databases, and pre- and post-processing

Executive Summary

The AI Village welcomes the NIST RFI on Artificial Intelligence Risk Management Framework. Such a framework is long overdue to confidently adopt AI in production systems, including mission-, security- and safety-critical applications.
Our response is articulated on five axes: (i) concrete and current risk, (ii) actionability, (iii) organizations focus, (iv) case specific and (v) existing and upcoming standards.

More specifically each of those five axes tries to answer a specific need of the framework:

1) Concrete and current risk: This addresses the lack of adequate and easily manipulable risk definition in the case of AI systems and the current divide between risks faced today in production by organizations and the hypothetical and future risks considered in a number of Reliable AI academic publications. As a strongly industry focused organization, we believe the former should be given immediate priority and the latter should be used primarily as a support vector for threat anticipation.
2) Actionability: This axis addresses the current difficulty in making AI risk assessments and mitigation operational including both the methodological and tooling gaps in

conducting end-to-end AI risk assessments. We identify gaps in enforceability as well as current risk and testing methodologies that will need to be filled ahead of proper AI risk assessment engagements.

3) Organizations focus: By this axis we address the incentives, budget obtention, team reorganizations, business integration and related organizational challenges that implementers will eventually face in their respective organizations. The framework will need to provide answers on those points to ease its deployment in complex governance schemes.
4) Case specific: By this axis we seek to address specific challenges that will arise on a case by case basis and which a generic framework might be incapable of covering. In our response, we take specific attention in addressing both large and small organizations, specific industries like finance, ICS and pharmaceuticals etc.
5) Existing and upcoming standards: Our submission draws heavily from a number of existing and upcoming security, privacy and AI regulations that are listed in appendix A. We curated this list based on their relevance and implementation likelihood in target organizations. The upcoming AI Risk framework should tightly integrate with those to limit conflicts and necessary additional work, thereby facilitating adoption.

Detailed responses to RFI:

1. **The greatest challenges in improving how AI actors manage AI-related risks – where "manage" means identify, assess, prioritize, respond to, or communicate those risks.**

    a. There exists a lack of coordination and alignment between current organizational governance processes and the management of an AI lifecycle. A disconnect further exists between the governance and oversight of different types of AI systems. An AI system used in health care is fundamentally different from one used in industrial controls, for example. NIST must be careful in not overgeneralizing the means to identify and respond to AI risks - an RMF could end up being too watered down and generic to be useful across industries.
    b. A lack of meaningful dialogue between industry and academia has led to a disconnect in the appreciation of AI risks. Academic papers often highlight AI risks that aren't immediately salient to real world deployments yet, as with anything security-related, remain a distant possibility. NIST should aim to decouple these by categorizing risks in a fashion that distinguishes risks realizable in current systems and hypothetical risks that could arise should a number of (currently unrealistic) assumptions be met. In the latter case, NIST should provide an evaluation of how likely these assumptions are to be met in the future.
    c. The diffusion and fragmentation of responsibility across functional areas within an organization with regard to AI means that it is not clear who will functionally own

the creation and implementation of an organization's RMF - security, IT, data science, product, compliance. NIST can and should provide general guidance on this.

d. Most organizations lack internal or external channels over which to communicate AI risks.
e. There is a notable lack of tools to identify AI risk as well as standard tools to perform and scope assessments. This will ultimately hinder progress on an AI RMF and practical assessment implementation.
f. There is a lack of security controls around AI systems. Additionally, there are organizational disconnects between AI developers and those who design and implement security controls.
g. There is a lack of appropriate framing of AI risks (i.e., impact, likelihood, severity, mitigation methods). This is best explained by the lack of definition and distinction between model level risks and system level risks. Interaction in-between components may lead to a different risk profile at the level of the AI component (model) and of the system.
h. The lack of external assessment capabilities leads to a conflict of interest, especially when the assessment is performed by the AI system producer/vendor itself. Lessons should be drawn from the role of private risk rating agencies in the 2008 financial crisis as these external assessment capabilities are designed.
i. Organizational challenges are fundamental. There is a highly variable composition of what makes up an AI Team - engineers, marketing, mathematicians, IT experts, etc. We suggest that NIST designs a taxonomy/a basic composition (i.e. a RACI chart) of an "AI Team". We propose that an emphasis on AI security and risks be made available through dedicated roles in this team. This will be especially important when an enterprise constructs an AI Security team. An afterthought is often the legal team which needs to understand the nuances of data sharing. We advise NIST to also include compliance and legal aspects in an RMF as well as an AI Team taxonomy.
j. Plain language is essential. Semantics and definitions of AI and AI-adjacent functions get confusing and vary considerably across organizations. We encourage NIST to follow the Plain Writing Act of 2010 which was designed "to improve the effectiveness and accountability of Federal agencies to the public by promoting clear Government communication that the public can understand and use."

2. **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. All the listed attributes considered for the Framework should be considered in the AI lifecycle context, i.e. as attributes that are continuously evaluated over the

course of the AI lifecycle rather than as static gates/checkpoints to more comprehensively achieve Reliable AI

b. The AI characteristics listed (accuracy, explainability etc) all need prescriptive ways of assessing data bias with regard to them. NIST should take the lead in prescriptive guidance for data bias assessment, and we commend NIST for seeking input to the NIST SP 1270.
c. AI Teams need quantitative guidelines for assessing and identifying model drift when a model is retrained periodically (for currency/temporality/freshness), and for the identification of software vulnerabilities. AI is often a blackbox added into a software stack, complicating the finding of vulnerabilities.
d. Expected inputs should be tracked as a characteristic of an AI because anything that doesn't comply with it will result in unexpected outcomes potentially invalidating the rest of the characteristics.
e. All potential outputs should be a characteristic of an AI. Any developer of a component making a dependency on an AI needs to know what may be provided as an input to their component, in order to avoid various software bugs, some of which can result in serious security issues.

3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: Transparency, fairness, and accountability.

1. The principles of transparency, fairness, and accountability are agreed upon by most organizations. We also propose the additional principles of openness, privacy and security. Although the definitions of transparency, fairness, and accountability could be expanded (in that order) to encompass these latter sub-principles, this extension is not widely recognized in the academic literature.
2. Security and privacy go hand in hand as fundamental principles of AI trustworthiness. An AI that can be trusted should offer privacy to the individuals on whose data the system may have been trained, and should offer integrity in its predictions. Whenever possible, predictions should offer an appropriate level of confidence, and systems should do their best to inform users what data is being leveraged and what objective is being optimized. Appropriate levels of confidence are nonetheless hard to define and, where possible, NIST should provide guidance on how to define appropriate levels of confidence.
3. Instead of a redefinition of principles, the NIST framework should concentrate on directly reusing pre-existing sets of comprehensive principles and providing guidelines on how to implement these principles. An organization should then align this set of principles with its internal organizational structure and processes.
4. Human rights and Privacy rights are non-optional attributes. For AI discovery in life-critical areas such as pharmaceuticals or medicine, data sharing should be integrated into the principles. Questions nonetheless arise with regards to competency. Indeed, data sharing between organizations is contingent on

technicnical and governance competency of the organization. Data removal should also be considered if a participation consent is retracted.

4. **The extent to which AI risks are incorporated into different organizations' overarching enterprise risk management—including, but not limited to, the management of risks related to cybersecurity, privacy, and safety.**

    a. Deep integration of AI risk considerations into existing risk management approaches in the organization is preferable for expediency in the application of risk-mitigation measures and in getting internal financial and other support, including resources. This is necessary to move ideas into practice rather than leaving them as line items that need to secure funding before adoption. This is an essential consideration given the relative lack of adoption of such deep integration and real-world deployment of these risk-mitigation approaches given the structural challenges of fragmentation in larger organizations. The concern also manifests itself in smaller organizations where a limitation in terms of resources restricts investments in a multitude of risk-mitigation strategies. A deep integration can then make it easier to justify paying attention to these novel concerns and construct a holistic strategy that is backed by the existing resources allocated to risk management within the organization.
    b. Such a deep integration also calls forth a demand for interdisciplinary expertise, namely one that straddles existing risk-management domains like cybersecurity, privacy, and safety with artificial intelligence, especially combining that with domain expertise since the manifestation of AI risks can differ based on the domain of application. This can be done through more holistic training both in the traditional domains where expertise in AI can be built up and in the domain of AI pedagogy where instruction on cybersecurity, privacy, and security is provided as a regular part of the curriculum.
    c. Justification for investment in this integrated approach is supported by the fact that AI risks contain immediate business and reputational risks. Examples of popular failures leading to immense reputational harm (Microsoft Tay, Google Photos, Amazon's resume scanning experiment etc). In other cases, failures from AI systems can harm the ability of the organization to bid for contracts from the government which create direct financial impacts.
    d. Folding in the consideration of AI risks within existing functions of an organization like Compliance, Risks, Legal, and Communication generate benefits along the lines of utilizing their existing machinery to put AI risk mitigation strategies effectively into practice quickly and they negate the need to create a new organizational unit which comes with challenges of resource allocation, but more importantly creates yet another silo limiting the efficacy of risk management in AI which necessitates considerations to be accorded across a broad set of stakeholders that span different functions within the organization.
    e. AI risks will directly impact the risks of the AI system (in our end-to-end definition), so industry teams will incorporate them into overarching enterprise

risk management. However, providing a means to incorporate AI risks may prevent each industry team from finding a different way to glue systems together in a holistic manner.

5. **Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above.**

    a. It will be hard to have a single set that comprehensively captures all the ***operational aspects*** as specified in the desiderata for the Framework. Instead, developing scenario- and domain-specific guidance from a broader Framework is what will lead it to being actionable.
    b. We do not believe the current NIST Risk Management Framework meets the minimum attributes required to properly evaluate an AI system.
    c. A framework must be enforceable to have impact. We draw a parallel to the impact that GDPR had with its 4% rule, a major change from prior privacy laws for computer systems which have existed in Europe since at least 1978 (France). Agencies such as the FTC should be charged with enforcement and levying of sufficient penalties for non-compliance.
    d. There should be a semantic layer specific to various industries which may deploy AI in mission critical systems: finance, pharmaceuticals, IT, ICS, transport, energy etc.
        i. NIST should enable the community to run through the scenarios with different drivers to see what ones are generalizable for a final RMF.
        ii. Current RMFs and other frameworks often lack specifics on how they should be implemented across industries. (EG) - I.e., scope of HIPAA is specific. CSF is pretty specific.
        iii. Current definitions of AI risk are nebulous at best. Risk should be conceived both at the level of the AI component but also at the system level to provide adequate granularity. There should be definitions of both risk **to** AI and Risk **from** AI. Adequate definitions of risk are of paramount importance but existing frameworks are not readily applicable!
        iv. Current frameworks are either too broad (NIST) or too specific (HIPAA, finance specific frameworks, etc).
        v. Both the PTES and NIST SP 800-115 provide guidance on security testing and assessment that is not applicable to AI systems. This is among others because of the probabilistic nature of machine learning systems.
    e. The EU has released a draft regulation (see appendix A) on trustworthy AI which should be taken into account, debated and adapted to the specifics of the US.
    f. We stress the relevance of biometrics ISO standards (24745 and 19792) as they are among the few places where model risk is taken into account both individually and in a system context (e.g. these include risk to databases holding

biometric features). A very recent ISO standard (24029) has also provided a methodology for AI trustworthiness assessments.

6. **How current regulatory or regulatory reporting requirements (e.g., local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles**.

   There exist a multitude of relevant industry standards: ISO 2700x, 24745, 19792. We provide an extensive list in Appendix A. Observe that some of these are not legally binding but in practice have similar practical effects as regulations. Indeed, organizations which don't follow these frameworks may lose market competitiveness or be barred from operating.

   a. The identified frameworks tend to predate the use of AI. These standards would benefit from either (a) a retrofit of interactions with AI or (b) precisions on how they are applicable to new AI techniques through the upcoming NIST AI framework.
   b. Financial standards and regulatory requirements are peculiar in that they do not mention AI explicitly but rather "new technologies" for FATF standards and risk models for Basel III. In both cases AI is covered in its broad understanding. Moreover, traditional statistical models, time series and monte carlo analysis may eventually be replaced with cutting edge AI/machine learning systems which will have to be properly assessed.
   c. A cohesive Framework can help align the regulatory requirements at various levels, if there is misalignment, especially for solutions that have cross-jurisdictional scopes, it will become incredibly hard to get consistent reporting on the key attributes that one would want from the relevant stakeholders of that AI system
   d. The US needs clear, national-level standards with regard to data handling and data privacy to not just encourage but regulate transparency.
   e. Data storage for training data should have strict, enforceable controls. Examples of such controls can be found in ISO 19792 & 24745 which relate to the protection of biometric authentication systems.
   f. We firmly believe that FIPS 140-2/3 inspired classification of the architecture and implementation process of AI techniques would prove beneficial. Additionally an Evaluation Assurance Level definition for the conduct of AI risk assessment would help industry adoption of better AI practices.

7. **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. The most relevant existing AI frameworks outside of the NIST AI attack taxonomy are International.

b. We highlight the Canadian AI Impact Assessments, especially those similar to the Algorithmic Impact Assessment Tool put out by the Canadian Government as good examples of such assessments. (https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html)
c. We highlight ISO 24029 and, in general, the standards to be published as part of the European document "AI Watch: AI Standardisation Landscape state of play and link to the EC proposal for an AI regulatory framework". The comprehensive list of other related frameworks is provided in Appendix A.

8. **How organizations take into account benefits and issues related to inclusiveness in AI design, development, use and evaluation—and how AI design and development may be carried out in a way that reduces or manages the risk of potential negative impact on individuals, groups, and society.**

a. Organizations can answer the inclusiveness and other attributes-related questions by following the motto "nothing about us without us" and implementing a design process like Community Juries (https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/#:~:text=Community%20jury%2C%20an%20adaptation%20of,use%20cases%20and%20product%20design).
b. A number of larger organizations have entire teams dedicated to this effect. Smaller organizations may not have the resources for hiring internal teams to oversee and assess ethics or AI security. This can disadvantage smaller organizations (especially if they are committed to equity and security of AI). Industrial Control Systems, finance and pharmaceuticals also face unique challenges linked to their activity sectors. More specifically, some financial organizations are considering machine learning models such as Deep Neural Nets and Federated Learning (or combinations of blockchain and DNN) to respond to among others FATF regulations and exchange information on illegal transactions.
c. The Framework should ensure that its requirements can be graduated, for instance with maturity models, based on the size and type of organization. The framework should pay attention to potential loopholes. It might be valuable to at times use a one size fits most approach. Impact of risk may have high variance and should be taken into account. A risk-based approach to diversity and inclusion should also be considered.
d. An AI Impact assessment methodology including each intended use should be woven into the framework. This methodology should include intent-based risk limitation and organizational and industry context.
e. "Tech Against Terrorism", a UN supported counter terrorism project, pushed for developing tech for compliance and threat monitoring in/with larger organizations and making it open-source. This is one example of how less-resourced organizations could also adhere to stronger standards.

f. AI assurance efforts should employ a “sterile environment” analogy to take into account robust supply chain, as well as physical, and traditional software security, all of which may impact AI resilience.

9. **The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, “AI RMF Development and Attributes”).**

    a. Efforts seem in their broad strokes appropriate to us. Nonetheless, we advocate for a stronger emphasis on **machine learning security**. Indeed, it is a horizontal concern spanning all other attributes and is, in that sense, foundational. The focus on cybersecurity is potentially optimal as, by trying to do too much, the framework may lose focus.
    b. We consider point 4 (adaptability to organizations) to be a double edged sword as the framework might no longer become actionable. The framework could and should behave like the combination of the NIST Cybersecurity framework with the CIS controls, one providing policy guidance and the other operational controls.
    c. We believe that a checklist of common biases, security issues and risks would be a beneficial output from the Framework. This would help with providing rigorous and comprehensive AI risk assessments similarly to the role that CIS 20 controls, or the OWASP top 10 provide for traditional security program, operational policy and security assessment implementation.

The AIVillage* as an organization takes the position that points 11, 12 and to a lesser extent 10 are out of scope for a risk framework. As an organization, we believe that these points could be addressed separately for instance in the form of special publications referred to in the framework. The AIVillage would happily respond to separate and specific requests for information on these points.

*The AI Village is a community of hackers and data scientists working to educate the world on the use and abuse of artificial intelligence in security and privacy. We are academics, IT specialists, security experts, students, philosophers, and concerned citizens.

Appendix A

The following list of laws, rules, regulations, industry standards and frameworks may affect AI in the US (and organizations with overseas dealings):

- AICPA
- Basel III (wherever AI models are used in Banking)
- CCPA
- CIS Controls/CIS Top 20
- CIS RAM

- COPPA
- FATF standard: (INTERNATIONAL STANDARDS ON COMBATING MONEY LAUNDERING AND THE FINANCING OF TERRORISM & PROLIFERATION)
- Federal Reserve Board SR 11-7
- FedRAMP
- FERPA
- FISMA
- GDPR - for US organizations with overseas dealings that concerns EU residents
- GLBA
- HIPAA (PHI)
- ISO 27001, 27081, 27701,31000, 19792, 24745, 24029
- ITAR
- NERC CIP Standards
- NIST Standards such as SP 800-115
- PCI
- PTES: Pentest Execution Standard
- PII as defined by OMB M-10-23
- SOX
- SOC 1 and 2
- Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS
- AI Watch: AI Standardisation Landscape state of play and link to the EC proposal for an AI regulatory framework
- Canadian Algorithmic Impact Assessment Tool