Response to the NIST RFI on an Artificial Intelligence Risk Management Framework

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421-6426
781-981-5500

Contributors:
Kendra Kratkiewicz, Diane Staheli, William Streilein, Dennis Ross, Michael Yee, Olivia Brown, Sanjeev Mohindra, Paul Metzger, Joseph Zipkin

1 Introduction

NIST has issued a call for feedback intended to inform a draft for an Artificial Intelligence Risk Management Framework (AI RMF). The guidance provided by this framework will help technology developers, users, and evaluators improve the robustness and trustworthiness of the AI systems they work with. MIT Lincoln Laboratory (MIT LL) is responding to your request to provide input and feedback.

To set the stage for our response, we reiterate the importance of risk management for AI systems and summarize high-level concerns. Several unique aspects of artificial intelligence technologies necessitate the development of an AI-tailored risk management framework. As new techniques are developed, and systems are built and deployed, the emergent properties of these systems can result in unintended consequences – particularly in unforeseen circumstances. Often, the models and algorithms used in these systems can be "black boxes", yielding little insight into how and why particular decisions were made, and with little ability to understand what the system has learned or will learn over time. Furthermore, the nature of human interaction with AI systems is a field of active research, and new guidelines and best practices are still being explored. High profile public failures of AI systems with negative consequences for human safety and lack of ethical considerations have eroded trust in AI utilization.

Risk management in general-purpose systems engineering is already complex, and is only compounded by the additional complexity of AI systems development. Further investigation is needed into how systems engineering techniques can be applied to reduce risk for AI. Some specific approaches for investigation could include those for distributed systems (chaos engineering), software engineering (fuzz testing, CI/CD), and cybersecurity (layered defense approach, penetration testing).

We have organized our response to the NIST RFI into five major categories that we believe capture the major components of a risk framework definition and implementation. MIT LL develops techniques, tools, and metrics for assessing AI robustness, resilience, explainability, and ethics. Through organization of workshops, we actively support a growing community of stakeholders exploring AI system robustness. These responses are based on lessons learned across our studies, research programs, and community activities.

1. **Managing risk relative to the AI lifecycle**
   Risk management requires consideration of the full lifecycle - from conception through design, development, testing and evaluation, deployment, and operational monitoring.

2. **Metrics in support of the RMF**
   Measuring risk quantitatively should be a key aspect of the RMF, so that goals can be clearly defined and progress assessed.

3. **Tools for evaluation and assessment of risk, as well as for mitigation through the AI lifecycle**
   The proposed framework should include tools, techniques, and both development and test harnesses that enable application of the risk principles to real-world AI capabilities.

4. **Human Computer Interface (HCI) aspects of AI risk**
   An effective way to manage risk from AI systems is to consider the context that a human-machine team provides. Throughout the AI lifecycle, there are opportunities to leverage HCI insights to assess and manage risk.

5. **Various organizational roles within the RMF ecosystem in support of the AI RMF**
   A variety of organizational types will be required to realize the RMF. While leading edge capabilities will largely come from industry and academia before they are leveraged by mission users, such as the government, FFRDCs can help assess, mature and manage the risk of those technologies prior to transition.

We believe the proposed RMF should not stand alone, but should be included within a general framework for AI engineering that includes not only AI design principles and best practices, but also guidelines for robust, fair, and ethical use and performance. In this way, risk consideration can more closely follow the AI lifecycle throughout rather than being an afterthought.

2 Managing Risk throughout the AI Lifecycle

Risk management requires consideration of the full lifecycle - from conception through design, development, testing and evaluation, deployment, and operational monitoring.

2.1 Pre-design Considerations

Some high-level questions should be addressed during the earliest planning stages of a potential AI-based solution, before design even begins in order to determine if AI is appropriate for the use case at hand and to uncover potential unintended consequences. These issues include:

- **Legal**:
  - Are there any data/privacy regulations that would affect access to necessary and sufficient training data, or that would apply to a deployed system's inputs or outputs (direct or inferable)?
  - Could the system exhibit bias or discrimination against protected classes?
  - Could the developer or operator of the system be held liable for consequential mistakes?
- **Societal and Ethical:**
  - Who might be adversely affected by the system, and what is the potential for societal harm? As described in *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle*, decisions made in many phases could contribute to this harm. (Suresh & Guttag, 2019)
  - Could there be significant public concern or opposition that might be difficult to manage, jeopardize operation, or damage the developer's or operator's reputation (including the government's)?
  - Is it possible or necessary to inform or educate the public in a way that could address foreseen concerns?
- **Consequences and Tolerances:**
  - What are the potential consequences of system errors (i.e., misclassifications, incorrect outputs), whether benign or due to adversarial attack?
  - What are the ramifications of "CIA" attacks or failures?
    - Confidentiality: exposure of confidential information
    - Integrity: manipulated data or outcomes
    - Availability: system inaccessible due to attack or other failure
  - What are the ramifications of inappropriate trust in the system – either overconfidence or lack of confidence?
  - How will consequence severity be quantified (e.g. severity levels) to enable evaluation and comparison? The European Commission's regulatory framework proposal on AI, for instance, defines four levels of risk, from minimal to unacceptable. (European Commission, 2021)
- **Requirements:**
  - What requirements must the system meet in order to avoid, or mitigate to an acceptable level, the risks that come out of these assessments?
  - How should these requirements change as a system progresses through various technology readiness levels (TRLs)?

2.2 Design and Development

Designers and developers of AI-based systems face many challenges in achieving and balancing accuracy, robustness, privacy, and fairness, all in the face of rapidly evolving research. In order to minimize design and development based risks, the following should be considered:

- **Training**: Require "responsible AI" training for AI system designers/developers, similar to training required for human subjects research. Illustrate such training with real-world failure examples, such as the Microsoft Tay chat bot or Google's racially biased face recognition.
- **Documentation**: Require minimum standards of documentation (across all phases) to facilitate reproducibility, testing, auditing, and generally more reliable engineering. Datasheets for datasets (Gebru, et al., 2018) and model cards for model reporting (MItchell, et al., 2019) are two examples of proposed, standardized documentation.
- **Best Practices**: Require consideration of and adherence to best known practices, consulting a checklist or guide (across all phases). For instance, the NeurIPS conference provides a checklist of guidelines for authors who should answer each question with yes/no/NA and provide justification of the answer. (NeurIPS, 2021)

During design and development, it is important for best practice guidelines to help developers identify potential vulnerabilities or areas of concern, and appropriate mitigations. A non-exhaustive list of such concerns includes:

- **Data**:
  - Input sensor reliability
  - External data dependencies, data provenance, and data poisoning
  - Data selection with respect to representativeness, comprehensiveness, bias, filtering, and sampling
  - Data labeling and metadata
  - Data fusion
  - Data normalization, encoding, and featurization
  - Data partitioning
- **Learning algorithm:**
  - Suitability for task
  - Randomness and reproducibility
  - Noise
  - Hyperparameters
  - Overfitting
  - Catastrophic forgetting
  - Confidentiality (algorithmic leakage)

Insider threat should be considered in all phases, including access to the data and learning algorithms.

2.3 Testing and Evaluation

Testing and evaluating AI-based systems is equally challenging, and there is a dire need for establishing scientifically rigorous best practices in this area. The testing reported in many academic papers is ad hoc and results are difficult to compare between approaches. Large datasets that are difficult to vet fully, and often-opaque system operation complicate the testing, evaluation, and debugging of AI systems.

While AI-specific testing and metrics are needed, existing methods and metrics from other disciplines (such as traditional software testing and safety engineering) could be adapted and applied in machine learning. Potential test and evaluation approaches could include:

- Verification of properties of an AI system, to ensure it behaves as intended (Katz, Barrett, Dill, Julian, & Kochenderfer, 2017)
- Metamorphic testing, a type of property-based testing used for scientific software (like AI) where exact input/output correspondence may be unknown, but certain properties about the relationship should hold (e.g. small perturbations to the input should not cause large swings in output)
- Targeted stress testing, to ensure resilience to possible or likely perturbations for a given deployment/use case
- Broad stress testing, to ensure scalability and resilience to a wide variety of general or common perturbations (such as data dropout and noise)
- Adaptive stress testing, searching for any possible failure modes

In addition to evaluating system performance for speed, resources, scalability, and accuracy, we should also subject AI systems to robustness testing (adversarial and otherwise), which implies having previously set robustness requirements. The development of this discipline could draw heavily upon examples from cybersecurity.

One challenge with a cybersecurity analog is evaluating the performance of a system unwittingly trained on poisoned data, resulting in misleading results during testing and evaluation. This is similar to the challenge of anomaly detection in cybersecurity (e.g. in network traffic), when the “norm” against which the system was developed unwittingly included adversarial activity.

Finally, dynamic systems that include online or active learning pose a particular challenge for testing and evaluation, since it is unknown how the system will evolve over time.

2.4 Deployment

Additional risks to consider during the deployment phase concern the system inputs and outputs in addition to the model itself.

Adversaries may manipulate the contents and timing of an input stream, or specific input samples, in order to degrade system performance or availability, or cause random misclassifications to erode user trust. Alternatively, the goal may be a targeted misclassification to achieve a specific outcome. Adversarial inputs can also be targeted at explainable AI, potentially achieving a targeted and misleading explanation, which could be another way to erode user trust.

A non-exhaustive list of deployed model and output concerns includes:

- Improper re-use of a model trained for another purpose
- Unwitting transfer of a poisoned or "Trojan" model
- Technical debt related to model provenance (i.e., open-source models)
- Generalizability vs specificity
- Model inversion (the ability to recover the training dataset)
- Model extraction (the ability to extract information and parameters necessary to reproduce the model)
- Membership inference (the ability to determine if a given sample was included in the training data)
- Proper tuning of hyperparameters and thresholds
- Confidence scores and explanations
- Opaqueness of model operation and outputs
- Mistakes, whether benign or adversarial
- Output feeding back into input (particularly in online learning systems)

Again, insider threat is a risk in many phases, including deployment where parameters, thresholds, and outputs could be manipulated, and human-machine teaming affected.

2.5 Operational Monitoring

Drawing upon lessons from cybersecurity, we must continue to evaluate risk even after deployment, actively monitoring an operational system in order to detect, respond to, and recover from failures, whether benign or adversarial in nature.

- **Detect**: Examples of detection activities include
    - Monitoring inputs to detect suspicious behavior (e.g. systematic query patterns designed to extract information, anomalies and outliers, or intentional manipulation)
    - Monitoring for data distribution shift (e.g., new cameras/sensors generating inputs)
    - Monitoring system to detect model or concept drift (e.g., identify model performance drop due to learned pattern changing)
    - Monitoring system performance (e.g., timing, resources) to detect degradation or unusual behavior

    - Human auditing of AI decisions, depending on particular application and consequences of mistakes (e.g. random audits, low-confidence decision audits, or even auditing all high-stakes decisions)

- **Respond**: A response plan should be developed prior to deployment so that if undesired system behavior or adversarial attack is detected, the plan can be put into action as soon as possible. A response might include taking the system offline, bringing a backup system online, reverting to alternative procedures, or continuing to operate but with additional caveats or restrictions. The appropriate response is highly dependent upon the application and the consequences of the undesired behavior. Incident reporting may be advised or required.
- **Recover**: Following an immediate response to a detected failure or breach, there is a need to eventually return to normal operation. The exact path to recovery will depend on the nature of the failure or attack, the application at hand, and the consequences of the errant behavior. If some type of poisoning was discovered, it may be necessary to retrain the model on cleansed data before being brought back online. If there was an adversarial evasion attack, recovery may involve performing additional adversarial training to make the model more robust. Again, a recovery plan should be developed prior to deployment to enable speedy execution when needed.

3 Metrics

Measuring risk quantitatively should be a key aspect of the RMF, so that goals can be clearly defined and progress assessed. No single, universal metric will suffice; rather, we require a suite of metrics across the entire AI system lifecycle.

Additionally, we will need to understand how various goals and metrics interact with each other, and which should be prioritized at the expense of others when they are in conflict. For instance, if improving a model's robustness decreases its fairness or privacy, how do we make that tradeoff? Quantitative metrics will be important to such decision making and cost/benefit analyses. As such, it is equally important to ensure that metrics and their implications can be presented to non-ML experts, which often include key decision makers, clearly and accessibly.

Some specific areas where metrics are applicable and needed include the following, some of which can be borrowed and adapted from prior work:

- Quantifying potential societal harm (e.g. economic)
- Quantifying cost of mistakes (e.g. in dollars, lives, opportunities), including false positives and false negatives
- Setting acceptable thresholds or levels of acceptable risk
- Quantifying costs of attacks/breaches (e.g. data exposure, system downtime)
- Quantifying severity of attack consequences

- Assessing fairness and bias (Mitchell, Potash, Barocas, D'Amour, & Lum, 2018) (Glymour & Herington, 2019)
- Assessing quality of development practices
    - Tracking adherence to best practice guidelines
    - Quality and sufficiency of documentation
    - Quality of data and dataset assembly
    - Well-considered and appropriately applied learning algorithm
- Assessing quality of test and evaluation procedures themselves
    - Appropriateness and rigor of methods and metrics used
    - Sufficient coverage of components and functionality
    - Adequacy of robustness assessment
- Assessing model performance and robustness
    - Model tuning and accuracy
    - Calibrating uncertainty or confidence estimates
    - Amount of information that can be inverted, extracted, or inferred
    - Efficacy of explanations provided with outputs
- Assessing operational monitoring
    - Efficacy of detection methods for identifying anomalous or suspicious behavior
    - Efficacy of audit procedures
    - Number and characteristics of failures/breaches
    - Efficacy and timeliness of responses to failures/breaches
    - Efficacy and timeliness of recoveries from failures/breaches
    - Effectiveness at enhancing human task or mission performance

4 Tools

Specialized tools are required to effectively design, develop, test, evaluate, and analyze AI systems. These tools may interact with the algorithms, models, or underlying data of an AI system to effectively evaluate the metrics and validate the defined parameters of the RMF. This comprehensive suite of tools should be interoperable, expandable, and flexible enough for the wide variety of available AI.

Specific tool categories and considerations include:

- A suite of tools should cover a broad spectrum of testing, monitoring, and metrics approaches, including those outlined in Sections 2.3 (Testing and Evaluation), 2.5 (Operational Monitoring), and 3 (Metrics).
- Similar to existing systems like static code analysis in software engineering, tools should be standardized and automated to provide analysis to a wide variety of AI systems with minimal ad hoc customization.
- The community should develop an enumeration of potential exploits and threats to AI systems. This federated system cataloging threats and exploits to AI systems is envisioned as the AI equivalent of the Common Vulnerability and Exposures (CVE),

Common Weakness Enumeration (CWE), and Common Attack Pattern Enumeration and Classification (CAPEC) (The MITRE Corporation, 2020). CVEs for AI systems may be particularly challenging, since vulnerabilities may be dependent on specific training data and training processes, which shape the final model. A related existing effort is MITRE ATLAS, which aims to produce an adversarial ML threat matrix similar to the ATT&CK matrix for cybersecurity (The MITRE Corporation, n.d.).

- Borrowing from the cyber ranges used to measure the risk of cyber systems, tools to evaluate AI in the presence of adversarial input and attacks should be created. AI testbeds will enable us to more accurately assess risk and compute established metrics on the proposed AI-based solution. To facilitate this type of evaluation, we need:
    - A common introspection API to enable automated testing and analysis of AI systems. This API should expose system and data status for the purposes of making systems compatible with test and evaluation as well as benchmarking tools.
    - A suite or toolbox of common adversarial attacks to be used for robustness testing and benchmarking, rather than individual, ad hoc approaches. Some existing efforts in this area include:
        - Armory (Two Six Labs, 2021) and the related Adversarial Robustness Toolbox (LF AI, 2021) from the DARPA GARD program (DARPA, n.d.)
        - AutoAttack, an ensemble of attacks to estimate adversarial robustness which often performs more reliable assessment of adversarial robustness techniques than researchers themselves (Croce & Hein, 2020)
        - CleverHans, a library to benchmark machine learning systems' vulnerability to adversarial examples (CleverHans Lab, 2021)
    - Benchmarking suites and metrics enable comparisons to baselines and between tools, enabling evaluations of multiple AI implementations. These benchmarking tools should have the capability to measure progress on a given risk-related task, and should be easy to integrate with and run against real-world systems and datasets to help organizations assess their own risks. A scoring system should be developed along with the metrics for use throughout the RMF.
- Tools and practices for reverse engineering AI systems are necessary for evaluating models due to proprietary protections or the “black box” nature of AI models. Such tools could assist red teams in performing AI system assessments, and may be useful for finding undeclared or hidden AI components in the supply chain.
- A vetted and centralized artifacts repository, such as the DoD’s Iron Bank, could provide access to hardened and verified open source and COTS tools, training data, and pre-trained models to facilitate safer development and deployment of AI solutions.

5 Human Computer Interaction

The combination of advanced and effective AI development, tightly coupled to mission success through human-machine teaming, presents an opportunity to create a disruptive advantage for many of the nation's and DoD's hardest problems. However, the close coupling of humans and systems has the potential to introduce an additional element of risk relative to human adoption and system acceptability, or from an increased attack surface due to uniquely human vulnerabilities.

To ensure effectiveness, additional investments in human-centered, AI-integrated risk management processes are required. Potential mitigations include defining and developing:

- Sustainable, repeatable AI capability development and deployment processes to ensure that we understand appropriate uses for AI (as well as identify cases where AI should not be applied), and to ensure that AI capabilities are a good fit for mission needs. The application of human-centered AI system development techniques should also ensure explainability, interpretability, appropriate uncertainty and error bounding, and usability of AI systems by both AI experts, domain experts, and novices to both. Ideally, these processes should be derived from the considerable existing body of knowledge – covering research methodologies and human-in-the-loop evaluation techniques -- from the user-centered design and human-computer interaction community. Lessons learned from operational deployments should inform the development of standards, best practices, and design guidelines for how human-machine teaming technologies are integrated into existing government platforms and tools.
- Experimentation, modeling/simulation, and instrumentation for assessing humans, systems, and environments can provide the means to conduct experimentation to predict how AI system and human-machine teams will perform under a variety of circumstances. These experiments and simulations will provide the means to predict the expected impact of AI at all levels of system maturity, understand the potential for emergent properties, and reduce the risk of unintended consequences.
- Quantitative and qualitative metrics and measures of effectiveness, performance, and risks for hybrid human-AI systems in a mission context are needed.
- Little attention has been paid to the security of AI and human-machine teams in red/blue or adversarial contexts. Integration of humans into the system can create new opportunities for system exploitation by the introduction of new kinds of attacks via adversarial design patterns, cyber deception techniques, cognitive or behavioral attacks or social engineering. Further research is needed to characterize the risk of hybrid human-AI systems.

6 Organization Roles

A variety of organizational types will be required to realize the RMF. While leading edge capabilities will largely come from industry and academia before they are leveraged by mission

users, such as the government, FFRDCs can help assess, mature, and manage the risk of those technologies prior to transition.

The DoD research enterprise is a powerful and unique set of organizational resources. If properly employed, it can make significant contributions to DoD and national AI strategies. As trusted USG advisers, government labs and FFRDCs can provide, for example, testbeds for evaluating and prototyping new AI computing architectures, advanced AI hardware prototypes, sensor and tool development for data collection and simulation, applied AI algorithm research and development activities, support for AI standards development, and very strong public (commercial and academic) engagement.

FFRDCs can also play key roles in helping to accelerate the adoption of AI and human-centered technologies to operations, providing a roadmap for emerging human-centered research and technology development, seeding and running academic research challenges, partnering with academia on mission-specific challenges (e.g., accelerators and incubators), bridging operations and research (including integration with and among commercial and private entities), and facilitating operational adoption of AI technologies. Being at the juncture between academia and the DoD, FFRDCs are also able to engage both communities in mediated discussions centered on the ethical and safe use of AI technologies.

7 Conclusion

MIT Lincoln Laboratory was pleased to have had the opportunity to contribute a response to the NIST RFI on an Artificial Intelligence Risk Management Framework. We hope you will find our input useful and welcome further interaction as you continue with this nationally important task.

For further discussions or questions on our submission, our point of contact is Kendra Kratkiewicz (kendra@ll.mit.edu)

8 References

CleverHans Lab. (2021). *CleverHans*. Retrieved from CleverHans GitHub: https://github.com/cleverhans-lab/cleverhans

Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter free-attacks. *International Conference on Machine Learning* (pp. 2206-2216). PMLR.

DARPA. (n.d.). *GARD: Guaranteeing AI Robustness to Deception*. Retrieved from GARD Projecthttps://www.gardproject.org/: https://www.gardproject.org/

European Commission. (2021). *Regulatory Framework Proposal on Artificial Intelligence*. Retrieved from European Commission, Digital Strategy: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Gebru, T., Morgenstern, J., Vecchione, B., Vaughn, J., Wallach, H., Daume III, H., & Crawford, K. (2018). *Datasheets for datasets.* arXiv preprint arXiv: 1803.09010.

Glymour, B., & Herington, J. (2019). Measuring the Biases that Matter. *Proceedings of the conference on fairness, accountability, and transparency*, (pp. 269-278).

Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017). Reluplex: An efficient SMT solver for verifying deep neural networks. *International Conference on Computer Aided Verification*, (pp. 97-117).

LF AI. (2021). *Adversarial Robustness Toolbox*. Retrieved from Adversarial Robustness Toolbox: https://adversarial-robustness-toolbox.org/

MItchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., . . . Gebru, T. (2019). Model cards for model reporting. *Proceedings of the conference on fairness, accountability, and transparency*, (pp. 220-229).

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2018). *Prediction-Based Decisions and Fairness:.* arXiv preprint arXiv:1811.08867.

NeurIPS. (2021). *NeurIPS 2021 Paper Checklist Guidelines*. Retrieved from NeurIPS 2021: https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist

Suresh, H., & Guttag, J. (2019). *A Framework for Understanding Sources of Harmthroughout the Machine Learning Life Cycle.* arXiv preprint arXiv:1901.10002.

The MITRE Corporation. (2020, 12 22). *CVE-CWE-CAPEC Relationships*. Retrieved from CVE: https://cve.mitre.org/cve_cwe_capec_relationships

The MITRE Corporation. (n.d.). *ATLAS*. Retrieved from MITRE ATLAS: https://atlas.mitre.org/matrix

Two Six Labs. (2021). *Armory*. Retrieved from Armory GitHub: https://github.com/twosixlabs/armory

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.