Response to NIST Request for Information
on the Artificial Intelligence Risk Management Framework

Raymond Sheh
Research Professor, Georgetown University, Washington DC
Guest Researcher, National Institute of Standards and Technology, Gaithersburg MD
Senior Lecturer (Adjunct), Curtin University, Western Australia
raymond.sheh@georgetown.edu

Karen Geappen
BSc (Comp Sci) (Hons), GDip Arts (Strat&Mgmt), FGIA
Independent Consultant
karen@geappen.net

15th September 2021

The deployment of any system carries with it an element of risk. In a large and increasing number of applications, this risk is well worth taking in exchange for improved performance, utility, economy, or simply the ability to do things that haven't been done before.

While Artificial Intelligence (AI) systems in general, and Machine Learning (ML) systems in particular, are becoming increasingly common across all sectors of the community, the tools to analyze the risks inherent in these systems, and determine the acceptable levels of risk for different applications, have not kept up. The NIST AI Risk Management Framework (RMF) Request for Information (RFI) document highlights some of these and the development of the RMF will go a long way to ensuring that risk is appropriately managed in the increasing use of AI.

We feel that across the industry as a whole, there are still significant gaps in what should be considered for a comprehensive view of the risks associated with AI systems, and a realistic assessment of the level of risk that any given application can tolerate. We highlight the following significant gaps that we have identified and that should be included in the RMF.

1. **Defining of AI Terms Relating to Risk:** As a topic that exists at the intersection of many disciplines, the semantics around attributes in AI that are important to risk management are poorly defined and inconsistently used. We feel that it is vital for the RMF to place a particular focus on defining standard terminology around such topics as performance measurement, traceability, interpretability, explainability, transparency, repairability, and so-on, in a specific and technically relevant manner. Without such definitions, their use within the rest of the RMF becomes open to unnecessary ambiguity. These terms should be able to be understood by both a technical and business process audience. The first author's previous work, such as [1,2], is an attempt at filling in some terminology gaps around explainability, particularly relating to risk, but the terminology in the RMF must go well beyond this.

2. **Expanded List of Stakeholders:** The RMF RFI as currently written focuses on the needs of AI designers, developers, users, and evaluators. However, there are many other stakeholders who dictate, influence, and/or bear the risks associated with the use of AI. These stakeholders are likewise involved in the design, development, use, evaluation and final decision making related to AI systems.

   We feel that it is critical that these stakeholders be included in the RMF, be it as consumers of the document, or mentioned in guidelines and case studies to bring awareness of them to those consuming the document. For example, a doctor may be a user of an AI based medical diagnosis assistant but their patient ultimately bears the consequence of the risk. Similarly, those involved in the development and implementation of governance, regulations, policies, procedures, requirements, procurement decisions, and insurance decisions, particularly where AI may be involved but are not AI focused, must also accurately understand AI risk despite not being a member of the aforementioned groups.

   Their informed acceptance, or rejection, of such risks can sometimes be the most significant driver in the overall adoption of AI systems for any given application. The development of a risk framework should explicitly include the language, guidelines, and other tools that support at least the training and education necessary to inform such stakeholders of the implications of AI risk in their work.

3. **Accountability and Responsibility:** Accountability is a vital part of trust and enforces risk management. This reflects as much on the AI system as it does on the regulatory infrastructure and cultural norms around the development, procurement, deployment, maintenance, and investigation of AI.

   The RMF should reflect the difference in accountability for different stakeholders and provide a common language, grounded in terminology that is meaningful to both technical and business audiences, for them to express and understand each others' responsibilities. For example, developers might be held accountable for accurate descriptions and disclosure of inductive, design, data, and other biases that would be difficult or impossible for downstream stakeholders to determine for themselves. Similarly, those making decisions on the use of AI might be considered responsible for assessing the disclosed information in their business risk context.

4. **Over-trust:** The RMF RFI focuses on the issue of insufficient trust in AI systems. Over-trust of systems is also a risk that we feel needs to be emphasised, and one that seems to have attracted less attention by the field as a whole. This is especially imperative when considering chains of suppliers to a final AI product.

5. **Communication of Risk Across Sectors:** Many systems in society span across groups with different risk profiles. The language that is developed and disseminated in the RMF

to properly communicate the risks of AI systems, and the risk tolerance of different applications, should also be meaningful for the communication of risk relating to AI systems across stakeholders in different sectors of society. This is particularly important when the entity making the risk acceptance decision is one, or several, steps (and sectors) removed from those who bear the consequences of that risk.

6. **Risks Across the Lifecycle:** Risk needs to be managed across the entire lifecycle of the AI system. The RMF RFI focuses on the risks managed in the development, evaluation, selection, procurement, and deployment phases of AI systems. We feel that it is lacking in addressing the risks associated with the ongoing maintenance of AI systems, particularly as their behaviour can change through their useful life. The RMF RFI also appears to neglect to cover risks associated with the end-of-life and replacement of AI systems. Furthermore, the RMF should explicitly include both planned end-of-life as well as end-of-life due to critical failure of the AI system, or any systems that the AI system depends on, and for which direct replacement may no longer be economically feasible.

   The behaviour of the system at end-of-life can be very different to designed and documented specifications due to the learning process that many such systems undergo. Appropriately replacing it, especially if caused by a critical failure rather than a planned upgrade, can carry substantial additional risk as compared to more traditional IT systems.

7. **Business/Process Continuity Risk:** A feature of AI is that it allows organisations to do what they could not otherwise do. However, their complexity, ability to learn over time, opacity, and specificity, can also make them difficult or impossible to replace or work around at short notice if they fail or are otherwise rendered inoperable, such as for regulatory reasons. This is particularly the case in critical ML systems where, by definition, critical knowledge is incorporated into the AI systems' ML model and often not present or documented anywhere else. A failure of this system thus also risks losing this valuable, and potentially irreplaceable, data and derived information such as learned models. Determining the risk associated with maintaining continuity when an AI system becomes unavailable is often poorly understood, yet it is crucial for an informed decision as to the risk of the use of AI systems. This is closely related to the aforementioned Risks Across the Lifecycle.

8. **Defensibility:** Security risk, cyber and otherwise, is mentioned in the RFM RFI, but only in the context of resilience. To manage AI risk, its level of defensibility must also be accurately characterised. Defensibility requires an understanding of the attack surface of the system and the ways in which attempted and successful attacks can be detected, defended against, logged, and audited after the fact. Such capabilities are often taken for granted in conventional IT systems and yet can be difficult or impossible in AI systems. It should therefore form part of the trade-off when designing or procuring such systems, and so that defenders of such systems understand what is required to perform

their part of overall risk management.

To date, the defensibility of AI systems is poorly understood, even by traditional “Blue Team” cyber security practitioners. The usual tools for anomaly detection or tracing the cause of behaviours often become impossible if the system learns and thus its behaviour changes through time, and especially if its decision making is done on the basis of a “black box” machine-learned model. The use of AI in such a manner can also reduce defensibility and increase the security risks in other systems. The less well defined behaviour of an AI system can mask incorrect behaviour, malicious or otherwise, in related systems. An AI Risk Management Framework should explicitly address the definition and quantification of such risks. In particular, “Explainability” itself is not necessarily sufficient for defensibility in this context. There are many different definitions for explainability, not all of which assist in defensibility of AI systems.

9. **Robustness and Resilience:** AI systems in general, and ML systems in particular, make decisions based on complicated, difficult to understand, and often opaque and time varying manners. After all, if this were not necessary, the system might not need to incorporate AI and could be developed more traditionally. Guarantees of robustness and definitions of valid operating parameters can be difficult or impossible to obtain. Determining the robustness and resilience of such systems, particularly to rare or unexpected events, be they malicious, accidental, or happenstance, is a vital component of managing risk associated with AI systems. We feel that this deserves greater focus in the RMF.

10. **Accurate Performance Metrics and Measurement:** Beyond failure, change and degradation in performance plays an important part of risk management. In many situations, a degradation in performance in one system increases the risk in a downstream system. Defining, characterising, and measuring the performance of AI systems, in both normal and unusual operating situations, both during procurement as well as throughout the lifecycle of the system, is necessary to manage risks associated with degraded performance. Current understanding of such performance measures, particularly outside of isolated testing of individual components, is poorly understood. It is strongly recommended that the RMF provide guidelines for the initial and continued performance measurement of AI systems as part of whole, and interconnected, systems.

11. **Consumer-Industry Agnostic:** The RMF RFI acknowledges that it should be platform and technology agnostic. We would recommend adding Consumer and Industry Sector agnostic to the list. This may be a significant challenge, particularly when combined with the broader stakeholder engagement recommendation earlier in this document, and may warrant splitting parts of the RMF into multiple sections for different stakeholders. Furthermore, in addition to principle level statements, we suggest that intended outcomes of applying these principles are also included. This will provide readers with context to better understand how they apply to their own industry.

12. **Non-government regulatory requirements:** The RMF RFI focuses on regulatory and reporting requirements from a government perspective. In many cases, the most effective regulatory and reporting requirements come from within industry, such as medical, insurance, or finance. Care should be taken to ensure that the RMF is similarly useful within such non-governmental bodies that also generate regulations and guidelines.

We strongly feel that effort spent ensuring that the RMF addresses these points will be well spent in increasing the relevance and impact of the RMF in assisting all stakeholders to properly understand the risks that they are creating, signing up for, and exposing others to. We would be delighted to discuss this topic further with those involved in developing the RMF.

[1] Sheh, R. and Monteath, I., 2017. Introspectively assessing failures through explainable artificial intelligence. In *IROS Workshop on Introspective Methods for Reliable Autonomy* pp. 40-47.

[2] Sheh, R. and Monteath, I., 2018. Defining explainable ai for requirements analysis. *KI-Künstliche Intelligenz*, *32*(4), pp.261-266.