Subject: DM feedback on the NIST AI RMF second draft

Dear NIST team developing the AI Risk Management Framework,

Thank you for the opportunity to provide comments on the second draft of the AI Risk Management Framework and Playbook. DeepMind shared a response to the RFI in September 2021 and feedback to the concept paper as well as the initial draft earlier this year. We appreciate that additional detail will be provided in the next iterations of the draft RMF and are sharing below some high-level feedback.

Introduction

As with previous drafts, we support the development of flexible and voluntary AI standards. Considering the pace of AI development and the fact that, in many areas, the field is not yet mature enough to design appropriate regulatory requirements, the development of guidance to improve companies' and individuals' understanding and management of risk is a crucial first step. These norms could then, if needed, mature into regulatory requirements at the appropriate time. The opportunity for many stakeholders to input on the various drafts in a transparent and clearly timed manner is also essential and something we have valued in the way NIST approached the development of the AI RMF.

AI Risks & Trustworthy AI

We support the changes made to the trustworthy AI section and the new structure around the seven elements of trustworthy AI - this new classification is better aligned with the OECD's approach of capturing the main characteristics of AI systems. This section however seems to suggest that AI systems are either trustworthy or not, depending on the seven characteristics. It might be more accurate to present trustworthiness as a spectrum that considers these characteristics within a particular context in which an AI application is being used. Ultimately, trustworthiness will also be determined by the users of the system.

As per our first NIST submission, we also want to reiterate the need to incentivise deliberation over complex longer-term risks from advancing AI capabilities. We support the inclusion of ongoing monitoring and information-sharing, given that new AI-related risks and mitigation strategies are discovered and studied over time as these systems evolve. Equally we think that guidance on tools, benchmarks and audit mechanisms may help different actors better understand the range of options available to address different harms.

The new emphasis on test, evaluation, validation and verification (TEVV) compared to previous drafts of the RMF is a positive development and we agree that socio-technical considerations throughout the AI lifecycle are important. However, we believe the preceding design stage deserves special attention as well, as choices made at the initial stage of problem formulation and outcome definition will matter greatly. We'd recommend reviewing language to be more balanced in the way TEVV is mentioned.

More concretely, on some of the changes and new additions in this section and the seven characteristics:

-
-
- Overall, within the characteristics, it seems that whether a system is performant - be it accurate or high-performing
- - is not sufficiently emphasised, although we noted it was under the 'valid and reliable' characteristic. Data governance also deserves treatment as a more fundamental concept that runs across all the different characteristics. Thinking about the governance
- mechanisms as a key component of the RMF should also help give it a more robust structure.
-
-
-
- **'Valid and Reliable' -**
- The mention of human-AI teaming aspects in this section is crucial. Accuracy should be determined within
- the more holistic environment of actual use, for instance, if a medical prediction device is developed to be used by nurses, then the accuracy isn't just of the model itself, but of the model coupled with the nurses' interpretation in the realm of real use.
-
-
-
- **'Safe' -**
- The mention of a statement from ISO/IEC TS 5723:2022 that AI systems 'should not, under defined conditions,
- cause physical or psychological harm or lead to a state in which human life, health, property, or the environment is endangered' doesn't seem to consider risk trade-offs rather than the absence of risk. The definition of a safe system will depend on how we
- understand this risk in relation to the benefit of the system overall - clarifying this point in this section would be helpful
-
-
-
- **'Fair - and bias is managed'**
- -
- NIST rightly notes that fairness, bias and discrimination are complex concepts that are defined and used differently,

- across different contexts, for example across consumer, financial, and data privacy law, as well as statistics, computer science, and cognitive science. For this reason, we support NIST's attempt to develop a standard for 'identifying and managing bias in
- AI' that demarcates this risk into challenges relating to systemic bias, computational bias and human biases. It would be helpful to further illustrate some of the sub-categories of risk within these three categories, supported by examples, as some are not
- immediately clear: for instance, what does NIST mean by 'computational bias' within datasets? Similarly, NIST could highlight that the need for ongoing monitoring of AI systems is particularly true for fairness risks, as some may only be visible over long
- timescales - for instance a loan-approval system that disproportionately favors certain groups, but in ways that only become evident over longer time horizons. In other instances, efforts to mitigate certain fairness risks can lead to negative longer-term
- side effects - DeepMind research
- has highlighted how efforts to measure and mitigate language model toxicity, at the language mode training stage, could potentially result in downstream harms, such as a reduction in text about, and dialects of, marginalized groups.
-
-
-
- **'Secure and Resilient'**
- -
- This section seems overly high level and there is a need to expand on some of the terminology: for instance, what does confidentiality, integrity and availability mean in terms of an AI system? More NIST publications or frameworks that capture security requirements
- could also be referenced in this section such as
- NIST
- CSF
- and how it fits into the AI risk management framework
- for addressing security risks.
-
-
-
- **'Transparent and Accountable' -**
- Documentation best practices with respect to AI systems continue to evolve, particularly in the AI research world. Efforts to document the characteristics of
- AI
- models have expanded to

- datasets
- and reinforcement
- learning agents, as well as to documenting the broader technical
- and sociotechnical systems that AI models are incorporated into. NIST could incorporate guidance on how and when to use such resources, as well as broader data management practices, perhaps referencing NIST's
- Research
- Data Framework (RDaF), to help organizations better gauge expectations
- around transparency and to prepare accordingly. As
- documentation practice develops and becomes increasingly common in AI research, NIST could explore and define best practices over time, and provide guidelines as to what, when and how to document. Building on the notion of responsibility being shared among
- 'all AI actors', considering early onwards how documentation can contribute to understanding how that responsibility is and should be distributed would also be helpful.
-
-
-
- **'Explainable and interpretable' -** The
- terms 'explainability' and 'interpretability' are used inconsistently across the AI community, including in ways not captured by the current NIST definitions. For example, some conceptualisations of explainability capture not just 'mechanistic' attempts
- to identify causal pathways or features underlying a model's predictions,
- but also the critical work, informed
- by social science, to ensure that the subsequent explanation
- provided to a user is accessible, and useful, for their specific goals. Other practitioners stress the need for post-hoc 'analysis' and 'probing' of AI systems and applications', and their outputs. Given these challenges, it would be helpful for NIST to highlight
- how their definitions differ to others in the field. NIST could also provide practitioners with practical examples and illustrations of the different types of explainability and interpretability challenges they may face, across different contexts and audiences,
- as well as guidance on the trade-offs and relationships that may exist - for example, between interpretability, accuracy, and fairness.
-
-
-
- **'Privacy-enhanced'**

- - The mention that 'processing of data could create privacy-related problems' might
- be too narrow a view for how privacy risks might emerge: since explicit combinations of data sources can create privacy problems that the two sources in isolation didn't have, interactions between AI systems represent implicit combinations of data sources
- that could generate privacy risks that each of the systems in isolation don't have. It might therefore not be enough for assessors to look at data processing activities; they might also have to look at the use and interactions between AI systems. We suggest
- favoring outcome-based approaches, as it is unlikely that a 'one size fits all' approach to privacy would adequately address the range of potential privacy risks.
-
-
-
- **'Human factors'**
- - The concept of 'human in the loop' system, mentioned in the human factors section,
- should be defined and discussed in more detail as this is currently the main mechanism for managing risk in deploying safety critical systems, a mechanism that is also limited by the extent the human can understand what the system does. The definition should
- clearly make a distinction between systems that can autonomously make a deferral decision to a human expert and systems that are merely used by a human decision maker as an additional opinion. In addition, a number of AI systems are unlikely to require much
- human oversight, such as models used to improve video compression.
-

AI RMF profiles

Considering the current focus on general purpose systems, we'd encourage the inclusion of a profile on those systems, which could facilitate cross-border interoperability as the EU is currently looking at regulatory requirements. A profile on general purpose systems would also bring to light some of the challenges we'd encounter when governing these systems - in this case, and considering the large number of actors involved in the GPAIS value chain, from which perspective should a profile be drafted? It's unclear whether a use case profile would effectively provide ecosystem-wide allocation of responsibilities. In addition, the experimental and iterative nature of AI development resembles more R&D than traditional product development; the risk management system needs to strike the right balance between structure/certainty and flexibility/modularity.

More broadly, on this point around AI actors, and despite the incorporation of the OECD definition of AI actors in the second draft, it is unclear what the various roles of organizations will be with respect to the design, development and distribution and use of an AI system within the AI RMF. These distinctions will be however critical in informing a risk management program. For instance, the mention of 'joint responsibility of all AI actors' could be misrepresented. The fact that the draft RMF does not distinguish between AI research and commercialisation is another area where a profile/use case would be appropriate.

RMF playbook

We've noted the efforts put in developing the playbook and in making it an interactive and accessible platform - what the playbook shows are some of the trade-offs that might need to be made between sharing substantial information as guidance and the possibility that too much information could become overwhelming to the RMF audience, or shift the perception of the framework to that of a box-checking compliance exercise, rather than a resource to enable deeper deliberation on the more complex risks facing an organization One recommendation could be to earmark more clearly which section is likely to be more relevant to which stakeholders, so that individuals can quickly grasp what part of the playbook they should be focusing on.

While we welcome the enhanced level of detail in the Playbook, we believe it's important to provide sufficient examples to illustrate what adequate implementation looks like. In particular, further guidance translating socio-regulatory requirements into technical implementation would be beneficial.

To make the playbook more actionable, NIST could look to collate and share state of the art AI risk evaluation and mitigation tools, across the risks covered by the RMF. This could build upon existing repositories, such the Partnership on AI's work to aggregate and compare Explainable AI tools, and help users address the 'choice overload' posed by the recent proliferation of such tools.

NIST could work with partners - in particular under-represented groups - to develop a representative set of case studies, showing how organizations plan to use the RMF, or how they currently do AI risk management. This would enable RMF users to build up a clear picture of what the RMF categories and subcategories might look like in practice. The case studies would also explain how risks and responsibilities will differ depending on where an organization sits in the AI supply chain - for example, an organization developing a new labeled dataset, or running AI experiments with human feedback, will want to closely consider the risks posed by data enrichment activities. Case studies could also demonstrate how RMF use will differ across sectors, but also for different actors within a single AI supply chain, from underlying dataset, model and compute developers, through to application deployers, and affected users. In doing so, they could highlight which actor is best-placed to act at different stages, but also which risks are dependent on multiple actors and their interactions, and illustrate different approaches to shared challenges. In addition, the case studies could highlight how decisions and trade-offs affect the resulting benefits and harms of an AI application, and how they are distributed. For example, our recent case study documented how DeepMind worked with third parties to identify and mitigate potential unintended risks posed by releasing our AlphaFold protein structure prediction model, as well as certain decisions we took to try and ensure that the benefits would be shared more equitably.

Conclusion

Considering the extensive work undertaken by NIST to develop the AI RMF, and its potential to act as a long-term guiding resource, we suggest that NIST also considers ways to share the work with relevant international partners. For example, the UK is currently outlining their approach to AI governance and regulation and could potentially benefit from similar resources.

We look forward to continuing to input on the NIST AI RMF, and we would welcome the opportunity to have a call with you to discuss our comments further.

Kind regards,
Alexandra

--

Alexandra Belias

International Public Policy Manager