Federation of American Scientists

1112 16th Street NW, Suite 600, Washington, D.C. 20036
P: 202-546-3300 • F: 202-675-1010
fas.org • fas@fas.org

Public Comment by the Federation of American Scientists (FAS) Submitted to the National Institute of Standards and Technology Regarding the Second Draft of the AI Risk Management Framework and the First Draft of the NIST AI RMF Playbook

The Federation of American Scientists is a 501(c)(3) nonprofit organization that combines research, analysis, and policy to provide critical, data-driven information to government officials. Founded in November 1945 to meet national security challenges with evidence-based, scientifically-driven, and nonpartisan policy analysis and research, FAS is devoted to the belief that scientists, engineers, and other technically trained people have the ethical obligation to ensure that the technological fruits of their intellect and labor are applied to the benefit of humankind. We greatly appreciate NIST's efforts to develop this flexible and voluntary approach for managing risks from AI. We think this framework will offer strong guidance to the American AI enterprise, and that it will serve as an important step forward in setting standards for AI actors to responsibly measure and manage risks throughout the AI system lifecycle.

We commend the focus NIST has placed on a useful for framing the Framework especially right at the beginning, and want to highlight a few aspects in particular that we hope will be present in future versions:

a) An acknowledgement of AI's potential to bring a "wide range of innovations with the potential to benefit nearly all aspects of our society and economy – from commerce and healthcare to transportation and cybersecurity."[1]
b) An appreciation of how greatly risks from AI can vary along a variety of dimensions e.g. "long- or short-term, high- or low-probability, systemic or localized, and high- or low-impact."[2]
c) An understanding of the inadequacy of current methods for measuring, quantifying, and managing risks from AI, due to the difficulties that are present in determining and predicting the behavior of many types of AI systems.

[1] NIST AI RMF Draft 2, p. 1
[2] *Ibid.*

Federation of American Scientists

1112 16th Street NW, Suite 400, Washington, D.C. 20036
P: 202-546-3300 • F: 202-675-1010
fas.org • fas@fas.org

Understanding Risk, Impacts, and Harms

The RMF lays out "examples of potential harms related to AI systems" and discusses harms to individuals in terms of harms to their civil liberties, rights, or physical safety (Figure 3). However, this does not mention the reputational harm that can be caused by "model hallucinations,"[3] a phenomenon in which AI systems generate untrue statements given certain prompts or may make inaccurate predictions in an attempt to fit the model on an input. If a chatbot built on a large language model generates untrue statements about an individual, this could be harmful to that person's reputation while at the same time not posing direct threats to their physical safety. To its credit, the RMF does recognize this as a potential harm to an organization ("Harm to an organization's reputation"). This harm should also be considered at the individual level when implementing risk management practices.

We suggest explicitly calling attention to this problem. This could be done by a modification to Figure 3—by adding reputation to the list of potential harms to individuals, so that it reads: ("Harm to a person's civil liberties, rights, physical safety, or reputation.")

Challenges for AI Risk Management

In this section, the framework lays out many challenges faced by AI actors that are "managing risks in pursuit of AI trustworthiness."[4] One such difficulty is the inscrutability of AI systems, which the Framework mentions "can complicate the measure of risk" due to a "lack of explainability or interpretability."

However, this section misses key difficulties that arise when managing AI risks, and we recommend that it describe some additional challenges for AI actors to take into account. One challenge that this section should include is the inadequacy of many interpretability and explainability toolkits. Most interpretability and explainability toolkits are unreliable and can contribute to a false sense of risk management and transparency. Tools like SHAP, LIME, gradient based saliency methods, and others, while popular interpretability tools, may often be fundamentally wrong tools to employ for certain models and provide no reason to believe that their results point to any reasonably accurate interpretation of a model's functionality.[5]

---

[3] https://www.unite.ai/preventing-hallucination-in-gpt-3-and-other-complex-language-models/
[4] NIST AI RMF Draft 2, p. 8
[5] Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020, November). Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning* (pp. 5491-5500). PMLR.

Federation of American Scientists

1112 16th Street NW, Suite 400, Washington, D.C. 20036
P: 202-546-3300 • F: 202-675-1010
fas.org • fas@fas.org

AI Risks and Trustworthiness

Accuracy

One of the key tenets to trustworthiness is accuracy. But it not only matters that a model is accurate on a held out set of data (i.e. the "test" set), it should also be critical to ensure that the model's worst group accuracy is above a certain acceptable threshold. We recommend the RMF disentangle the concept of accuracy as a singular number on a dataset or data distribution and go even further than its current section "Fair–and Bias Is Managed" to explicitly recommend that researchers pay attention to the demographic attributes of individuals who are asked to evaluate or beta-test their systems before such systems are deployed in the world. The Food and Drug Administration recently issued a draft guidance, "Diversity Plans to Improve Enrollment of Participants from Underrepresented Racial and Ethnic Subgroups in Clinical Trials"[6] which recommends that sponsors of medical products develop and submit a Race and Ethnicity Diversity Plan to the agency early in clinical development, based on a framework outlined in the guidance. Practitioners designing, implementing, and evaluating AI models should be encouraged to design similar plans that would help put a focus up front on the diversity of the cohort contributing to the training data as well as the diversity of cohort evaluating and auditing the resulting models.

Interpretability and Explainability

As AI increasingly penetrates critical areas such as medicine, the criminal justice system, and financial markets, the inability of humans to understand these models is becoming a major problem. While there is growing recognition of the need for model interpretability, there is still no agreement on what interpretability actually means, or on how to achieve it.[7] In fact, in academic literature, the concepts of interpretability and explainability appear simultaneously important and slippery.[8] It's important to be careful when interpreting model results. For instance, a saliency map is a local explanation.[9] If you move just one pixel on an input image, you could get a very different saliency map. This contrasts with linear models, which model

---

[6] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/diversity-plans-improve-enrollment-participants-underrepresented-racial-and-ethnic-populations

[7] Supra note 3.

[8] Lipton, Z. C., & Steinhardt, J. (2018). Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*.

[9] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, *31*.

1112 16th Street NW, Suite 400, Washington, D.C. 20036
P: 202-546-3300 • F: 202-675-1010
fas.org • fas@fas.org

global relationships between inputs and outputs. In a study on local explanations,[10] the authors explain the decisions of any model in a local region near a particular point by learning a separate sparse linear model to explain the decisions of the first. Even though this method can provide explanations for non-differentiable models, it's more often used when the model subject to interpretation is in fact differentiable. In this case, it's not clear what benefits this approach offers over a plain gradient. In this paper, the explanation is offered in terms of a set of superpixels. Whether or not this is more informative than a plain gradient may depend strongly on how the superpixels are chosen. Moreover, without a rigorously defined objective, it's hard to say which hyper-parameters are correct.

In the absence of clear definitions and objectives, a focus on these concepts of explainability and interpretability may serve little purpose. The AI RMF has the opportunity and a responsibility to get ahead of the curve. There is a growing body of literature proposing purportedly interpretable algorithms, but it is clear that interpretability is not a monolithic concept. It is made up of several distinct ideas that must be disentangled before any progress can be made. The objectives and methods put forth in the literature investigating interpretability are diverse, suggesting that interpretability is not a monolithic concept but several distinct ideas that must be disentangled before any progress can be made. We recommend that the RMF be edited to reflect this.

Safety

We appreciate the effort that has gone into this section, and especially commend NIST's encouragement of maintaining "the ability to shut down or modify systems that deviate from intended or expected functionality."[11]

We think this section could be improved by explicitly drawing attention to the concept of AI Value Alignment, i.e. the task of getting AI systems to reliably pursue the goals which humans intend. This is a concept that many researchers find useful in shaping their efforts to create AI systems which are in line with the mission of minimizing risks (and maximizing positive benefits) from AI systems.[12] It therefore would be useful to draw explicit attention to it in the framework.

---

[10] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
[11] NIST AI RMF Draft 2, p. 13
[12] https://openai.com/alignment/, https://www.deepmind.com/publications/artificial-intelligence-values-and-alignment

Federation of American Scientists

1112 16th Street NW, Suite 400, Washington, D.C. 20036
P: 202-546-3300 • F: 202-675-1010
fas.org • fas@fas.org

We recommend that the name of this section be changed to "Safety & Value Alignment," and a short description of "Value Alignment" could be provided. Resources which may be useful are referenced below.[13]

Performance Guarantees and Contracts

One aspect missing from the RMF is a discussion of how the burden to educate the end users on when to expect an AI-enabled technology to work (or fail) should fall on the developers of the technology. Absent education and clearly worded, easy to understand contracts about performance guarantees and model reliability, there is a risk that the burden for risk management would be completely transferred to the end user. For instance, if a developer claims their model is certifiably robust to certain attacks, the onus of communicating that to the end user and ensuring that a lay-user without knowledge of advanced statistics and machine learning is able to understand what that entails.

Additional Comments

1. The RMF must explicitly call out the use of AI for law enforcement and defense purposes in that there must be a significantly high bar for risk management in those contexts versus non-defense, non-law enforcement use cases. While the Department of Defense has issued Responsible AI Guidelines[14] and is working to implement them, many local, state and federal law enforcement agencies have not yet established clear guidance for their use of AI-enabled technologies. In fact, as many of these law enforcement agencies have been partnering with third-party vendors for facial recognition and other AI-enabled surveillance tools, the guidelines and regulations come from a patchwork of local ordinances or state laws. NIST can help lawmakers in these states, and implementers at agencies look to a common standard as they seek to engage in responsible use of AI-enabled technologies. NIST should further emphasize that these agencies have a significantly greater responsibility while employing such technologies and using the AI RMF around these technologies.
2. While the AI RMF takes significant steps in highlighting how risk management is integral to every part of the AI lifecycle, we recommend that the RMF incorporate valuable insights offered by researchers Thomas Krendl Gilbert, Sarah Dean, Tom Zick, and Nathan Lambert in a recent white paper, in which they discuss new forms of risk presented by Reinforcement Learning (RL) models that exacerbate the harms already generated by standard machine learning tools. They correspondingly present

[13] https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6, https://brianchristian.org/the-alignment-problem/
[14] https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf

Federation of American Scientists

1112 16th Street NW, Suite 400, Washington, D.C. 20036
P: 202-546-3300 • F: 202-675-1010
fas.org • fas@fas.org

a new typology of risks arising from RL design choices, falling under four categories: *scoping the horizon*, *defining rewards*, *pruning information*, and *training multiple agents*.[15]

3. While we find it very encouraging that "the Framework and supporting resources will be updated, expanded, and improved based on evolving technology, the standards landscape around the globe, and stakeholder feedback,"[16] we recommend building in an additional commitment to update this framework regularly (by considering feedback and updating the framework accordingly at least once every 12 months, similar to the current plan to update the Playbook every 6 months), due to the inherent difficulty in ensuring that guidance on "best practices" keeps apace with rapid progress in the capabilities of AI systems.

Comments on the Initial Draft of the NIST AI RMF Playbook

General Feedback

Here, we first provide feedback on the specific questions that NIST requested input on here. Each heading corresponds to one of NIST's specific requests.

The Playbook's relative usefulness as a complementary resource to the AI RMF

It is likely that this playbook will be a useful resource to complement the AI RMF. It is more interactive and easy to navigate than the Framework, and it provides more "actions" for organizations to consider taking, whereas the original RMF relies more on giving examples of potential outcomes. Thus, for organizations aiming to improve their risk management practices, the Playbook will serve as a very useful "go-to" resource.

[15] Gilbert, T. K., Dean, S., Zick, T., & Lambert, N. (2022). Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems. *arXiv preprint arXiv:2202.05716*.
[16] NIST AI RMF Draft 2, p. 3

Federation of American Scientists

1112 16th Street NW, Suite 400, Washington, D.C. 20036
P: 202-546-3300 • F: 202-675-1010
fas.org • fas@fas.org

Whether the guidance is actionable, especially as related to organization size.

The guidance is generally actionable, but could be improved. There are some pieces of guidance in the "Actions" section of the Playbook that are vague and therefore may be difficult for organizations to implement effectively in practice if they are not made more specific. Below we provide a few examples of actions from the Playbook that are too vague to be easily actionable.

- "Assess system benefits and negative impacts in relation to trustworthy characteristics."
- "Collaboratively establish policies that address third-party AI systems and data."
- "Establish policies that facilitate inclusivity and the integration of new insights into existing practice."

It may be worth making many of these "Actions" more specific. Another way to improve the guidance would be to provide case studies of how different organizations have approached risk management when deploying AI systems in a variety of contexts.

Suggested presentation alternatives for the forthcoming first version of the Playbook, to ensure effectiveness and ease of use.

We think that one of the most significant drawbacks of the Playbook is that it is currently not very easy to navigate, especially relative to lots of other online resources. This may decrease its usefulness as a voluntary resource. Here are a few potential ideas for improvements:

- **Make the website more interactive and tailored to individual users' needs.**
    - One potential way to do this could be to present website users with a set of questions about their situation. (E.g., In what context are they deploying an AI system? How large is their organization? How many parameters are in their model/how much compute are they using?) Then, direct users to a tailored version of the Playbook that provides risk-management practices and considerations that are most relevant to their situation, while communicating that this is an example to help them guide through the process and not a definitive answer on what they should or should not do. A recognition that the needs for one user may very well be different and not fully captured is critical and should be communicated up front. This could help make the resource more easily usable for a wide variety of AI actors.
- **Make greater use of videos**

Federation of American Scientists
1112 16th Street NW, Suite 400, Washington, D.C. 20036
P: 202-546-3300 • F: 202-675-1010
fas.org • fas@fas.org

    - Key concepts and actions that are useful in AI risk management could be explained in videos. NIST has successfully used videos[17] in the past to communicate some key ideas, and could do so for some key concepts in the AI RMF Playbook.
- **Make greater use of images and diagrams**
    - Another way the website could be improved is through greater use of images and diagrams.
- **Offer multiple website mock-ups for feedback.**
    - If wishing to get greater feedback on usability of the Playbook, it may be worthwhile to provide a few mock-ups of the website for feedback, then allowing those in the AI community to vote on their preferred design. This may end up producing a better product than incremental suggestions for improvement on one prototype.
- **Allow for the whole playbook to be viewed on "one page."**
    - While using this playbook, authors found it difficult to navigate to a piece of guidance they had previously read because it was hard to remember where specific pieces of guidance were embedded. This would have been easier if they could view the entire playbook in one page, allowing for easier searching of key terms.
    - This would also provide the benefit of allowing the entirety of the playbook to be printed out. For some users, allowing a printable version of the playbook will make it easier to reference and use. Currently, one cannot easily print the entirety of the Playbook.

Section-by-section feedback

Here, we provide feedback on specific sections of the playbook's content, for the published "GOVERN" and "MAP" functions.

GOVERN

In this section, we believe it is useful to mention the importance of instituting good information security practices[18], especially for high-risk AI systems. If higher-risk systems can be stolen by outside actors, this could prove catastrophic. It would further be useful to

[17] https://www.youtube.com/c/NIST
[18] https://csrc.nist.gov/glossary/term/infosec

Federation of American Scientists

1112 16th Street NW, Suite 400, Washington, D.C. 20036
P: 202-546-3300 • F: 202-675-1010
fas.org • fas@fas.org

provide specific reference to the NIST Cybersecurity Framework,[19] with organizations progressing up from Tier 1 to Tier 4 of this framework according to the risk factor associated with their AI systems. Some information on the importance of cybersecurity for AI systems can be found in "How to improve cybersecurity for artificial intelligence" by Josephine Wolff.[20]

(4.1) *"Establish policies that incentivize safety-first mindset and general critical thinking and review at an organizational and procedural level."*

> **Comment:** We suggest that NIST provide further examples of what it could look like to establish and incentivize a safety-first mindset and general critical thinking, especially if there are examples of such training programs that NIST could reference, either here or in the Resource Center.

MAP

We generally appreciate the guidance offered by this section. We do think this section could be improved in some ways as we lay below.

(5.2) *"Likelihood and magnitude of each identified impact based on expected use, past uses of AI systems in similar contexts, public incident reports, stakeholder feedback, or other data are identified and documented."*

> **Comment:** While it is reasonable to rely on track records to assess the risk factor level of most types of AI systems, this method will tend to underestimate the expected impact of risks that arise from emergent behavior of particularly advanced models, for which it may be difficult to find relevant track records.[21] We think it would be useful for NIST to provide an account of how to manage risks of systems for which there are no (or very limited) useful precedents that allow for accurate risk assessment.
> **Suggestion:** We suggest adding a new sentence to suggest something along the lines of: "If there is limited relevant data to assess the likelihood of different possible impacts, then actors should aim to provide an 'upper bound' on possible

[19] Barrett, M. P. (2018). Framework for improving critical infrastructure cybersecurity version 1.1.
[20] https://www.brookings.edu/research/how-to-improve-cybersecurity-for-artificial-intelligence/
[21] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. and Chi, E.H. (2022) Emergent Abilities of Large Language Models. *arXiv.* pp.arXiv-2206.

Federation of American Scientists

1112 16th Street NW, Suite 400, Washington, D.C. 20036
P: 202-546-3300 • F: 202-675-1010
fas.org • fas@fas.org

negative impacts, for example by limiting the scale of an AI system's influence until a low enough level of risk can be guaranteed."

"Likelihood estimates may then be assessed and judged for go/no-go decisions about deploying an AI system. If an organization decides to proceed with deploying the system, the likelihood estimate can be used to assign oversight resources appropriate for the risk level."

Comment: We suggest NIST include a method of assessing such "go/no-go" decisions, and for assessing whether catastrophic risks may be present. As mentioned above, arriving at likelihood estimates for such risks and for systems for which there is limited precedence is difficult, and is a task that it could be useful for NIST to provide guidance on.

Suggestion: Provide guidance on "go/no-go" decisions in instances where catastrophic risks may be present. One helpful resource may be "Actionable Guidance for High-Consequence AI Risk Management."[22]

(5.3) *"Make a go/no-go determination based on magnitude, and likelihood of impact. Do not deploy (no-go) or decommission the system if estimated risk surpasses organizational tolerances or thresholds. If a decision is made to proceed with deployment, assign the system to an appropriate risk tolerance and align oversight resources with the assessed risk."*

Comment: In this section, we believe it would be useful to emphasize the importance of assessing likelihood levels of risks that may emerge, but for which there is not yet sufficient track record to make a reasonable judgment. (Similar to the point made above for Section 5.2)

Thank you for your attention to these important matters.

Sincerely,

Joshua Schoop
Principal Director
Technology and Innovation
Federation of American Scientists

[22] Barrett, A. M., Hendrycks, D., Newman, J., & Nonnecke, B. (2022). Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. *arXiv preprint arXiv:2206.08966*.