Response to the request for Information: Artificial Intelligence Risk Management Framework

To:
Mark Przybocki
National Institute of Standards and Technology
100 Bureau Drive, Stop 200
Gaithersburg, MD 20899

RE: Request for Information: Artificial Intelligence Risk Management Framework

Dear Mr. Przybocki,

The Future of Life Institute (FLI) is grateful for the opportunity to contribute comments on the initial ideation of the National Institute of Standards and Technology's (NIST's) Artificial Intelligence Risk Management Framework (AI RMF). Having contributed input on numerous past documents from NIST on AI, we believe that the effort to create an initial version of the legislatively mandated AI RMF will be a critical capstone to the valuable work NIST has already engaged in on AI. We therefore look forward to working closely with NIST as the RMF matures in the coming year.

In order to make our comments easier to read, we are submitting a document version of our comments that are structured with the core categories of NIST's response template (requested topic, response/comment, and suggested change) for each of our ten comments, numbered by roman numerals. We are also submitting the same content inside the NIST's requested response template spreadsheet.

For any questions, please contact Jared Brown, Senior Adviser for Government Affairs, at jared@futureoflife.org.

About the Future of Life Institute (FLI)

FLI is one of the world's leading voices on the governance of AI. The Institute is an independent nonprofit that works on maximizing the benefits of technology and reducing its associated risks. FLI created one of the earliest and most influential set of AI soft law governance initiatives - the Asilomar AI principles - and maintains a large network among the world's top AI researchers. The Institute, alongside the governments of France and Finland, is also the civil society champion of the recommendations on AI in the UN Secretary General's Digital Cooperation Roadmap. For additional information about FLI's policy work related to AI, please see https://futureoflife.org/policy-work/.

I. COMMENT

Requested topic:

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

Response/Comment (Include rationale):

One of the greatest challenges in managing AI risks is the evaluation of effects from AI systems that are characterized as being long-term, low probability, systemic, and high impact. Proactively tackling scenarios that can represent catastrophic or even existential risks to society within NIST's AI Risk Management Framework (RMF) should incorporate two lenses: an emphasis on managing the aggregate risks from low probability, high consequence effects of AI systems, and the need to proactively ensure the alignment of evermore powerful advanced or general AI systems.

The first lens NIST should consider is the aggregate systematic impact of small effects by AI systems that, when deployed on a massive scale, can lead to harms of a societal magnitude. To properly evaluate the long-term implications of any AI system, it is imperative that NIST develop indicators, best practices, guidelines, and standards, among other tools, with the objective of enabling organizations to gauge the aggregate effects of their AI-based products and services. For example, consider the possibility of a catastrophic safety failure in the performance of autonomous vehicles for conditions that are hard to model or anticipate in a training environment. When this technology is deployed on a massive scale, society may encounter the safety failures with a low, but unacceptable, frequency. Another instance to consider are the aggregate effects on user polarization and radicalization that can occur with content recommendation algorithms in social media. Though the vast majority of users may be relatively unaffected by being unintentionally recommended polarizing or radicalizing content, if even a small percentage (e.g., 0.1%) does evince negative effects, it can create a societal-wide consequence.

The second lens addresses the prospect of a rapidly evolving slate of AI system capabilities. One concerning class of AI systems can be characterized as increasingly generalized in their purpose. When powerful versions of these systems are not properly aligned with intended objectives, the systems can become extremely unsafe or unpredictable when deployed.

As described by the Office of Management and Budget in their regulatory guidance to federal agencies, "there is a risk that AI's pursuit of its defined goals may diverge from the underlying or original human intent and cause unintended consequences-including those that negatively impact privacy, civil rights, civil liberties, confidentiality, security, and safety" [1]. Certain plausible misalignments of incentives within the learning performed and decisions made by these powerful AI systems may even endanger the sustainability of human life, as has been argued by numerous leading scholars in AI research [2]. In this regard, NIST must develop mechanisms in the RMF that inform stakeholders on how to deal with this possibility and proactively prevent such outsized risks from becoming a reality.

Suggested Change:

To tackle the challenge of long-term, low probability, systemic, and high impact AI risk, NIST should proactively incorporate two lenses into its RMF: tools and methods for evaluating the aggregate risk of AI systems' effects on society and the alignment of evermore powerful advanced AI systems, including those that are used as a foundation for other AI systems. Concretely, this could entail: training on pitfalls and safety remediations for powerful and potentially misaligned systems; increased attention to developing further technical and governance remediations for such issues; increasing sensitivity to post-market effects of more general AI systems through monitoring to identify precursor safety problems; increasing NIST's investment in AI safety research; introducing process standards for characterizing risk level and for proper design and implementation of AI systems' incentives; and developing safety requirements for applying for government AI research and development funding. Further, as has been agreed upon by thousands of leading AI experts and technologists in the Asilomar AI Principles (specifically its 21st principle), it is imperative that NIST incorporate catastrophic and existential risks into its purview and subject them to planning and mitigation testing efforts commensurate with their expected impact [3].

[1] U.S. Office of Management and Budget, "Guidance for Regulation of Artificial Intelligence Applications", M-21-06, p. 13, available at https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf.
[2] See, for example, Russell S (2019), Human Compatible: Artificial Intelligence and the Problem of Control. Penguin, New York; Christian, B. (2020). The Alignment Problem: Machine Learning and Human Values. W. W. Norton & Company; and Critch, A. and D. Krueger, AI Research Considerations for Human Existential Safety (ARCHES). arXiv preprint arXiv:2006.04948, 2020.
[3] Created at an FLI organized workshop in 2017, the Asilomar AI Principles are signed by over 1,700 leading AI and robotics researchers, and over 3,900 other prominent individuals. For more, see: https://futureoflife.org/ai-principles/.

II. COMMENT

Requested topic:

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

Response/Comment (Include rationale):

AI actors are challenged to manage the risks of AI systems when there is uncertainty over how the systems will be used, or misused, during deployment. This problem will amplify with the growing trend in AI research and development toward increasingly generalized AI systems capable of performing a wider range of tasks and activities within a single system or model. Such systems may also serve as foundations for other AI systems and applications [4]. These systems pose a unique challenge in managing AI risk in that there is no single use for them and the range of uses and misuses is extremely uncertain. While developers may have some scenarios in mind for how the systems might ultimately be used most beneficially or harmfully, the creativity of the public can easily go beyond the imagination of AI actors (especially if the system is open-sourced). High uncertainty about the end use cases of a product or service is not unique to the broader field of risk management, but it is relatively unique to the nascent field of AI risk management, which has typically depended upon a more rigid understanding of how a system will be employed to assess its likely risk [5]

In NIST's RFI, in the section on Genesis for Development of the AI Risk Management Framework, it states that "With broad and complex uses of AI, the Framework should consider risks from unintentional, unanticipated, or harmful outcomes that arise from intended uses, secondary uses, and misuses of the AI" and that the RMF should "be adaptable to many different organizations, AI technologies, lifecycle phases, sectors, and uses."

[4] Bommasani R et al. (2021) On the Opportunities and Risks of Foundation Models. arXiv, https://arxiv.org/abs/2108.07258.
[5] For example, see the European Commission's proposal on AI, which depends heavily on use case to determine whether an AI system is "high-risk." See, EU (2021) 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. European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC020.
.

While this is correct and commendable, we believe the RMF needs to go further than just considering a wide range of possible uses. The RMF must not depend on AI actors having high certainty of the expected use and misuse cases of AI systems. Put another way, conceptually, "use" should not be a variable that demands certainty or definition in the mathematical formula employed to calculate the risk of an AI system.

Suggested Change:

There are several ways in that the RMF can facilitate resolution of this problem. For example, the RMF should further develop the explainable AI principle of knowledge limits which NIST has described as when "systems identify cases they were not designed or approved to operate, or their answers are not reliable." Guidance on how stakeholders can implement knowledge limits will help AI actors understand when AI systems are being used in ways unanticipated by developers that may not be reliable or safe. Additionally, the RMF should prescribe mechanisms for red-teaming such considerations for systems of relevant risk level. The RMF would also benefit from including specific guidance for how to manage the unavoidable uncertainty about the uses of generalized AI systems that AI actors will encounter, including for systems used as a foundation for other applications.

III. COMMENT

Requested topic:

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

Response/Comment (Include rationale):

As has been identified by NIST in prior work by its experts, there are limitations to current methods of evaluating user trust in an AI system [6]. AI developers often have a poor model of how the eventual users will interact with an AI system, which in turn can affect the user trust potential and the perceived system trustworthiness. Situations in which there is a mismatch in expectations about the trustworthiness of an AI system between the developer and an end user are more likely to result in the system being misused unintentionally in ways that the developer did not expect. The end user may also give greater amounts of trust to the outputs of the AI system than the developer intended or believed is warranted. Of particular concern, users must have a clear understanding of, and accurate expectations for, the loyaltyof an AI system, as described in Comment V of our submission. As a characteristic of trustworthiness, loyalty is especially pertinent when AI systems serve as part of or in place of human systems with a high degree of fiduciary responsibility, such as doctors, lawyers, therapists, and financial advisers. In general, a trustworthy AI system is characterized by being transparent about "who they are working for," and disclose conflicts-of-interest between their users and creators when those exist. In addition, the user must always be aware of when an AI system is in use. If this is purposefully obfuscated or left undisclosed, users may be easily manipulated or misled.

Suggested change:

The NIST AI RMF should develop tools and guidance for its stakeholders on how to accurately predict or understand the trustworthiness expectations of likely users of AI systems. This can include guidance on communicating about the pertinency and sufficiency of the characteristics of trustworthiness to end users (as identified in prior NIST work). NIST should include as one of these characteristics the concept of AI system loyalty and develop requirements for disclosing the use of AI systems to maintain end user trust.

[6] B. Stanton and T. Jensen (2021). Trust and Artificial Intelligence. NIST. https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8332-draft.pdf

IV. COMMENT

Requested topic:

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;

Response/Comment (Include rationale):

We believe that NIST should include an analysis of the AI system's generality as a characteristic essential for evaluating its trustworthiness. More general systems will require more sophisticated and comprehensive training, testing, and quality assurance processes, and should be expected to make unexpected decisions or actions of a qualitatively different nature from very narrow systems. An assessment of generality would evaluate where a system lies on the continuum between weak to strong AI, or narrow to general AI. An assessment of generality may also implicitly take into account, for example and among other capabilities, the system's: ability to do transfer learning, AutoML, perform robustly in few-shot or zero-shot conditions, operate across modalties, having advanced planning capabilities, and whether it is capable of self-improving or self-modeling [7]. Generality measures would not depend on such features, but are quantitative metrics and applicable for any AI system.

Some systems that have high generalizability may also be used as a foundation for other AI applications, as has been recently noted by researchers at Stanford University [8]. A generality assessment is needed because of a range of issues that may arise in the deployment of systems with greater generality, particularly with regard to challenges identified in our comments to Topic 1 of the RFI.

[7] For additional research on how to assess generality, see F. Martínez-Plumed and J. Hernández-Orallo, "Dual Indicators to Analyze AI Benchmarks: Difficulty, Discrimination, Ability, and Generality," in IEEE Transactions on Games, vol. 12, no. 2, pp. 121-131, June 2020, doi: 10.1109/TG.2018.2883773; J. Hernández-Orallo, "AI Generality and Spearman's Law of Diminishing Returns" in Journal of Artificial Intelligence Research, vol. 64, pp. 529-562, 2019; and S., Legg, and M. Hutter, "Universal intelligence: A definition of machine intelligence," in Minds and Machines, vol 17, pp. 391-444, 2007.
[8] Bommasani R et al. (2021), Id.

For example, systems with higher generality are more likely to produce unexpected aggregate or compound societal risks and have significantly greater uncertainty with respect to their likely use cases. In many ways, AI systems with high generality will demand a different set of risk management techniques than those that have low generality (i.e., very narrow AI systems with highly controlled use cases and use environments). For example, experts agree that AI systems designed to recursively self-improve in a manner that could lead to rapidly increasing quality must be subject to strict safety and control measures [9]

Suggested Change:

Add a new characteristic of AI trustworthiness in the form of "generality" and develop appropriate guidance and tools for evaluating and managing its risk.

[9] See principle 22 of the Asilomar AI Principles, which are signed by over 1,700 leading AI and robotics researchers, and over 3,900 other prominent individuals. For more, see: https://futureoflife.org/ai-principles/.

V. COMMENT

Requested topic:

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;

Response/Comment (Include rationale):

Outside of transparency, fairness, and accountability, we recommend that NIST consider loyalty as a principle that fundamentally shapes the trustworthiness of AI systems [10]. Our proposal stems from the importance of informing individuals about the incentives that motivate the decisions or actions of an AI application. Specifically, we say that an AI agent is loyal to another entity insofar as the agent successfully serves or adopts that end user's goals and interests.

Incorporating loyalty into AI systems is important because of the differences in how they make decisions compared to their organic counterparts. Unlike humans, AI systems are not intrinsically self-interested. In other words, its developers make the determination for what incentives or objectives ought to be followed. If aligned, the system's goals (or objective function) will attempt to satisfy an individual's objectives.

In contrast, a conflict of interest will arise if an AI system attempts to be loyal to the goals of both an individual and another party (e.g. the company that created it) and the interests do not align. In addition, a system may be disloyal by representing itself as loyal to an individual, while prioritizing the interests of other parties instead, which may be discovered later or not. Therefore, loyalty can affect "perceived trustworthiness" because if a user perceives that a system (correctly or not) is serving an interest other than their own, they are less likely to trust it.

However, loyalty need not be binary. Transparently demarcating degrees of loyalty is crucial to avoid scenarios where blind allegiance to an entity leads to non-physical harms, including the loss of user trust in the system. In effect, an AI system could be loyal to several parties or incorporate social maxims (e.g. against breaking international human rights law).

[10] See: Aguirre, Anthony and Dempsey, Gaia and Surden, Harry and Reiner, Peter Bart, AI loyalty: A New Paradigm for Aligning Stakeholder Interests (March 25, 2020). U of Colorado Law Legal Studies Research Paper No. 20-18, Available at SSRN: https://ssrn.com/abstract=3560653 or http://dx.doi.org/10.2139/ssrn.3560653

Suggested Change:

NIST should incorporate the concept of loyalty into its slate of principles for managing the trustworthiness of AI systems.

VI. COMMENT

Requested topic:

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;

Response/Comment (Include rationale):

The Partnership on AI (PAI) has begun to maintain an AI Incident Database that shares useful information about "AI deployment harms or near harms across all disciplines, geographies, and use cases.[11]" In the NIST RMF, this practice and tool could be advanced to create standardized ways for stakeholders to submit information about observed incidents, either about their own AI systems or other systems that they have encountered. A version of this tool for the NIST RMF could be kept confidential or have its data anonymized as needed for encouraging participation from industry, allowing for lessons to be generated without risking reputation harm to particular AI systems.

Suggested Change:

NIST should evaluate creating a reporting and database tool in its RMF to consolidated information about "AI deployment harms or near harms across all disciplines, geographies, and use cases," as has been started by PAI. However, unlike PAI's open-source dependent database, NIST should look to maintain a version of this tool that is confidential and anonymized so that industry practitioners can submit standardized information in a way that does not risk economic or reputational damage to their product.

[11] See : The Partnership on AI's AI Incident Database (AIID) is available at https://incidentdatabase.ai/

VII. COMMENT

Requested topic:

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;

Response/Comment (Include rationale):

The Partnership on AI's AI Incident Database (AIID) is available at https://incidentdatabase.ai/

Suggested Change:

Examine the database developed by Arizona State University that compiles the largest sample of soft law programs dedicated to the management of AI I with the goal of expanding the purview of prospective NIST standards.

[12] See: Arizona State University. Soft Law Governance of Artificial Intelligence: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3855171

VIII. COMMENT

Requested topic:

The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, "AI RMF Development and Attributes");

Response/Comment (Include rationale):

The eight listed attributes for the AI RMF are commendable, but we believe there is a chance that the final attribute, "Be a living document" could lead to an over acceptance of risk in instances where there is scientific uncertainty about an aspect of AI trustworthiness. The eighth attribute states:

"The Framework should be capable of being readily updated as technology, understanding, and approaches to AI trustworthiness and uses of AI change and as stakeholders learn from implementing AI risk management."

We concur with the above statement. However, the RFI then states that NIST "expects there may be aspects of AI trustworthiness that are not sufficiently developed for inclusion in the initial Framework." We are are alarmed by the possible implications of this statement, particularly because the word choice of "sufficiently developed" is highly subjective. To be clear, being a living document does not mean that the RMF should ignore, or leave unmanaged for a later date, risks that be characterized by their considerable uncertainty in the science of AI, especially if these risks can produce significant harm. For example, many problems with the trustworthiness of AI systems are often identified in the scientific literature prior to there being a definitive analysis of how and in which circumstances the problem occurs probabilistically, or without there being definitive techniques for mitigating the problem. [13] The NIST RMF cannot simply ignore such problems for future iterations of the living document until there is greater certainty about the likelihood of the risk or means for its mitigation. As a matter of risk management, it would be especially negligent to ignore uncertain risks if the possible harms could be irreversible and catastrophic [14].

Suggested Change:

NIST should clarify its intentions with regard to how it will address established, qualitatively identifiable risks that are not "sufficiently developed" with regard to the RMF's attribute of being a "living document." We propose that a ninth attribute may be necessary to highlight this need, which could state "9. Effectively communicate novel AI trustworthiness research and means for managing uncertainty." Through this attribute, the RMF should, at a minimum, properly communicate to stakeholders any possible risk, even if it is only to say that the possible risk requires additional attention and research in their development of AI products and services. Other risk management frameworks, for instance in the field of medical devices and pharmaceuticals managed by the Food and Drug Administration, use techniques such as alerts, statements, and official safety communications to highlight particular issues on a rolling basis as new information comes to light about particular possible risks. NIST should develop similar methods for the RMF to enhance the attribute of it being a living document, and specifically embrace ways of communicating potential risk, especially potential catastrophic risks, even when the science may not met the "sufficiently developed" threshold. The RMF should also avoid strong assumptions regarding upper limits on future AI capabilities, as has been affirmed by leading AI scientists, roboticists, and other technologists [15].

13] For example, see the problem of "underspecification" in D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al., (2020) Underspecification presents challenges for credibility in modern machine learning. arXiv preprint, arXiv:2011.03395.

[14] See particularly the work of Cass Sunstein, the former Administrator of the Office of Information and Regulatory Affairs, in Cass R. Sunstein, Irreversible and Catastrophic, 91 Cornell Law Review. 841, 848 (2006), and Cass R. Sunstein, The Catastrophic Harm Precautionary Principle, Issues in Legal Scholarship (2007).

[15] See principle 19 of the Asilomar AI Principles, which are signed by over 1,700 leading AI and robotics researchers, and over 3,900 other prominent individuals. For more, see: https://futureoflife.org/ai-principles/.

IX. COMMENT

Requested topic:

Effective ways to structure the Framework to achieve the desired goals, including, but not limited to, integrating AI risk management processes with organizational processes for developing products and services for better outcomes in terms of trustworthiness and management of AI risks. Respondents are asked to identify any current models which would be effective. These could include – but are not limited to – the NIST Cybersecurity Framework or Privacy Framework, which focus on outcomes, functions, categories and subcategories and also offer options for developing profiles reflecting current and desired approaches as well as tiers to describe degree of framework implementation;

Response/Comment (Include rationale):

Recent research in the design and development of AI soft law mechanisms provides insights into the incentives required for a credible and effective AI soft law program, such as NIST's framework. In this context, incentives refer to the alignment of interests that catalyzes action.

In terms of incentives, there appear to be three reasons that explain why soft law AI programs are implemented or incorporated by stakeholders. NIST should consider each of these in the design of their work. First, soft law is used by governments as a warning system. In essence, it serves to caution stakeholders as to the enforcement or creation of hard law. Second, members of society can act to preempt government action. This means that organizations join forces to create soft law with the purpose of dissuading public authorities from enacting hard law. Lastly, soft law is dependent upon self-interest. In these cases, entities will readily implement any soft law governance effort that produces tangible benefits.

Suggested Change:

NIST should examine research outcomes on the incentives structure that successfully aligned the interests of stakeholders and compelled the implementation/enforcement of AI soft law programs. This is particularly relevant in the framework that is currently under consideration.

[16] C.I. Gutierrez (2021). Identifying Incentives for the Enforcement of Artificial Intelligence Soft Law Programs. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3897486

X. COMMENT

Requested topic:

The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

Response/Comment (Include rationale):

To build an effective framework, we suggest that NIST provide guidance in two key areas that would benefit its design and implementation.

First, it should explicitly state the organizational roles of AI actors required for implementing the framework. In effect, the RMF should delineate the accountability of AI actors as defined by the OECD AI Principle 1.5, which has been endorsed and adopted by the U.S. government. The OECD defines AI actors as "those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI." The actors "should be accountable for the proper functioning of AI systems ... based on their roles, the context, and consistent with the state of art."

Concretely, a core purpose of the RMF should be to provide practical guidance on how US stakeholders can fulfill this OECD principle through governance. For example, it should include the roles and responsibilities needed for activities related to implementation, monitoring and evaluation of long-term safety risks, auditing, and external communications in the form of risk reporting. Particularly important is the composition of the teams that design and deploy AI systems. These should contain individuals with diverse experiences and knowledge to inform the organization about potential AI risks. In addition, we believe individuals developing AI systems need to have options to communicate risks that are disregarded by their superiors. Providing this recourse is fundamental to ensure that risk-laden AI systems are not brought to market. To do so, options such as an internal ombudsman or whistleblower protection need to be explicitly considered.

Second, the framework should suggest governance processes that enable organizations to identify and avoid the use of AI systems that generate safety and security risks. Once again, the RMF should strive to identify how US stakeholders can implement a principle of the OECD, in this case, 1.4 on "Robustness, security, and safety."

In particular, the RMF should help guide how AI actors can ensure AI systems are "robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk." In particular, the RMF must focus on issues or factors that pose significant local, regional, or global risks impacting social norms, rights, or values. To do so, it may include methods and resources for characterizing risks and identifying appropriate remediations for these risks.

Suggested Change:

It is in NIST's interests to consider two areas where guidance on their framework would benefit its stakeholders: 1) organizational roles necessary for its implementation; and 2) processes to identify safety and security risks of AI systems.