Elham Tabassi
Chief of Staff, Information Technology Laboratory
National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, MD 20899

Subject: First Draft of the NIST AI Risk Management Framework

Dear Ms. Tabassi,

The Future of Life Institute would like to thank NIST for allowing organizations from all sectors to jointly participate in the development of the AI RMF. The evolving capabilities of AI's methods and applications make it necessary for the US government to provide stakeholders that develop and/or deploy these technologies with a soft law alternative to harmonize the assessment of their impact on individuals and groups. It is our hope that this effort culminates in a document that emphasizes the long-term, low probability, and high-impact outcomes that are possible with our reliance on increasingly powerful AI systems.

Attached to this letter, you will find our contribution to this effort. It consists of five sections that answer several of the questions posed in the first draft of the AI RMF. We hope that our feedback provides a useful external perspective for improving the current draft and building a version 1.0 that addresses our concerns. As you will see below, our comments focus on issues such as: catastrophic and unacceptable risks, loyalty of AI systems, the documentation of risk calculus, and the risk management of general purpose systems, among others.

We thank you for this opportunity and ask that you contact Carlos Ignacio Gutierrez at carlos@futureoflife.org if further information on our response is needed.

Regards,

The Future of Life Institute

1. Whether the AI RMF appropriately covers and addresses AI risks, including with the right level of specificity for various use cases.

FLI commends NIST on highlighting "low probability" events with a "high likelihood for adverse impacts" or ones that are "not easily foreseeable" in the first draft of the AI RMF (page 5 lines 28-30 and page 6 lines 12-14). Our organization believes in the value of emphasizing the potential for AI systems to catalyze catastrophic risk, as was stated in the AI RMF concept paper (page 2 line 15). We suggest the re-incorporation of this idea when version 1.0 of the document is published. Doing so will serve to contextualize the possibility that ever-more capable AI systems may incite global long-lasting direct and indirect negative social effects. In addition, NIST should consider incorporating into its AI RMF the notion that low probability problems with small-to-medium consequences for an individual can aggregate into large societal issues when AI systems are deployed at scale.

With its AI RMF, NIST also has an opportunity to shape how stakeholders gauge the risks generated by their systems. This document would benefit its target audience by characterizing a class of risks that create an unwanted level of harm that should be subject to significant additional safety measures, or even be considered for a complete or partial prohibition. A good example of this classification is present in the evaluation of algorithmic risks by the Data Ethics Commission of the German Government (see here). Therefore, regardless of how an entity proposes to evaluate or mitigate risks in this category, it is our view that NIST can proactively deem applications and methods that fit this description as unacceptable and create a negative stigma towards their development and deployment.

Lastly, section 4.2.2 discusses risk thresholds that serve as operational limits or triggers for concrete decision points. Considering that these thresholds will most often be crossed within the development and prototyping of a system, FLI recommends that the AI RMF specifically mentions that exceeding an entity's self-imposed tolerance would typically prompt an iteration on improvement of the system so that it returns within the tolerance levels. This recommendation may help prompt R&D departments to adopt the AI RMF earlier in the system's lifecycle.

See: Opinion of Data Ethics Commission. German Federal Government. https://www.bmj.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2#page=19

3. Whether the AI RMF enables decisions about how an organization can increase understanding of, communication about, and efforts to manage AI risks.

All types of organizations will employ the AI RMF to map, measure, manage, and govern their AI risks. To minimize the incentives for under-reporting negative outcomes, NIST should request that entities document their calculus for weighing or balancing a system's positive and negative risks. This is most relevant for categories 3 and 4 of the Map function. Explicitly documenting this process is important because it evinces the evaluation of what is deemed an acceptable risk. This type of analysis is inherently a subjective process and requesting that it be documented highlights how these individuals value the "upside risk" of their systems versus the remaining residual risk. In addition, NIST should provide clear guidance for deciding when an entity ought to stop the development or refrain from deploying an AI system and note that the upside and downside risk should not be netted together.

To increase an entity's understanding of risks that present themselves in aggregate contexts, FLI recommends providing guidance or examples on the typology of risks that manifest in the context of AI systems. This would provide a concrete grounding for an assessment of systemic, societal, medium to longer-term, and uncertain to lower-probability harms. In considering the analysis of risks, it is relevant to note that many can aggregate to substantial probabilities and harms. Specifically, in the Map function's risk scanning phase, informed by knowledge of how the systems and models in question work to the extent practicable, red-teams should explore the potential for each of the following risks individually and in combination:

- **Accumulated risk**: small harms accumulating over time to form a major harm;
- **Accrued risk**: where events that are low-probability in the short-term, but high-impact, can accrue and build to significant-probability in the medium term;
- **Correlation risk**: where there are adverse events that are not evident in unit tests or accuracy tests, but can be expected to emerge from correlated decisions or correlated actions with a large number of users, instances, or executions of the system;
- **Latent risk**: where harms that will not manifest significantly or at all on system training or release may still be expected to appear with distributional shift, new use cases, or qualitative shifts in capabilities arising from quantitative scaling;
- **Compounding risk**: where harms would be expected to manifest only when either other problems occur or unexpected, but conceivable conditions or interactions, manifest; and,
- **Adversarial risk**: where harms manifest due to the lack of robustness in the system when in the presence of optimization pressures for inputs to induce those harms.

4. Whether the functions, categories, and subcategories are complete, appropriate, and clearly stated.

In the Map function, page 5 line 12, directly after "model management", we recommend adding "and improvement" in order to underscore that this RMF is to be used alongside an iterative process to reduce risks, rather than simply documenting them before release.

In table 3 of the Manage function we recommend replacing "Mechanisms are in place and maintained to supersede, disengage, or deactivate" with "Mechanisms are in place and maintained to temporarily or permanently supersede, disengage, or deactivate." FLI sees the need to underscore that these actions may be temporary to foster greater engagement in the development of plans that consider the breadth of contingencies and suffuse an ethos of improvement.

7. What might be missing from the AI RMF.

Because of the influence the AI RMF will hold, FLI believes it must cover both the range of risks and how they are likely to evolve as AI systems grow in capability. We see two significant "holes" in the present framework. The first is that, in many cases, AI systems will have users whose interests differ from those charged with their design, development, and deployment (together known as "providers"). Two problems may arise from this statement:

(a) there may be a *conflict of interest* between the end user and provider. This leads to a mismatch between the appearance, on the product provider's part, and the presumption, on the user's part, that the system is serving the interests of the user, while in reality serving those of the provider where they conflict.

(b) there is motivation and opportunity for the provider to *manipulate* the user to act in the interest of the provider and not in the user's own interest.

As AI systems become more capable, these risks will be pervasive and the AI RMF's current definition of "trustworthy" would not serve to address them fully. Although the quality of *transparency* comes closest, we would argue it is still insufficient. Our preferred method to address these issues is through the framework of AI *loyalty* and *disloyalty* (see Aguirre, et al. 2020).

Put briefly, loyalty is defined as an AI system that places the goals and interests of its user first and foremost (like a human fiduciary), and transparently discloses conflicts of interests where they cannot be avoided or resolved in favor of the user. It by all means avoids *disloyalty*, taken

to be a mismatch between the expectation of loyalty and the level of loyalty actually provided. Without accounting for the effects of disloyal AI systems, there is nothing stopping, for example, AI therapy, medical, legal, etc. systems, from acting in ways completely in breach of the fiduciary duties human therapists, doctors, and lawyers hold.

Even for roles without a human fiduciary counterpart, conflicts of interest and manipulation are likely to be encouraged by the same market forces that have led to breaches of privacy or security. These reduce trust in AI systems as a whole, doing widespread harm to both the public and good industry actors. Loyalty is FLI's chosen lens to address these risks. Regardless of NIST's adoption of this idea, it is hard to imagine how a system could be considered "trustworthy" if it embodies a strong conflict of interest or is manipulative to its users.

The second concern is centered on "general purpose," "multipurpose," or "foundation model" AI systems. Increasingly, very large pre-trained language models like BERT, GPT-3, Gopher and PaLM, and multimodal models like Dall-E(2), and "Socratic" models, are being developed and deployed as commercial systems. Given the scope of applications (as well as recent impressive demonstrations of capability) it is important for the AI RMF to acknowledge and highlight the risk profile of these systems and how they may evolve.

These systems are likely to require more sophisticated evaluation, testing, and quality assurance processes because they are expected to make decisions, take actions, generate content, or provide recommendations in ways that are qualitatively different from narrow systems. For instance, because of the range of potential applications, the AI RMF should discuss the issue of knowledge limits, which NIST describes 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 systems that recognize knowledge limits will increase how downstream developers or users understand unanticipated and potentially unsafe outcomes. Further, establishing the principle of knowledge limits in a narrow AI system may be substantively different than in a general purpose one. This is because the range of possible uses of the latter demands a greater ability to anticipate a larger universe of outcomes than that of a narrow system.

The scope for risk and harm in a system clearly increases with its capability and range of applicability and action. Therefore, the RMF should be consistent with widely-agreed norms that aim to protect society from the risks of general purpose AI systems. For example, the OECD principles state that "AI systems should be 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." Meanwhile,

the Asilomar Principles highlight that "risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact." Because general purpose AI systems have a wider range of foreseeable uses and misuses, NIST should consider the increased burden of managing their risks to guarantee their robustness, security, and safety.

See: Aguirre, Anthony, Peter Bart Reiner, Harry Surden, and Gaia Dempsey. " AI Loyalty by Design: A framework for governance of AI." (2021). Oxford Handbook on AI Governance (Oxford University Press, 2022 Forthcoming) Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3930338

See: OECD Principles, https://oecd.ai/en/dashboards/ai-principles/P8

See: 21st Asilomar Principle, https://futureoflife.org/2017/08/11/ai-principles/

8. Whether the soon to be published draft companion document citing AI risk management practices is useful as a complementary resource and what practices or standards should be added.

FLI believes that the Practice Guide can be very useful to implementers of the AI RMF. Therefore, the document should include an expanded version of the aggregate risk scanning methodology outlined in answer to question 3, along with examples of each category of risk. FLI also recommends referencing and highlighting examples of unintended harmful behaviors of AI systems, such as those found in the Partnership on AI's incident database (https://partnershiponai.org/aiincidentdatabase/).