January 25th, 2022

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

Subject: NIST AI Risk Management Framework Concept Paper
Via email to AIframework@nist.gov

Dear Ms. Tabassi,

The Future of Life Institute (FLI) applauds and will continuously support NIST's plans to generate a consensus multistakeholder approach towards the responsible and thoughtful design, development, and deployment of artificial intelligence (AI) systems. We believe that the release of this concept paper represents an important step in the right direction toward trustworthy AI systems. It continues a process that may eventually formalize how organizations and individuals should consider risks to both mitigate this technology's negative effects and maximize socially beneficial outcomes.

Enclosed you will find our comments to the concept paper. Out of all of these, we would like to especially highlight two ideas. First, we strongly support your sensible inclusion of the "aggregate risks from low probability, high consequence effects of AI systems, and the need to ensure alignment of ever more powerful advanced systems." In our opinion, a tool that disregards any of the aforementioned ideas performs a disservice to entities attempting to effectively prepare for the actual consequences of AI's methods and applications. Hence, we thank the NIST team for emphasizing these risks.

Second, we understand that the predominant use for the AI RMF will be directed at rather narrow or weak AI. Despite this, we strongly encourage NIST to consider and incorporate "foundation models" and increasingly general-purpose AI systems into all aspects of the AI RMF process. As AI evolves, it is critical for NIST to serve a proactive role in identifying and managing novel forms of this technology. By accounting for increasingly general-purpose AI in such processes, either as a stand-alone product or one that serves as a "foundation" for more narrow or weak AI systems, FLI believes that NIST can adequately prepare AI RMF stakeholders for technologies that may have enduring effects on society.

We thank you for the opportunity to provide our feedback on the AI RMF Concept Paper. Please contact Carlos Ignacio Gutierrez at carlos@futureoflife.org if further information on our response is needed.

Regards,

Anthony Aguirre, Vice President and Head of Policy and Strategy

Jared Brown, Director for US and International Policy

Carlos Ignacio Gutierrez, AI Policy Researcher

Richard Mallah, Director of AI Projects

FLI Comments to Specific Parts of the AI RMF Concept Paper

Page 2, Lines 14-17

We are encouraged by NIST's mention of managing catastrophic scenarios resulting from "low probability" and "high consequence effects of AI systems" in this document. One of FLI's institutional objectives is to underscore such risks with the goal of increasing social awareness of them and assisting in the advocacy and development of the necessary tools, policies, and guidelines for their mitigation. Their inclusion in this concept paper represents an important step to mainstream these ideas. We support NIST in further examining how organizations can improve their preparedness with the first draft of the AI RMF.

An additional issue we would like to comment on is related to the idea of ensuring the "alignment of ever more powerful advanced AI systems." This is particularly the case as consequential decision-making via this technology continues to complement and substitute the work of humans. In this regard, FLI believes it is fundamental to ensure technological alignment with beneficial social objectives. Technical safety forms a key basis by which proper alignment of AI systems should be performed to reduce risk. In systems that are more general purpose, there is a fair amount of overlap between safety and ethics, but safety bears on all the complexities of AI. While safety for AI systems is not a solved topic, many observed and expected pitfalls have been identified and consolidated, and techniques are available to mitigate many such issues (even among those still deemed "open problems"). For more context on this, see:

- Dafoe, Allan et al., Open Problems in Cooperative AI (December 15, 2020), https://arxiv.org/abs/2012.08630
- Hendrycks, Dan et al., Unsolved Problems in ML Safety (September 28, 2021), https://arxiv.org/abs/2109.13916
- Amodei, Dario et al., Concrete Problems in AI Safety (June 21, 2016), https://arxiv.org/abs/1606.06565

An intermediary step to ensure that narrow or weak AI systems comply with this condition is to consider the concept of AI loyalty (see Aguirre, et al. 2020). Loyalty represents the idea that AI systems should be designed to successfully and transparently serve the goals and interests of their end users. While this may seem obvious, it is lacking in many extant commercial AI services, which can contain fundamental conflicts-of-interest. The intersection between loyalty and transparency are key means to avoid the risk of *dis*loyalty, which is a mismatch between the goals and interests a system appears to serve (usually the user's) and those it *actually* serves. This could be due to a failure of alignment or competence, or because it is actually serving the interests of another party (e.g. a company or other system provider). Loyalty in AI systems can be decomposed along several parts that include elements of trustworthiness and codes of conduct governing human roles with fiduciary responsibility.

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

Page 2, Line 40

FLI largely agrees with the scope and audience as stated in the concept paper. The exception is that we would like the document to be more explicit about how the designer population includes not only technical architects, but also product specifiers and system specifiers from business functions, especially Product Managers. To most succinctly and generally address this clarification, we recommend adding "specifying" to this set of functions: "specifying, designing, or developing AI systems."

Page 3, Lines 11-13

FLI appreciates that in the framing of risk, NIST seeks to broaden analyses to both negative and positive influences and factors. Indeed, in many fields, potential positive outcomes or factors are referred to as "upside risk." A holistic perspective is valuable when considering potential effects. However, FLI is concerned that when there is a common category for both upside and downside risk, especially if this is referred to as "risk," the commercial and intra-organizational pressures on AI RMF implementers can lead to disproportionate consideration and documentation of upside versus downside factors.

Considering the above, we have two suggestions. First, to minimize the incentives for under-reporting negative outcomes, NIST should emphasize the importance of identifying them in order for an AI RMF to effectively serve an organization. Second, stakeholders should be asked to document their thinking or calculus for weighing or balancing a system's positive and negative risks. Doing so is important because it evinces how implementers determine what they deem as acceptable risk, which inherently is a subjective process that highlights how they value the "upside risk" of their systems versus the remaining residual risk.

Lastly and relatedly, FLI believes that the AI RMF represents an opportunity for NIST to acknowledge the existence of **intrinsically dangerous applications and unsafe development practices for AI**. As illustrated by the Data Ethics Commission of the German Government in their evaluation of algorithmic risks (see here), there are systems that are so risky, even as prototypes, that they should be subject to significant additional safety measures, or even considered for a complete or partial prohibition. Based on the concept paper, the AI RMF will invite implementers to determine if their system is too risky to proceed with. FLI suggests the inclusion of characterizations and the explicit representation of the above-mentioned pinnacle level of risk as a class to consider. In addition, while NIST itself is not a regulatory agency, the AI RMF should acknowledge the possibility of there being applications or technical development practices that are unacceptable regardless of how a company mitigates their risk, potentially even to put forward an explicit list of these in its first iteration or future versions of the document. Among the AI technologies that concern FLI are: physical and cyber weaponry, AI-powered viruses and hacking tools, and recursive self-improvement as a development technique.

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

Page 4, Lines 23-34

FLI agrees that the identification of risks is an important step to inform stakeholders on the suitability of designing, developing, and eventually deploying AI systems. We ask that the first draft of the AI RMF clearly underscores the relevance of the negative direct effects and externalities of this technology. In addition, it should provide clear guidance for deciding when to stop the development or refrain from deploying an AI system.

Projection or even visualization of realistic potential dynamics, interactions, or effects is something that many people find quite challenging. Intra-organizational pressures may likely discourage the investment needed to uncover legitimate, but hard-to-ideate risks. For this reason, FLI recommends that NIST provision an addendum of common pitfalls and side effects for the AI RMF implementer to consider, with the clear stipulation that the examples provided are illustrative and by no means comprehensive. Such a section will assist in the risk ideation process. For example, NIST could illustrate how a content recommendation algorithm may create societal risk through hard-to-predict ways, as we have seen with the large scale deployment of these models in social media.

Page 4, Lines 35

With respect to the Mapping Function, FLI is supportive of the direction described. However, it is important for NIST to clarify that "domain" and "use" may actually be, and are often likely to be, plural. Using the singular form of these terms can imply that AI systems have a singular intended purpose or application, which will impact how stakeholders analyze their benefits and risks. Consideration of the set of domains and uses of the system should be propagated throughout the document by making these words plural.

Page 5, Lines 1-3

Since the scope of this AI RMF are narrow and generalized systems, FLI sees a need for explicitly stating a system's ability to run autonomously in the list of considerations for risk sources. Hence, the statement that reads "the way in which the AI system is used" would be amended to "the way in which the AI system is used or runs autonomously."

FLI fully supports the guidance on the enumeration of risks. To minimize the level of misunderstanding regarding the definition of "enumeration," as to some it may be taken as a count, we recommend referring to this process as "listing and enumeration" rather than just "enumeration."

Page 5, Line 5

With respect to Note 3, FLI would like to applaud the encouragement of diversifying the set of perspectives. For ideation, analysis, prioritization, and contextualization of AI systems, both expert and situated tacit knowledge are important.

In cases where the intended stakeholders are a very large and diverse group, FLI finds it important for NIST to mention the inclusion of experts who study the stakeholder groups' characteristics and considerations. Including these social scientists in this process enables the representation of the interests of a larger swath of the population that may not be directly represented in an organization's team.

Page 5, Line 12

With respect to the Measure Function, FLI agrees that tracking post-deployment risks is crucial and we applaud its mention. In order to encourage the integration of a more comprehensive tracking plan and procedure on the part of implementers, FLI encourages NIST to add to the list of detailed considerations in Note 4 a mention of "appropriate frequency of evaluation", i.e. addressing temporality and the periodic nature of such considerations and analyses.

Page 7, Table 1

FLI suggests the following changes to Table 1:

ID	Category	Sub-category
Map: Context is recognized, and risks related to the context are enumerated.		
1	Context is established, ~~and~~ understood, and documented.	
2	AI capabilities, targeted usage, goals, and expected benefits over status quo are understood and documented.	
3	Technical, socio-technical risks[7] and direct/indirect harms from individual,organizational, and societal perspectives are enumerated and classified according to their temporality.	
Measure: Enumerated risks are analyzed, quantified, or tracked where possible.		
	Methods and metrics for quantitative or qualitative measurement of the enumerated risks, including sensitivity, specificity, and confidence levels for specific inferences are identified and applied to the enumerated risks.	

	The likelihood of events and their consequences to internal and external stakeholders are assessed and documented.	
	The effectiveness of existing security controls is evaluated and compared to alternatives from best practices.	
	The methods and frequency with which enumerated risks are assessed are documented as a tracking plan and that plan is followed.	
Manage: Enumerated risks are prioritized, mitigated, shared, transferred, or accepted based on measured severity.		
	Cost/benefit analysis (including the cost of not using AI or an assessment of whether an AI system should be developed or deployed in the first place) is performed. Subjective determination of what is acceptable risk for the AI system is explained and documented.	
	Appropriate responses to enumerated and measured risks are identified, assessed considering alternatives, prepared with the participation of relevant internal and external stakeholders, ~~and~~ implemented, and evaluated in a pre-defined time period.	
Govern: Appropriate organizational measures, set of policies, processes, and operating procedures, and specification of roles and responsibilities are in place.		
	The resources – including engineering tools and infrastructure and engineers with appropriate AI expertise required for risk management, including contingencies – are identified.	
	Ongoing monitoring and periodic review of the risk management process and its outcomes are planned, with responsibilities clearly defined, and with the periodic risk assessments informed by updated data in and about the system and its usage.	
	The risk management process and its outcomes are documented and reported through transparent mechanisms as appropriate.	
	Decision making throughout the AI lifecycle is informed by a demographically and disciplinarily diverse team including expertise in relevant risks and other stakeholders.	