All comments will be made public as-is, with no edits or redactions. Please be careful to not include confidential business or personal information, otherwise sensitive or protected information, or any

Comment Template for Responses to NIST Artificial Intelligence Risk Management Framework Request for Information

General RFI Topics (Use as many lines as you like)	**Response #**	**Responding organization**	**Responder's name**	**Paper Section (if applicable)**	**Response/Comment (Include rationale)**
Emphasis on managing risks to society	1	Global Catastrophic Risk Institute	Seth Baum		We appreciate that the RMF mentions risks to society, and we recommend that the RMF appropriately account for risks to society. AI technology already has major impacts on society, including some significant harms. These impacts are very likely to increase as the technology progresses. Therefore, it is vital for the RMF to put risks to society front and center. We recommend that as part of considering risks to society, the RMF include explicit consideration of catastrophic risks. AI technology can pose catastrophic risks to society when applied to high-stakes domains such as critical infrastructure. There may additionally be catastrophic risks to society from the most advanced AI systems, such as "foundation models" that have broad capabilities and applications. Finally, there is the risk of intentional harmful misuse of AI technology. See Brundage et al. (2018) and Bommasani et al. (2021). Additionally, we recommend that the RMF include explicit consideration of future-oriented dimensions of risks. An essential attribute of AI technology is that it is constantly changing. A framework that is based on a static snapshot of AI risks is bound to miss numerous important novel risks posed by future AI technology, including more extreme catastrophic risks posed by the most advanced future systems. Additionally, some risks posed by current AI technology may have important future effects, such as risks to long-term environmental change. It is important for the RMF to account for these risks. References in this table cell: Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. (2021), On the Opportunities and Risks of Foundation Models. arXiv, https://arxiv.org/abs/2108.07258 Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, et al. (2018) The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and
Sensitivity to low-probability / high-severity risks	2	Global Catastrophic Risk Institute	Seth Baum		We believe it will be important for managing societal risks of AI that the RMF include risk assessment and prioritization procedures that provide appropriate sensitivity to low-probability / high-severity risks, especially for high-stakes AI applications. With guidance from the RMF, organizations should not only focus on what seems like the most likely scenarios. Instead, organizations should consider the full range of important risks. One simple approach would be for the RMF to include a commonly used risk analysis formulation that defines the risk of an event as the probability of occurrence of an event within a specified time period, multiplied by the consequence of that event if it occurs. In other words, the risk equation is Risk = Probability x Consequence, or R = P x C. (A variant uses frequency instead of probability, or R = F x C.) Some form of this basic formulation is often used in cybersecurity, engineering, business, public health, and other domains; see, e.g., Endorf (2007, p. 135), Morgan (2017, p. 293), PMI (2017, p. 435), and Stine et al. (2021, p. 32). With that approach, high-consequence events can be rated as important risks even if they are unlikely. As previously mentioned, we recommend that the RMF appropriately account for risks to society and not just risks to the organization. A conceptually straightforward way to do that would be for the risk equation consequence term to include consequences to society, not just consequences to the organization. With this risk formulation, the RMF would have a simple conceptual approach for accounting for low-probability risks that could be important in the context of the deployment of AI systems at large scales and/or in interaction with critical societal systems. We recognize there would be challenges with implementing this approach (see, e.g., Morgan 2017, Ch. 10). We discuss those further below, under RFI topic #1 (challenges).

Continuous updating	3	Global Catastrophic Risk Institute	Seth Baum		We appreciate that NIST already states that the RMF would include regular updating. We want that to remain, and to be applied throughout the RMF. We recommend that risk analyses include risks at multiple stages of an AI system lifecycle, i.e., the sequence of activities that take an AI system from its initial conception to its final use (Cihon et al. 2021), and that they be updated at key stages periodically (e.g., at least annually) or as new information becomes available. This relates to RFI topics 1 and 5, in that a key challenge of AI risk management is its great uncertainties (related to topic 1) and the value of using adaptive strategies to update approaches as new information becomes available (related to topic 5). References in this table cell: Cihon P, Schuett J, Baum SD (2021) Corporate Governance of Artificial Intelligence in the Public Interest. Information 12 (7) 275, https://doi.org/10.3390/info12070275
Responses to Specific Request for information (pages 11,12, 13 and 14 of the RFI)					
1. The greatest challenges in improving how AI actors manage AI-related risks – where "manage" means identify, assess, prioritize, respond to, or communicate those risks;	4	Global Catastrophic Risk Institute	Seth Baum		One important challenge for risk assessment is that many AI developers will not be well equipped to accurately estimate the consequences or probabilities of events, especially of novel or rare events for which little or no real-world empirical data would be available. AI developers may be better able to assess specific factors that would affect consequences and probabilities of events, rather than directly estimating consequences and probabilities, though there may not be consensus at this time on how various factors would affect risks. See the literatures on risk analysis for rare events and elicitation of expert judgment, e.g., Morgan and Henrion (1990, Ch. 7) and Morgan (2017, Ch. 9). It could be valuable for NIST to do research aimed at filling some of these gaps, based on AI risk models from the field of AI safety as appropriate. References in this table cell: Morgan MG and Henrion M (1990) Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press, New York Morgan MG (2017) Theory and Practice in Policy Analysis. Cambridge University Press, New York
2. 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;					
3. 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;					

4. The extent to which AI risks are incorporated into different organizations' overarching enterprise risk management – including, but not limited to, the management of risks related to cybersecurity, privacy, and safety;					
5. 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;					
6. How current regulatory or regulatory reporting requirements (e.g., local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles;					
7. 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;					
8. How organizations take into account benefits and issues related to inclusiveness in AI design, development, use and evaluation – and how AI design and development may be carried out in a way that reduces or manages the risk of potential negative impact on individuals, groups, and society.					
9. The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, “AI RMF Development and Attributes”);					

10. 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; and					
11. How the Framework could be developed to advance the recruitment, hiring, development, and retention of a knowledgeable and skilled workforce necessary to perform AI-related functions within organizations.					
12. 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.					