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 information you do not wish to be posted.

Comments from the Center for Security and Emerging Technology (CSET) at Georgetown University.

Submit comments by August 19, 2021:

General RFI Topics (Use as many lines as you like)	**Response #**	**Responding organization**	**Responder's name**	**Paper Section (if applicable)**	**Response/Comment (Include rationale)**	**Suggested change**
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;		Center for Security and Emerging Technology (CSET), Georgetown University	Center for Security and Emerging Technology (CSET), Georgetown University		Challenge 1: Developing risk management guidelines with enough flexibility to adapt to continued progress in AI research and use. Rationale: AI R&D has progressed rapidly over the last 10 years and is likely to continue to do so. This means that a framework that tries to account primarily for risks that we can observe and anticipate in 2021 is likely to quickly become out of date as research and practice evolves. For example, a risk framework developed 5 years ago would not necessarily have accommodated the advances in natural language processing that we have seen since 2017, or the new use cases created by those advances. A forward-thinking framework needs to be structured around the changing nature of AI systems, and should therefore incorporate a mechanism to update the Framework regularly based on changes in the AI ecosystem. Challenge 2: Accommodating increasingly general AI systems (e.g. so-called "foundation models"). Rationale: At present, AI systems are generally deployed for relatively narrow use cases, so it is natural for risk management to incorporate information about the deployment context and use case of a given AI system. However, the Framework cannot assume that each AI system only has a narrow use case, due to the increasing generality of some AI systems, where one model can be adapted for use in very different contexts with very different use cases. Stanford has coined the term "foundation models" to refer to some such models, referring to the fact that many different products and services can be built on top of the same underlying model. In cases like this, some risks will be dependent on the use case and deployment context of the specific AI product in question, while other risks will derive from the underlying "foundation model."	Challenge 1: Ensure that a method to update the Framework (based on changes in AI research advances, usage patterns, and risk profiles) is "baked into" the Framework. This update process could incorporate trends in the research literature, information from incident reports, etc. Challenge 2: Ensure that the structure of the Framework is compatible with increasingly general AI systems, which may have a wide range of potential application areas, rather than assuming that all AI systems in question are "narrow" AI with only one use case.

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;		Center for Security and Emerging Technology (CSET), Georgetown University	Center for Security and Emerging Technology (CSET), Georgetown University		Alphabet subsidiary DeepMind (a leading AI research organization) includes "specification" in its taxonomy of AI safety characteristics (along with "robustness" and "assurance," which are already well captured in NIST's draft). This refers to the challenge of specifying a goal or objective such that the behavior of the system aligns with the operator's true intentions. Misspecification occurs when a system fulfils the literal objective it was given, but does so in an unintended or harmful way, such as a social media algorithm successfully fulfilling the objective of keeping users on-site by promoting radicalizing content.	Add specification as a characteristic to the Framework.
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;		Center for Security and Emerging Technology (CSET), Georgetown University	Center for Security and Emerging Technology (CSET), Georgetown University		Suggestion 1: The Framework should include definitions and templates that facilitate information sharing about AI risks and incidents. As NIST has recognized, AI's potential uses are vast and ever-changing, making it difficult to know in advance where problems are likely to emerge. Standardized ways to share information about incidents would be very valuable for identifying, assessing, prioritizing, mitigating, and communicating AI risk. At present, the closest thing available is the Partnership on AI's AI Incident Database (AIID, https://incidentdatabase.ai/), which collects information on AI-related incidents. But the AIID is limited to publicly available information, e.g. from media reports, and therefore often struggles to collect relevant information about the circumstances and causes of a given incident. Recommendations from NIST of a standardized format and/or venue to report information after an AI incident occurs would be very valuable in encouraging AI developers in the private and public sector to share relevant information, and increasing the consistency of the information shared. Suggestion 2: Use standardized ways to classify AI systems in order to more efficiently identify, assess, and communicate risk. That is, start by classifying a system according to multiple dimensions (e.g., breadth of deployment, type of data inputs, task) in order to assign it to a broad, pre-defined risk category. This broad category can then be used to determine how to proceed, e.g. systems in one category might be automatically required to undergo an in-depth risk assessment, whereas others might not. Organizations such as the European Commission and German Data Ethics Commission have proposed "risk level" classifications of this kind.	Suggestion 1: Include standardized templates for reporting information about AI incidents, which AI developers could voluntarily adopt. Suggestion 2: Map the Framework to an AI systems classification framework, e.g. along the lines of the OECD Framework for Classification of AI Systems (https://oecd.ai/wonk/classification).

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;		Center for Security and Emerging Technology (CSET), Georgetown University	Center for Security and Emerging Technology (CSET), Georgetown University		In fields adjacent to AI, U.S. government-run incident reporting systems enable the collection, structuring, and analysis of information about real-world failures. Developing a similar system to systematically track AI incidents would help regulators understand where to focus their efforts, encourage companies to improve their AI products, contribute to greater public awareness of AI's limitations, and inform technical initiatives to make AI safer and more secure. This type of reporting infrastructure could be combined with regulation mandating incident reporting (as is the case in cybersecurity, aviation, marine transport, chemicals, and occupational health and safety), or could be created for voluntary participation.	Consider developing infrastructure required for incident reporting as part of the Framework.
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.		Center for Security and Emerging Technology (CSET), Georgetown University	Center for Security and Emerging Technology (CSET), Georgetown University		The Global Partnership on Artificial Intelligence (GPAI) is undertaking a project on data justice, which aims to move beyond understanding data governance narrowly as a compliance matter of individualised privacy or ethical design, and to include considerations of equity and justice in terms of access to and visibility and representation in data used in the development of AI/ML systems.	Refer GPAI's work on data justice for guidance on inclusiveness in AI design.
9. The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, "AI RMF Development and Attributes");		Center for Security and Emerging Technology (CSET), Georgetown University	Center for Security and Emerging Technology (CSET), Georgetown University		We suggest adding two additional attributes: To ensure usability by being judicious in how much information about the AI system, and which kinds of information, are required to use the framework. Plain language and clear definitions are important, but are not helpful when being used to ask for information that is not available or easily accessible. Likewise, if the framework requires an excessive number of items, the need to consult multiple sources, etc., this will reduce usability (and thus reduce voluntary utilization of the Framework). In our own usability testing for several AI system classification frameworks, we also found that providing a rubric, or summary matrix, of the core framework dimensions and their defining categories that a user can refer to quickly made the framework more usable.	Add two attributes: The intended length and/or number of items or categories to be included. One option could be to designate some types of information necessary for the Framework as "core categories," with less essential information designated as such. An accompanying framework summary rubric or matrix for users to have on-hand when using the Framework.

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		Center for Security and Emerging Technology (CSET), Georgetown University	Center for Security and Emerging Technology (CSET), Georgetown University		Structure the Framework to accommodate, or ideally incorporate, existing processes for tracking or classifying AI systems to lower the cost of added reporting. A classification framework like the one CSET is developing in collaboration with OECD is one such complementary process that may be used for organizational management of AI systems and risks. If organizations are already providing information on a system (e.g., via the classification framework), it would be ideal for that process to automatically assign a system to a predefined risk category, as opposed to requiring the completion of an entirely new framework.	Structure the Framework to accommodate, or ideally incorporate, existing processes for tracking or classifying AI systems to lower the cost of added reporting.
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.		Center for Security and Emerging Technology (CSET), Georgetown University	Center for Security and Emerging Technology (CSET), Georgetown University		NIST could consider including education or training criteria for those working on AI-enabled capabilities or solutions on AI responsible use, ethics, and bias.	Include education or training criteria for those working on AI-enabled capabilities or solutions on AI responsible use, ethics, and bias.
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.						