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.

Comment Template for Responses to NIST Artifical Intelligence Risk Management Framework

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**
Keep focusing on and delineate the meaning of societal-scale issues, to include: risks to democracy and security; risks to human rights and wellbeing; and global catastrophic risks.	1	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		We appreciate that NIST has dedicated substantial attention to societal-scale issues in the AI RMF RFI, in addition to individual and group risks. We recommend that the focus of impacts on society remain and for the meaning of societal-scale issues to be expanded to include: 1.Risks to democracy and security such as polarization, extremism, disinformation, and social manipulation; 2.Risks to human rights and wellbeing including equity, environmental, and public health risks; and 3.Global catastrophic risks including risks to large numbers of people caused by AI accidents, misuse, or unintended impacts in both the near- and long-term. These categories are not mutually exclusive, and other categories also could be worth including. Risks to Democracy and Security Societal risks include that personalized disinformation (enabled by AI) on social media (e.g., through Twitter bots, synthesis of massive datasets from Facebook, deepfake videos) can sway elections (Brkan 2019) and incite genocide (Mozur 2018). AI-enabled automated surveillance systems could suppress dissent, and hackers can use AI to augment their capability for cyberattacks, including on critical infrastructure (Brundage et al. 2018). Risks to Human Rights and Wellbeing In addition to risks to democracy from AI-enabled disinformation, we have also seen throughout the COVID-19 pandemic the role of mis- and disinformation on public health outcomes, which is a major component of human rights and wellbeing. The 2021 National Defense Authorization Act (NDAA) authorizes the Secretary of Commerce to establish a National	We recommend that the meaning of societal scale issues be expanded to include: risks to democracy and security such as polarization, extremism, mis- and disinformation, and social manipulation; risks to human rights and wellbeing including equity, environmental, and public health risks; and global catastrophic risks, including risks to large numbers of people caused by AI accidents, misuse, or unintended impacts in both the near- and long-term.
Risk assessment approaches focused on intended use cases have important limitations.	2	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		Consideration of intended AI use-cases is valuable and necessary, but not sufficient, for identification and assessment of important AI risks. We appreciate that NIST goes beyond focusing on intended use cases in the AI RMF RFI section Supplementary Information, Genesis for Development of the AI Risk Management Framework. That section 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." However, NIST does not clearly indicate scope beyond intended use cases when the NIST AI RMF RFI section Supplementary Information, AI RMF Development and Attributes, attribute 5, states that "…The Framework should assist those designing, developing, using, and evaluating AI to better manage AI risks for their intended use cases or scenarios." A focus on intended use cases could miss other foreseeable use cases and misuses. The limitations of a use case focused approach become more important as new AI systems become increasingly general in capability, with greater potential for adaptation to new uses (and misuses) across application domains. As an example of new AI systems with increasing generality of applicability, GPT-3 generated text with performance comparable to, or in some cases better than, task-specific fine-tuned systems (Brown et al. 2020). For discussion of the importance of considering potential misuse of AI, see, e.g., Brundage et al. (2018). The EU AI Act also includes the general idea of considering "reasonably foreseeable misuse" along with an "intended purpose" of an AI system (EU 2021). We recommend that the RMF include clear, usable guidance on identifying and assessing risks of potential uses, yielding risk management strategies that would be robust in the face of high uncertainty about future potential uses and misuses beyond the AI designers' originally intended/planned uses. For example, to anticipate potential misuses,	We recommend that the RMF include clear, usable guidance on identifying and assessing risks of AI, yielding risk management strategies that would be robust despite high uncertainty about future potential uses and misuses beyond the AI designers' originally intended/planned uses.

The nascent but growing field of AI safety is providing insights about AI risks and risk management.	3	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		While much of the work in the field of AI safety is at an early stage, it has already yielded some general principles and tools that we expect could be useful to NIST stakeholders. For examples of resources that include concepts or tools for technical specialists in testing key aspects of AI safety, see Amodei et al. (2016), Ray et al. (2019), and OpenAI (2019a, 2019b). Work adjacent to the field of AI safety has also highlighted the distinctive risks of formal models and real-world systems. This includes distinguishing the optimization of some represented task as part of a model vs. establishing control and stability over the dynamics of the domain in interaction with a given AI system. For a sociotechnical presentation that highlights important dimensions of this problem, see Andrus et al. (2020) and Dean et al. (2021). The lack of clear or agreed-upon definitions for terms like "trustworthiness" and "safety" is now being examined by safety researchers (Dobbe et al. 2021). In addition, the Georgetown University Center for Security and Emerging Technology (CSET) briefs on AI safety provide summaries for broad audiences; see Rudner and Toner (2021a, 2021b, 2021c). Our points on this cross-cutting topic relate to several specific topics in the RFI, including: challenges in risk management (Topic 1), definitions of AI characteristics such as safety (Topic 2), AI risk management principles (Topic 7), and risk to society (Topic 8). We recommend that the NIST Framework consider the nascent but growing field of AI safety in informing its deliberations.	We recommend that the NIST Framework consider the nascent but growing field of AI safety in informing its deliberations.
NIST should continue to maintain awareness of progress in AI safety and other key fields, and update corresponding components of the RMF as needed.	4	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		The AI field has changed significantly over the last five years, and is likely to continue to change, perhaps even more dramatically. Ongoing research, particularly in such critical domains as AI safety, security, and capabilities will demand that the Framework is flexible enough to withstand potential shifts, and that NIST update corresponding components of the Framework as needed. To follow shifts across these fields and potential impact on the RMF, we recommend that NIST maintain close relationships with researchers in key fields, such as AI safety, security, and capabilities. These include researchers at three UC Berkeley research centers: the Center for Human-Compatible AI (CHAI), the Center for Long-Term Cybersecurity (CLTC), and the Center for Information Technology Research in the Interest of Society (CITRIS). Our points on this cross-cutting topic relate to several specific topics in the RFI, including: challenges in risk management (Topic 1), definitions of AI characteristics such as safety (Topic 2), AI risk management methodologies (Topic 5), and risk to society (Topic 8). We recommend that NIST maintain close relationships with researchers in key fields (including AI safety, security and capabilities) to follow shifts across these fields and potential impact on the RMF, and that NIST update corresponding components of the Framework as needed.	We recommend that NIST maintain close relationships with researchers in key fields (including AI safety, security and capabilities) to follow shifts across these fields and potential impact on the RMF, and that NIST update corresponding components of the Framework as needed.
Coordination of standards for risk identification and mitigation, to the extent possible.	5	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		The NDAA requests that NIST ensure the Framework "align(s) with international standards, as appropriate." Development and deployment of AI systems is often global. To better support efficiency and effectiveness in implementation of standards to identify and mitigate risks of AI, NIST should coordinate development of any AI standards with standards development organizations, including the Institute of Electrical and Electronics Engineers (IEEE), the International Standards Organization (ISO), the International Electrotechnical Commission (IEC), the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC), among others. While standards may provide guidance on appropriate criteria to evaluate AI, it is important that standards are carefully developed to ensure relevant criteria are considered. If criteria in the Framework and corresponding standards are too narrow, they may inadvertently overlook potential risks. NIST's commitment to a flexible Framework that is consistently updated is critical to ensure appropriate identification and mitigation of risks. Our points on this cross-cutting topic relate to AI RMF attribute #7, as well as RFI topics #1 and #5. We recommend that NIST be explicit about how and where the RMF will incorporate and coordinate with existing and future AI standards development and risk assessment.	We recommend that NIST be explicit about how and where the RMF will incorporate and coordinate with existing and future AI standards development and risk assessment.
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;	6	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		A general challenge is the identification, assessment and prioritization of risks that could have high consequences for society but may seem to be outside the typical scope of consideration by an organization's AI designers. One reason is that many high-consequence risks would involve novel or low-probability events, or systemic risks, that could seem very unlikely or outside the scope of the organization's direct responsibility. Moreover, organizations have limited resources for risk identification and risk mitigation. Furthermore, guidance available on identifying and assessing low-probability, high-consequence risks is likely less standardized and straightforward than typical guidance for identifying and assessing more common types of events (e.g., for standard information-system risk assessment). Thus, the RMF presents an opportunity for NIST to address these gaps and to guide organizations to consider risks of events with high consequences for society. The RMF also represents an opportunity within a voluntary framework to remind organizations of reasons why they should consider events with impacts to society, e.g., identifying risks to the organization's reputation if an AI project becomes associated with undesirable societal-level outcomes. However, there are substantial challenges in addressing risks to society within a voluntary framework. Yeung (2021, p. 20) argues that such approaches as taken in the voluntary Privacy Framework may not be sufficient for the AI RMF: "Because [risks from use of AI systems] might cause physical harm or violate fundamental values, NIST should also incorporate more stringent elements in the AI risk management framework than were in the privacy framework." As one way to address such challenges with voluntary frameworks, we suggest NIST consider coordinating guidance and other policy instruments including standards, at least for some domains. This could include collective proprietary attention to known risks, structured audits to help monitor poorly-understood domain dynamics, and/or certifications preceding deployment in high-risk settings.	We recommend that the RMF provide guidance on risk identification, assessment and prioritization processes to include risks that could have high consequences for society but may seem to AI designers to be outside the typical scope of consideration for their organization, such as events that would be novel or low-probability events, or systemic risks, or expected to be outside their typical time horizon.
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;	7	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		For definitions of AI safety (as well as reliability, robustness, security, and harmful outcomes from misuse), see AI safety research agendas and publications such as Amodei et al. (2016). Part of the work of safety is to build systems that remain under human control and are demonstrably subject to human oversight and periodic external evaluation. For one prominent example of technical work in this direction, see Hadfield-Menell et al. (2016). We suggest that NIST consider "assessment of generality" (i.e., assessment of the breadth of AI applicability/adaptability) as another important characteristic affecting trustworthiness of an AI system, or perhaps as a factor affecting one or more of the AI trustworthiness characteristics NIST has already outlined. If an AI has not undergone any assessment of its generality, that would suggest lower trustworthiness. If assessment indicates high generality of an AI, we expect it would be appropriate to conduct more in-depth risk assessment, more assessment of use cases beyond the originally intended use cases, longer time horizons in risk assessment, more continuing assessment, etc. (Ideally, a generality assessment process would be quick and low-cost for the majority of AI with low generality, while accurately identifying the smaller number of AI with high generality.) For discussion of AI generality, see e.g. Bommasani et al. (2021). For definitions of explainability, it is important to understand how the term has been used differently by various stakeholders and how in practice it has often failed to meet its objectives (Newman 2021). The definition of fairness is similarly contested (Mulligan et al. 2019).	We recommend that NIST consult with a diverse set of stakeholders, including risk-sensitive groups, for input such as on definitions of key terms to better understand how the terms have been used differently by various stakeholders. We also recommend that NIST consider "assessment of generality" (i.e. assessment of the breadth of AI applicability/adaptability) as another important characteristic affecting trustworthiness of an AI, or perhaps as a factor affecting one or more of the AI trustworthiness characteristics NIST has already outlined.
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;	8	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		Additional principles which should be considered are sustainability and inclusivity. For example, one of the OECD AI principles is, "AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being." Other AI risk and impact frameworks have also included these considerations (Yeung 2021). Over 170 sets of ethical AI guidelines have been developed (Algorithmwatch.org 2020). A growing consensus is emerging around the following principles: accountability, privacy and security, transparency and explainability, fairness and non-discrimination, professional responsibility, human control, and the promotion of human values such as civil and human rights. Organizations are taking concrete steps to operationalize AI principles. For example, the OECD Network of Experts on AI is creating a database of tools and practices to implement the OECD AI Principles (OECD 2021). For a more in depth case study on how organizations such as Microsoft are defining and managing AI principles, see Newman (2020). Finally, we recommend that NIST clarify two items in the RMF RFI regarding NIST's use of the terms "characteristics" and "principles". First, we recommend that the difference between principles and characteristics is made more clear. Second, where the RFI states that "These characteristics and principles are generally considered as contributing to the trustworthiness of AI technologies and systems, products, and services", we recommend you clarify to what extent NIST meant "considered by the public", or "considered by experts", or both; differentiating expert and public evaluations of trustworthiness seems both descriptively salient and normatively appropriate. (This relates to RFI section Supplementary Information: Genesis for Development of the AI Risk Management	We recommend that NIST consider including principles of sustainability and inclusivity. We also recommend that NIST clarify two items in the RMF RFI regarding NIST's use of the terms "characteristics" and "principles": 1. That the difference between principles and characteristics is made more clear, and 2. Where the RFI states that "These characteristics and principles are generally considered as contributing to the trustworthiness of AI technologies and systems, products, and services", we recommend you clarify to what extent NIST meant "considered by the public", or "considered by experts", or both.

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;	9	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		Research on organizational safety standards and the incorporation of AI technologies into the commercial aviation industry reveals how the opaque, unpredictable, and accident-prone nature of AI technologies results in slow adoption in safety critical domains. There is demand for collaborative AI safety standards that meet rather than relax aviation's high safety standards (Hunt 2020). References in this subsection: Hunt W (2020) The Flight to Safety-Critical AI: Lessons in AI Safety from the Aviation Industry. CLTC, https://cltc.berkeley.edu/wp-content/uploads/2020/08/Flight-to-Safety-Critical-AI.pdf	
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;	10	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		For effective risk identification, one best practice is to have risk identification processes performed by a team that is diverse, multidisciplinary, representing multiple departments of the organization, as well as including a correspondingly diverse set of stakeholders from outside the organization. See, e.g., guidance on including stakeholders during project risk identification (PMI 2017, section 11.2), as well as guidance on the ranges of types of stakeholders to include when identifying potential types of AI harm (Microsoft 2020). As we mentioned previously, one proposal to manage risks more effectively, reliably, and safely is to incorporate feedback from stakeholders and risk-sensitive groups, democratizing the structure of AI pipelines (Dobbe et al. 2021). The diversity of perspectives from such approaches can help identify a greater breadth and depth of risks that otherwise could be missed by a team without the same perspectives. It would be valuable for the Framework to include templates and definitions to facilitate information sharing on AI risk factors and incidents. Standardized tools for sharing information about incidents and risk factors could reduce costs and increase value of efforts to identify, assess, prioritize, mitigate, and communicate AI risk. For AI incident reporting, one leading effort is the Partnership on AI's AI Incident Database (AIID n.d). Reporting on AI risk factors potentially could adapt procedures and templates currently used in the cybersecurity community for vulnerability disclosure. NIST could provide standardized reporting formats or other means to help AI developers share information in consistently beneficial ways. As mandated in the NDAA, NIST should align its efforts with international standards, as applicable. In doing so, NIST will support the development of standards that support greater efficiency and effectiveness in risk mitigation. We recommend that NIST review the work of the IEEE Ethics Certification Program for Autonomous and Intelligent	We recommend that NIST consider having the RMF include guidance to have risk identification processes performed by a team that is diverse, multidisciplinary, representing multiple departments of the organization, as well as including a correspondingly diverse set of stakeholders from outside the organization. We also recommend that the RMF include standardized templates for reporting information on AI risk factors and incidents, that AI developers could adopt voluntarily.
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;	11	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		For a comparative analysis of AI risk and impact assessments from five regions around the world including Canada, New Zealand, Germany, the European Union, and San Francisco, California, see Yeung (2021). Please also see our discussion above of standards related to NIST AI RMF RFI topic #5. References in this subsection: Yeung LA (2021) Guidance for the Development of AI Risk and Impact Assessments, CLTC, https://cltc.berkeley.edu/2021/08/09/guidance-for-the-development-of-ai-risk-and-impact-assessments/	

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.	12	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		Case studies documented in Newman (2020) detail how institutions including Microsoft and OpenAI have tried to improve the inclusiveness of AI design, development, use, and evaluation and also reduce and manage the risk of potential negative impacts. At Microsoft for example, the Responsible AI Program includes the AETHER Committee, the Office of Responsible AI, a Responsible AI Standard, and a Responsible AI Champs community. Microsoft researchers have also documented the role of checklists in AI ethics and worked on "harms modeling" designed to help researchers anticipate the potential for harm and identify gaps in products that could put people at risk (Madaio et al. 2020, Microsoft 2020). References in this subsection: Newman J (2020) Decision Points in AI Governance: Three Case Studies Explore Efforts to Operationalize AI Principles, CLTC, https://cltc.berkeley.edu/ai-decision-points/ Madaio M et al. (2020) Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, https://dl.acm.org/doi/abs/10.1145/3313831.3376445 Microsoft (2020) Foundations of assessing harm, Microsoft, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/	
9. The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, "AI RMF Development and Attributes");	13	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		While the RMF attributes list currently includes using plain language that is understandable by a broad audience, it does not explicitly include being user-friendly more broadly. Enabling ease of use for diverse stakeholders - for example by including implementation guides - is advised in order to help NIST achieve its goals for the AI RMF. We recommend that NIST consider adding usability as an attribute of the AI RMF.	We recommend that NIST consider adding usability as an attribute of the AI RMF.
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	14	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		We commend NIST for planning to take an iterative approach with AI RMF development. We expect that appropriate, net-beneficial guidance addressing many key concepts (e.g., for some technical aspects of safety) may require more time to develop than would be feasible for inclusion in the initial Framework. We suggest that NIST consider clarifying its planned procedures for making RMF updates (how often, under what conditions, decision criteria), and how it aims to balance flexibility with standard-setting authority. For recommendations on linking the AI risk framework to procurement and purchasing decisions, see Yeung (2021). Yeung (2021, p.19) also discusses how the NIST Privacy Framework, as a voluntary framework, reminds organizations of reasons and incentives to consider risks affecting external stakeholders: "the framework points out how privacy risks can … impact the organization, such as its reputation taking a hit or revenue loss from customers moving elsewhere. This linkage to organizational impact helps to provide parity between privacy risks and other risks that organizations are managing and leads to more informed decision-making." Similarly, the NIST Cybersecurity Framework also mentions that cybersecurity incidents can affect an organization's reputation. However, Yeung (2021, p. 20) also argues that such approaches taken in the Privacy Framework may not be sufficient for the AI RMF: "Because [risks from use of AI systems] might cause physical harm or violate fundamental values, NIST should also incorporate more stringent elements in the AI risk management framework than were in the privacy framework." Analytic dimensions of AI risks and possible domain manifestations are now being explored and mapped by technical and sociotechnical researchers. See Dean et al. (2021).	We recommend that NIST consider clarifying its planned procedures for making RMF updates (how often, under what conditions, decision criteria), and how it aims to balance flexibility with standard-setting authority.
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.	15	(1): AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley; (2): Center for Human-Compatible AI, UC Berkeley; (3): CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley	Anthony Barrett (1), Thomas Krendl Gilbert (2), Caroline Jeanmaire (2), Jessica Newman (1), Brandie Nonnecke (3), Ifejesu Ogunleye (1)		It would be very valuable for the Framework to include a comprehensive set of governance mechanisms to help organizations mitigate identified risks. These should include guidance for who should be responsible for implementing the Framework within each organization, ongoing monitoring and evaluation mechanisms that protect against evolving risks from continually learning AI systems, support for incident reporting, risk communication, complaint and redress mechanisms, independent auditing, and protection for whistleblowers, among other mechanisms. On auditing see, e.g., Raji et al. (2020); on AI incidents see the AI Incident Database (McGregor 2020) and Arnold and Toner (2021). We also recommend that the Framework encourage organizations to consider entirely avoiding AI systems that pose unacceptable risks to rights, values, or safety; related considerations are included in other AI risk frameworks (Yeung 2021). For an example of a leading AI enterprise that reviews applications that would use their AI platform, and disallows unacceptable categories of use cases, see OpenAI (2020). Assessment frameworks that address this include explorations of the problem of "trustworthy" mechanisms for verifying development claims and Z-inspection as a domain-specific approach to risk diagnostics. See Brundage et al. (2020) and Zicari et al. (2021). We recommend that NIST include guidance on governance processes to support the successful implementation of the AI RMF. We recommend reviewing Moss et al. (2021), which outlines "10 constitutive components" of supporting accountability in impact assessments. NIST should provide guidance on ways to support accountability in the implementation of the RMF (e.g., suggesting personnel/management levels that will implement and oversee the	We strongly recommend that the Framework include a comprehensive set of governance mechanisms to help organizations mitigate identified risks. These should include guidance for determining who should be responsible for implementing the Framework within each organization, ongoing monitoring and evaluation mechanisms that protect against evolving risks from continually learning AI systems, support for incident reporting, risk communication, complaint and redress mechanisms, independent auditing, and protection for whistleblowers, among other mechanisms. We also recommend that the Framework encourage organizations to consider entirely avoiding AI systems that pose unacceptable risks to rights, values, or safety.