	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 Request for Infromation (RFI)			**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**
			Aryeh Englander		NIST needs to be proactively reaching out to relevant research communities and organizations, especially the AI safety research community, and the assured autonomy / AI safety engineering community. In particular, NIST should be proactively talking with the following organizations: Assuring Autonomy International Programme (https://www.york.ac.uk/assuring-autonomy/); Center for Human-Compatible Aritificial Intelligence (https://humancompatible.ai/); Center for Security and Emerging Technology (CSET, https://cset.georgetown.edu/); Anthropic (https://www.anthropic.com/); Stanford University Human-Centered Artificial Intelligence (HAI, https://hai.stanford.edu/); Consortium on the Landscape of AI Safety (CLAIS, https://www.clais.org/)	
			Aryeh Englander		The following references seem very relevant, and NIST should be incorporating the issues and suggestions discussed in these references: https://arxiv.org/abs/1905.04223v1, https://www.york.ac.uk/assuring-autonomy/guidance/amlas/, https://arxiv.org/abs/2004.07213v2, https://www.york.ac.uk/assuring-autonomy/guidance/body-of-knowledge/, https://arxiv.org/abs/2108.07258v2, https://cset.georgetown.edu/publication/ai-accidents-an-emerging-threat/	
			I-Jeng Wang		There is a need to imbue risk-sensitive behavior into advanced AI agent to enable shaping of risk-taking strategies consistent with human values.	

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;			Aryeh Englander		Difficult or impossible to test all possible situations including edge cases for very complex AI systems deployed in very complex environments.	
			Aryeh Englander		Many complex AI systems, especially deep learning networks, are essentially black boxes which are extremely difficult or impossible to understand with current techniques.	
			Aryeh Englander		Advanced machine learning systems are liable to discover very novel solutions that satisfy the objective we gave it but which may not satisfy what we actually want. For very advanced systems it becomes extremely difficult or impossible to precisely specify everything that we do or do not want the system to do, which could lead to the AI finding novel solutions that we very much do not want, in ways that we may not know about until it is too late to prevent.	
			Aryeh Englander		There is currently very little serious government-level discussion of globally catastrophic or even existential risks from very advanced AI systems, despite warnings from many experts that very advanced AI may pose such threats within the next few decades. Decision makers often dismiss such concerns as "science fiction" without actually looking at the relevant arguments and evidence.	

			Aryeh Englander		There is a very difficult challenge of setting up auditing, oversight, and governance mechanisms so that actors and organizations actually follow through with the principles they say they agree to. Very often organizations create lists of great principles that they will adhere to, and then those principles end up being mostly a PR piece and they only get very poorly instituted in practice if at all.	
			Aryeh Englander		Technology races between companies or nations is perhaps the greatest factor in AI risks. If implementing safety or ethical concerns turns out to be difficult or costly or if it negatively impacts performance, then a "race to the bottom" becomes highly likely, where competing companies or nations are greatly incentivized to cut corners in terms of safety or ethics. National and international risk mitigation frameworks, perhaps including treaties between nations, may be critical for solving this issue.	
			James P. Howard		The greatest challenge is in quantifying risk. Some risks are easily quantified becauase we know the associated risk of error. However, some risks are unquantifiable due to rarity, lack of data, or lack of knowledge the risk even exists. Catch-all risk management can attempt to capture this, but it is hit or miss at best.	
			Katie Zaback		Lack of quantification of uncertainty/risk (i.e. current AI systems often do not have quantified representation of risk/uncertaintly baked into the algorithm - it's either non-existent/secondary/"soft") Real-world risk is difficult to quantify; even more difficult - quantifying how to determine what is an acceptable "level" of risk [this is usually domain/application specific] Risks can be introducted not just in results/predicitions but in blind-spots, inherent bias or "invisible" risk that might be baked into the underpinnings of the algorithm or that data that drives the algorithm Related: Neural networks (and similar methods) will always carry risk of error (i.e. failure/incorrect predicitons/etc) will ALWAYS be a part of the system.	
			I-Jeng Wang		It is extremely difficult to predict risks associated with dynamic context-dependent adaptation of envisioned online or lifelong learning AI. This is especially concerning due to learned optimization. See https://arxiv.org/abs/1906.01820.	

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;			I-Jeng Wang		Uncertainty estimates on output of advanced ML techniques such as DNN shall be a key elements of any AI models deployed to safety critical domains. A comprehensive and effective uncertainty modeling framework/methodology is lacking and remains an open research problem.	

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;			Katie Zaback		Trustworthiness should be more granular than high-level markers of transparency and fairness. Details assesments related to data would be a good place to start (i.e. describe the data that was and was not present when training; this should be more than data on the *amount* of data, but also the *type* and *quality*). Trustworthiness also relates to human perceptions. For example, a pilot that is in a plane being controlled in part by an AI system might not trust that system with her life in practice, but would in a virtual setting. Thinking about these issues of trust when building (and implementing) these AI systems should be considered. (Good example: DARPA's Alpha Dogfight)	
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;			Aryeh Englander		Government organizations and other large-scale organizations need to be actively incorporating longer-term considerations of risks from very advanced AI that experts anticipate may be coming in the next few decades. There is considerable research that can be done now to mitigate those risks, yet very few organizations are thinking about them.	

			Katie Zaback		Again, this is application/domain specific. Some AI systems might be highly correlated with safety, but not with privacy. Some might represent significant cybersecurity risks, but not concerns of physical saftey. This presents difficulty when building a robust framework for risk.	
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;			Aryeh Englander		NIST may also want to consider the following risk framework for decisions related to longer-term AI risks: https://www.alignmentforum.org/posts/qnA6paRwMky3Q6ktk/modelling-transformative-ai-risks-mtair-project-introduction (the author of this comment is a POC for this project)	
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;			James P. Howard		NIST should look at the economic research on decision theory, especially decision-making under uncertainty. This is used in the economic field to make decisions without complete information. This is, conceptually, isomorphic to the question posed by NIST here. Risk management standards around uncertainty can be used to support how an organization responds to a risk (see, for instance, how an investment bank makes investment decisions).	

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.			Katie Zaback		Reframe "inclusiveness" as "anti-discriminatory measures." Inclusiveness downplays the harm that will be done by AI systems that are improperly trained. Make anti-discriminatory practices when designing/training/deploying AI systems a requirement of a good system - not something secondary to the design of the system. To do otherwise doesn't clearly communicate that a discriminatory system is not only harmful, but does not meet system requirements. For example, a facial recognition system that performs poorly on black faces is not a facial recognition system - it is a *white* facial recognition system (i.e. does NOT meet system requirements).	
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			James P. Howard		NIST could propose a maturity model for risk management under AI decision-making. For instance, at the low end of a maturity model, there is no risk management. At the high-end can be a complete response framework built into an organizational decision-making. By defining as a maturity model, an organization can adapt the requirements to their specific needs, even on a project-by-project basis. In addition, NIST can again look toward the financial community to structure these responses and look toward the Basel Accords for banking risk management and the Solvency II directive issued by the EU for insurance risk management.	

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.						