Subject: Comments on NIST AI Risk Management Framework Initial Draft Paper

From:
Sent: Friday, April 29, 2022 12:06 AM
To: aiframework <aiframework@nist.gov>
Subject: Comments on NIST AI Risk Management Framework Initial Draft Paper

Dear Ms. Tabassi and the entire NIST team developing the AI Risk Management Framework,

Thank you for the opportunity to provide public comment in response to the "Concept Paper on the AI Risk Management Framework." We strongly support the Framework proposed thus far and offer these comments as you prepare the AI RMF 1.0.

We strongly endorse the "Harm to a system" item in Figure 2. It's appropriate to think about the societal-scale impacts of the existence of new technologies.

We agree that many of the most critical risks of AI to individuals and society derive from the inadequacy of aligning the objectives of increasingly powerful AI systems with our societal values. For example, the COVID-19 pandemic is that when low probability, high consequence events occur, they have acute and disparate impacts on historically disadvantaged communities. The pandemic highlighted the need to do more to reduce the chances of these types of risks from occurring.

Therefore, NIST should focus on similar "low probability, high consequence" risks as a critical element of its Risk Management Framework. Guarding against these risks and the potential consequences for disadvantaged populations should prioritize organizations developing and deploying increasingly-powerful AI systems.

Model - Theft Risks:

One significant risk we encourage NIST to include in the AI RMF is the information security of AI systems. For example, a malicious actor could gain access to an organization's powerful AI system that would otherwise be out of reach. As a result, many top neural-network-based AI systems can be cheaply "fine-tuned" to fulfill an objective.

As NIST's Reva Schwartz said, "An AI tool is often developed for one purpose, but then it gets used in other very different contexts," and "Many AI applications also have been insufficiently tested, or not tested at all in the context for which they are intended. All these factors can allow bias to go undetected."

Current events have already demonstrated the reality of these risks, including the hacking of NVIDIA chips and industrial espionage by authoritarian governments. Untested models from leading companies may increase the odds that malicious or incautious actors first deploy untested AI. Acquired world-leading AI technology by theft could be used to profile Minorities, leaving insufficient time for bias testing, alignment research, and robustness checks.

Thus the theft of increasingly-advanced models poses a significant risk to society. We suggest that organizations adhere to the highest implementation tiers (e.g., Tier 4) of the AI RMF, **NIST should advise organizations to adhere to the guidelines necessary to fulfill appropriate corresponding tiers (e.g., Tier 4) of the NIST Cybersecurity Risk Management Framework. Additionally, it could be suitable for the AI RMF to guide specific risks analogous to the Cybersecurity framework guidance. For example, it could be appropriate for organizations to protect AI model weights in fashions, similar to how cybersecurity guidance recommends protecting cryptographic keys.**

Safe shareability:
Another significant risk we encourage the AI RMF 1.0 to address is "safe shareability" as a desirable sociotechnical characteristic that is missing and should be added as a header under the "sociotechnical characteristics" section. Here is some possible content to contribute, explaining this:

Ideas spread quickly in the modern world. Before embarking on a new AI project, organizations should assess: "If outsiders find out this technology exists in a cost-effective form, will it be good for society or bad for society?" If wrong, the technology will not be shareable, i.e., the fact that it exists and the techniques for building it will pose a risk to society. Below are two examples of this problem:

> **Example problem 1**: search-and-rescue drones. The X, Institute wins the ongoing NIST Prize Challenge for Unmanned Aircraft in Search and Rescue for first responder search and rescue operations. After announcing the winner's product, many people point out that terrorists could use the product to search and provide explosive charges. In particular, now that the technology is known to exist in a cost-effective form, enemies of the state may try to steal or rebuild the technology for themselves. This technology is not safely shareable, and therefore it was not a good idea to invent and publicize.
>
> **Example problem 2**: superhuman strategic advisors. The Y Institute develops a cost-effective AI tool to analyze a business and advise on effective social strategies to defeat its competitor businesses. Sometimes it gives unethical advice involving illegal sabotage operations or threats, so before announcing and releasing the product, the creators add an "ethics module" that filters out unethical suggestions. Once the product is known to exist in a cost-effective form, some people point out that a hacker can quickly turn off the ethics module or that the product could promptly be rebuilt from scratch without the ethics module, yielding an ethically unconstrained version. In particular, the unconstrained version could be used by enemies of the state to recommend strategies for overthrowing government or law enforcement. Unfortunately, this technology is not safely shareable, and therefore it was not a good idea to invent and publicize.

Summary:
Because ideas about what is feasible and cost-effective will leak and spread quickly, all else equal, it is better to develop safely shareable technologies. The above are some examples that are not safely shareable. This consideration poses an inconvenience for many potentially exciting AI technologies, but this inconvenience needs to be faced head-on. We realize that cyberweapons will not be 'safely

shareable' and will require an exception for this condition when developed under legitimate circumstances. However, any publically available guidelines from NIST should be asking scientists to consider the issue of safe shareability for any AI system being designed; it is often observed that scientists do not even ask themselves this question.

We encourage the NIST AI RMF 1.0 to issue profiles on AI risks that derive not from the specific use-cases. Still, the resources used and thus the capabilities are likely to be enabled, **particularly considering how the theft of an AI model could negatively affect disadvantaged communities and exacerbate discrimination.** For example, research has shown that as the computational and data resources used in training machine learning models increases, those models' capabilities will often increase as well. Therefore, we suggest the AI RMF include a specific set of profiles for AI systems utilizing large quantities of AI-enabling resources, especially when equivalent amounts are not yet in everyday use. Thus, a system's likely capabilities cannot yet be well-estimated. Similarly, we encourage the "Map" and "Measure" functions of the AI RMF to include categories of procedures for determining the capabilities of AI systems, including those not initially intended or anticipated by the designers.

We also recommend that the AI RMF 1.0 issue detailed guidance for AI systems that will, when deployed, encounter a sufficiently broad set of operating conditions that it is not feasible to test for all possible risks ahead of time preemptively. AI represents an unprecedented challenge for risk management, given the difficulty of anticipating AI system behavior in novel circumstances, combined with the exponential diversity of possible settings that an AI system may encounter during deployment. Finally, to assist organizations in managing numerous such risks, the AI RMF should incorporate best practices from other sectors facing many low-probability, high-consequence risks, like the NASA Risk Management Handbook and the AI risk mitigation experiences of mature self-driving car companies.

In conclusion, we encourage NIST to include detailed case studies of the proper application of the AI RMF to hypothetical and real-world AI systems to make it easier for stakeholders to apply these guidelines. Finally, we suggest NIST issue a Request for Information for case studies from organizations that have developed and deployed AI systems and completed a risk management process that they believe would comply with the AI RMF.

Thank you,