Response to "Request for Information: Artificial Intelligence Risk Management Framework" (86 FR 40810)

Andrew Grotto[1] Greg Falco Iliana Maifeld-Carucci
Stanford University Johns Hopkins University Johns Hopkins University

BLUF: Avoid the temptation to create a siloed AI risk vertical. Adopt a rebuttable presumption that AI risks are extensions of risks associated with non-AI digital technologies unless proven otherwise.

<u>Introduction</u>

Congress has directed NIST to work with the private and public sectors to develop a voluntary risk framework for AI—an initiative we applaud and support. It is a monumental undertaking: whereas NIST's previous initiatives to develop risk guidance and frameworks have focused primarily on a single digital risk vertical—such as cybersecurity[2]—Congress has in effect tasked NIST with developing a framework for managing digital risks generally. AI is, after all, a species of digital technology, and will exist alongside or as a component of non-AI (classical) digital technologies for the foreseeable future.

We think it is vitally important to the success of NIST's initiative, therefore, that NIST frames the AI risk challenge as an extension of its work to empower organizations to manage digital risks, and not as a replacement for it.

To advance this framing, we recommend that NIST and the broader community of stakeholders planning to contribute to NIST's initiative adopt a rebuttable presumption that AI risks are extensions of risks associated with non-AI digital technologies unless proven otherwise. In cases where AI exposes shortcomings in the incumbent catalogue of laws, standards, guidelines and best practices for

[1] Corresponding author (grotto@stanford.edu).

[2] *See, e.g.,* NIST, "Cybersecurity—Overview," *available at* https://www.nist.gov/cybersecurity.

managing digital risk, the first step should be to attempt to reform the incumbent catalogue to remedy the shortcomings in order to improve digital risk management for AI and non-AI technologies alike. If and only if remedying shortcomings within the four corners of existing governance measures proves intractable should the community develop new standards, guidelines, or best practices. And even then, a core design principle should be to strive for coherence across all digital governance measures. The goals should be to mainstream AI risk management, not establish AI as a siloed risk vertical, and to integrate it into digital risk management.

Risky Business

Artificial intelligence has enormous potential to improve lives. Achieving this potential requires establishing governance frameworks that optimize the benefits of AI while keeping the inevitable risks in check. There is a healthy public debate about how regulators, developers, and businesses should approach solving this optimization problem, which spans micro-risks to individuals and groups as well as macro-risks to whole societies. Foreseeable micro-risks include exposing people to privacy risks, unfair bias, and risks to personal health and safety. Foreseeable macro-risks range from job displacement caused by automation to military applications of AI that potentially undermine global peace and security.

It is a dizzying, complex array of risks. Fortunately, researchers from around the world have made dozens of thoughtful contributions on how to mitigate them, especially in terms of identifying general ethical principles for the design and use of AI systems. Arguably—and perhaps surprisingly—there appears to be broad consensus on the set of applicable general principles, which includes such principles as privacy, accountability, security, safety, explainability, and promotion of human values, among others.[3] Governments have even started to take these principles and develop regulatory frameworks around them, with the European Commission's

[3] See Fjeld, Jessica et al. 2020. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.* http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420%0AThis (April 29, 2021).

proposed Artificial Intelligence Act standing out as an especially ambitious and noteworthy example.[4]

Mind the Gap

General principles are an important starting point, but there is a big gap between general principles and actionable risk-based guidance and criteria that are clear and auditable. The challenge for NIST and its community of contributors is how to fill this gap in a manner that avoids treating AI risk as though it were its own siloed risk vertical, distinct from other digital technologies and the governance frameworks applicable to them.

The European Commission's proposed Artificial Intelligence Act makes this mistake. For example, it proposes a requirement that providers of AI systems establish "appropriate data governance and management practices" and use datasets that are "relevant, representative, free of errors and complete."[5] These principles are too general and do not provide actionable guidance. Additionally, the law would require that providers of AI systems develop technical documentation to demonstrate that the system conforms with these rules without providing any further guidance on what should go into the documentation beyond a generalized description of the AI system, its validation and testing data, and performance metrics.[6]

Making matters worse, these general principles are not obviously tethered to other European laws—such as the General Data Protection Regulation, the NIS and proposed NIS2 directives, and the proposed Digital Operational Resiliency Act, among others—and relevant international standards relating to digital technologies.[7] As presently constituted, the Artificial Intelligence Act treats AI risk as a siloed risk vertical.

[4] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM/2021/206 final).

[5] Art. 10.

[6] Art. 11.

[7] Section 1.2 of the Act's explanatory memorandum asserts that the Act will be consistent with other EU laws, but the text of the Act itself is largely silent on this ambition.

What is urgently needed—and what NIST and its contributor community should aim to produce—is actionable, practical guidance that is fully integrated into the broader ecosystem of digital risk management laws, standards, guidelines and best practices. Like any technology, AI presents risks to individuals and society. The fact that the technology is novel, however, does not necessarily mean that the risks associated with the technology are novel. Privacy, security, and other general principles applicable to AI have well-established antecedents in digital risk governance. An AI Risk Framework should build upon these existing governance frameworks, while identifying gaps and areas needing further refinement.

AI Risks: New or Compounding?

NIST and its contributor community will almost certainly identify substantial gaps and shortcomings in the incumbent catalogue of digital risk management resources. These gaps may be the result of the unique attributes of an AI system, such as its development and operational lifecycle. For the development of AI, there may be multiple parties involved in the data ownership and algorithm ownership which can lead to liability and insurance implications. Meanwhile, in the operational landscape of AI, there could be unanticipated externalities that impact successful operation, as well as issues with the projected scale of deployment, and the blurring of customer and vendor. Challenges with explainability in AI systems present another vexing risk management challenge. These and other characteristics about AI could necessitate the development of new taxonomies and methodologies for thorough characterization and measurement.

Some characteristics of AI, on the other hand, compound existing risks by changing their scale or scope. These risks include macro effects like job loss due to technological change, automated systems in warfare, and increased income inequality. Meanwhile, some risks will fall on businesses, customers, and regulators to manage such as privacy and data protection, or cybersecurity. Our proposal for a rebuttable presumption that AI risks are extensions of risks associated with non-AI digital technologies unless proven otherwise would help NIST and its contributor community sort through which of these risks truly requires new taxonomies and methodologies for managing risk and which can be accommodated within incumbent governance frameworks.

Submitted 09/15/2021