credo | ai

Sep 15, 2021

Mr. Mark Przybocki
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899
aiframework@nist.gov

Credo AI Comments on NIST’s “Artificial Intelligence Risk Management Framework” (Docket: 210726-0151)

Dear Mr. Przybocki:

Credo AI is pleased to submit the comments below in response to NIST’s Request for Information on the proposed Artificial Intelligence Risk Management Framework. As a startup working to empower organizations to build AI with the highest ethical standards, we have seen firsthand the challenges that companies face governing AI systems and mitigating AI risks in practice. We believe that NIST’s proposed Framework could have an outsized impact in helping companies to understand “what good looks like” in governing AI, to translate between policy requirements and technical implementation, and to create mechanisms for monitoring and governing AI models and solutions. We particularly believe that the NIST Framework can make a real contribution by considering the role of governance tools and practice throughout the lifecycle of AI models and solutions.

Our filing below offers an overview of the challenges and opportunities we see in the NIST Framework as it relates to ethical AI and its governance. We then directly address three of the questions posed by NIST in the RFI, where we believe Credo AI’s governance expertise and experience on the ground is most relevant to NIST’s inquiry.

Introduction: AI Risk Management and Ethical AI Governance

Credo AI is an AI governance startup with a mission to empower organizations to build AI with the highest ethical standards. Founded in 2020, Credo AI enables AI governance and oversight to promote trustworthy, fair, compliant, and auditable development and use of AI. Through a suite of AI governance, risk management and compliance tools and solutions, Credo AI allows enterprises to measure, monitor and manage AI introduced risks at scale. Over the past 18 months, Credo AI has worked with over 120 organizations including Global 2000 companies, nonprofits, regulatory bodies, research and academic institutions and others to gather insights around the governance of AI.

Artificial Intelligence systems -- including machine learning, statistical and Bayesian approaches, expert systems, reinforcement learning, and autonomous systems -- possess tremendous potential to improve our society and our way of life. We will not realize that promise unless AI systems account for issues related to ethics, bias, trust, privacy, and cybersecurity in their development, deployment, use, and maintenance. In practice, these risks are often compounded by the number and variety of AI models and systems that may be implemented today across an enterprise. Companies also face upstream risk as they rely on the AI systems and data from vendors. As a result, even organizations with a high degree of commitment to ethical AI often lack governance tools to enforce standards, ensure compliance with regulations, or even have visibility on the risks created by the AI systems they are deploying.

Credo AI applauds NIST's efforts to develop a general-purpose AI risk assessment framework. In our experience, such a framework could offer a helpful touchstone for organizations seeking guidelines for deploying ethical AI in a still-nascent regulatory environment. We also believe the proposed attributes for the Framework - including the use of an open and transparent process to create common definitions and standards that are widely accessible - are on point.

Based on our experience in the field with companies building good AI governance mechanisms, the NIST Framework could be of particular help to organizations in several ways:

- **Understanding "what good looks like" in ethical AI**. We see a large number of companies across multiple sectors -- financial services, defense applications, talent and recruiting, retail -- seeking to get ahead of their use of AI and deploy ethically. Today there are few regulatory standards to guide their approaches; while some are in the works, governments are appropriately wary of regulating too quickly in this rapidly evolving area. In other areas, such as employment discrimination or bank regulation, existing law may well provide guidance but enterprises are struggling to interpret those rules in the context of AI and machine learning. We have also seen a proliferation of proposals and research around AI risk management and ethical AI. Organizations are in need of tools to help them assess the many proposed approaches and create a practical framework for managing AI risk. NIST can help by offering a model.

- **Translating between policy requirements and technical implementation.** There is a genuine gap in understanding between AI practitioners and the broad range of stakeholders that have an interest in AI. Professionals in compliance, risk assessment, ethics, or public policy struggle to make their perspective relevant or actionable for AI development. In turn, the technical metrics and quantifications that are the language of the AI practitioner often fail to translate meaningfully into broader policy or business requirements. NIST could offer value in supporting better translation between these stakeholders.
- **Making room for human factors.** Organizations would benefit from clearer standards around "human" or nontechnical considerations, forces shown to have meaningful impacts on the outcomes of the AI development process. Among the non-technical considerations to prioritize: Use case decisions, as to when and where to apply machine learning to a problem; Data decisions, on the selection and manipulation of data to train a machine learning model; Evaluation metric selection for evaluating and validating machine learning models, an incredibly manual and human process; Team diversity, which studies have shown to be one of the most important determinants of ability to confront issues of harmful bias in AI; and approval & individual accountability, creating mechanisms for alloting accountability for the numerous decisions made throughout the ML development lifecycle.
- **Supporting mechanisms for monitoring and governing AI systems over time.** Process-based requirements for governing AI systems will be an essential element of responsible AI deployment. Risk assessment, regular bias testing, compliance assurance, or auditing are among the tools we see companies using today. NIST can aid organizations by offering a better understanding of what decision-making processes,

record keeping, and audit requirements will be considered best practice in mitigating AI risk and promoting ethical AI.

Based on these observations, we believe the NIST Framework can play an important role in helping organizations both manage risk and meet their goals for responsible and ethical AI deployment. We expect that frameworks such as this will play an important role as our societal understanding of AI risk evolves, and while regulatory guidance lags AI deployment. Below we offer responses to several of the questions posed in the RFI, focusing on issues relating to Credo AI's field experience with implementing ethical AI governance.

Response to the RFI's Specific Requests

Topic 1: Challenges in AI

The greatest challenges in improving how AI actors manage AI-related risks—where "manage" means identify, assess, prioritize, respond to, or communicate those risks.

Many are focused on the challenges subsumed by technical AI safety (e.g., robustness guarantees, generalization, transparency, fairness, and alignment). Overcoming these technical hurdles present significant obstacles and will require the focus and investment of a huge range of experts across the world. Even articulating the components of "responsible AI" is an evolving venture, resulting in a multitude of ethical frameworks and principles. While significant investment in these problems has started, broader societal forces are not necessarily aligned with progress towards ethical AI. For instance, individual actors may have greater incentive to race towards ever more performant AI systems rather than prioritizing responsible AI.

Given these dynamics, we believe the most pressing challenges facing AI are system-level problems, related to the processes and incentives surrounding AI development. If not changed, we will fall short of creating broadly beneficial AI systems. More concretely, these challenges can be conceptualized as closing principles-to-practice gaps, which encompass incentive alignment issues, information silos, disciplinary boundaries, as well as the broader challenge of connecting high-level value goals with the technical challenges facing the development of responsible AI systems. The most important of these challenges include:

- **Disciplinary divides in perspective and language.**
  We face a divide in AI sophistication between AI practitioners and a broad range of interested, non-technical stakeholders. In particular, professionals in compliance, risk assessment, audit, ethics, public policy and others, all have critical perspectives on the safe deployment of AI technologies. But communicating their perspective in a way that is actionable at the level of AI development is a significant barrier. In turn, the jargon, metrics and quantifications that are the language of the AI practitioner are not readily translated into a form meaningful to these diverse professionals. This communication barrier is not merely linguistic - there is often a genuine gap in both the values and systems-level understanding between different disciplines. We believe investing in *translation* between diverse stakeholders is the single most important challenge facing AI right now. Without effective translation, we will not be able to bring together the diverse sets of expertise and skills needed to tackle the other challenges facing responsible AI development.

- **Mismatch in the speed of AI and policy development**
  A second challenge is the speed of AI development relative to the policy environment. The unprecedented pace of advancement in AI raises a unique challenge; regulations may either be generic and broadly applicable or specific and prescriptive. In the former case, regulations may apply to many AI systems for years to come but lack teeth and ultimately do little to change the course of AI development. In the latter case, regulations risk becoming irrelevant quickly or styming innovation if they are enforced strictly . We believe there are sets of policies that could strike a balance, but it's likely that they will need to evolve in step with the underlying technology. More broadly, regulating AI systems is simply a new field. Determining which regulatory strategies are effective is ultimately an empirical venture, which our current regulatory systems may be ill-prepared to pursue. Proposals for novel regulatory solutions may deal with these challenges. For instance, third-party governance organizations have the breadth of perspective necessary to evaluate which policies work in different contexts and adapt over time. Government regulatory bodies could interact fruitfully with these organizations to direct policy innovation towards desired outcomes. While we are hopeful that innovative new regulatory frameworks can emerge, the exponential speed of AI innovation is an immense challenge for public policy or any regulatory approach.

- **Uncertain AI risk and unauditable systems**

  A third challenge is our current inability to evaluate AI system risk comprehensively and continuously. We anticipate that AI regulation will need iteration as we learn what approaches work and AI technologies evolve. However, without a clear understanding of which AI systems are risky we have no target with which to evaluate regulatory frameworks. Simply put, one cannot develop safeguards for systems one doesn't understand. Of course, many aspects of technical AI safety feed into this challenge. Fairness, transparency, robustness - these are all aspects of risk articulated by the AI community. However, we believe we are missing large aspects of the development process that are critical for evaluating AI risk. For example, neglected aspects of risk include characterizing the team developing the AI solution, precommitment to evaluation standards, and articulating data provenance. To account for these factors, we need clear audit trails that articulate how an AI system came to be, from conception to productionization. With full audit trails we will begin to have the data necessary to evaluate which system components are associated with poor outcomes when they arise. Audit trails are the foundational building block of an accurate understanding of AI risk.

- **Misaligned incentives**

  Finally, there is the overarching challenge that incentives are not aligned at the individual actor level to improve the safety of our AI systems. For instance, conflicts with transparency (e.g., protecting intellectual property) limit whether any claim about AI systems can be verified, a necessary condition to any other kind of oversight. That said, we believe the challenges above subsume these incentive alignment issues. In particular, we believe regulation and standard setting from governments and certification bodies can identify a set of constraints that will align individual organization incentives with the societal good. However, we believe that identifying those constraints is an immense challenge, and will require faster policy iteration dependent on the effective feedback of more professional disciplines, new, innovative regulatory systems, and better understanding about the ways AI systems are created.

Topic 5: Standards and Frameworks

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.

The “MLOps” landscape has exploded in the last few years. More and more tools are being developed to help organizations manage the complexity of the ML development lifecycle, from data management to development, deployment, and production monitoring. Increasingly, these tools are advertising a set of “governance” capabilities; however, these capabilities generally cater to a very technical definition of governance. These are tools designed to help data scientists and ML engineers identify technical issues with their ML models in production, including explaining model outcomes, detecting unintended harmful bias, and identifying data drift or other issues that might impact model performance.

While technical governance is a necessary part of effective AI governance, it is not sufficient. Governance needs to involve non-technical stakeholders, and governance must be applied to the *processes* and *objectives* of the ML development lifecycle in addition to the *technology* itself. There is a gap in the ecosystem of available tools; responsible AI needs governance tools that bring together multiple, diverse stakeholders to identify, assess, prioritize, mitigate, and communicate AI risk. NIST has an opportunity to help close this gap by establishing a set of standards that emphasize the importance of multi-stakeholder collaboration and effective translation of technical specifications into AI risk.

Topic 12: Governance in the Framework

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

As we have made clear in the other parts of our comment, AI systems are human systems; therefore, we firmly believe that the processes established to govern those systems must consider *human* inputs as much as they consider technical inputs. Human inputs must be considered across the entire ML development process. “Human-in-the-loop” inputs happen at every decision-point in the design and development of an AI system, even if they are not involved in the final output or inference that is made by the application or model.

We feel strongly that the Framework must include requirements around “human” or nontechnical considerations, as these forces are shown to have meaningful impacts on the outcomes of the AI development process. There are five non-technical considerations that we think should be prioritized by NIST’s Framework:

- **Use case decisions**
  When and where to apply machine learning to a problem is the first decision that is made in the ML development lifecycle. The use case for an AI system must play a major role in determining how it is governed. Use case decisions are not only driven by technical considerations but also business concerns, as well as the preferences and priorities of the AI system’s creators. Ignoring the human factors in deciding where and how AI is applied to the real world will lead to ineffective governance.

- **Data decisions**
  While there are many important technical considerations that should be accounted for in an effective governance framework for AI systems, we cannot ignore the human considerations that go into the selection and manipulation of data to train a machine learning model. Selection bias and labelling bias are two examples of issues that are introduced by human factors in the data collection and preparation process. An effective governance framework must ask about the human decisions that are being made about what data to use and how.

- **Evaluation criteria selection**
  We have consistently heard from our customers that the process of determining “what good looks like” when it comes to evaluating and validating machine learning models is an incredibly manual and human process. Business stakeholders, compliance stakeholders, and technical stakeholders must come together to determine what levels of performance, robustness, and fairness are required for a given use case, and which metrics of the many available are best for measuring those levels. Ignoring the effect of organizational power dynamics and individual preferences on this process is impossible, and an effective governance framework must address these human impacts.

- **Team diversity**
  Studies have shown that one of the most important determinants of a team's ability to confront issues of harmful bias in AI is that team's diversity. Less diverse teams have a harder time reducing unintended bias in their machine learning models than teams that are made up of members that come from a wide range of genders, ethnicities, and backgrounds. Furthermore, teams that do not represent the perspective of communities impacted by AI systems have a harder time predicting and mitigating potential harms that the system causes to those communities. We recommend that the Framework pushes organizations to consider the diversity of their teams when evaluating the risk of unintended, harmful bias and potential for unintended consequences in their AI systems.

- **Approval & individual accountability**
  As discussed above, throughout the ML development lifecycle, countless decisions are made that influence and determine the shape of the AI system that is ultimately deployed. Without a means of holding the individuals who make those decisions accountable for their outcomes, there is no incentive to ensure that those decisions are leading to good outcomes. We firmly believe that NIST's Framework must establish a standard for holding individuals and organizations accountable for the decisions that they make about when, where, and how to deploy machine learning models.

If the NIST Framework is going to effectively address all of the potential risks associated with the development and deployment of AI systems, it must acknowledge and address the role that human factors play in AI risk and provide guidance on how to manage these human factors during the ML development lifecycle..

Conclusion

Credo AI's experience in the field has shown us that a broad range of organizations are well along in deploying and relying on AI systems -- and are committed to managing risk and building ethically. NIST's proposed Framework can play a highly constructive role for these organizations by offering a better view of "what good looks like" in this field, and articulating standards and practices even as policymakers move deliberately in the coming years to regulate. We believe the Framework would benefit from directly addressing the governance processes that organizations will need and the tools -- such as bias assessments, external audits, and transparency practices -- to support that governance. Credo AI looks forward to working with NIST and the broader community to build a strong Framework with a standard of practice to help organizations build the ethical AI systems to which they aspire.

Respectfully submitted,

Navrina Singh
Founder and CEO
Credo AI

Susannah Shattuck
Head of Product

Ian Eisenberg
Senior Data Scientist

Alan Davidson
Advisor, Tech Policy