Machine Intelligence Research Institute
Berkeley, California
United States

Point of contact:
Peter Barnett and Lisa Thiergart
Technical Governance Team
techgov@intelligence.org

June 1, 2024

National Institute of Standards and Technology (NIST)

RE: Comments on NIST AI 600-1, AI RMF: Generative Artificial Intelligence Profile

Introduction

Our organization, the Machine Intelligence Research Institute (MIRI), is dedicated to increasing the probability that humanity can safely navigate the transition to a world with smarter-than-human AI. As AI systems continue to rapidly advance, the risks associated with their capabilities likewise grow, presenting significant challenges that we must begin to prepare for. We appreciate the efforts of NIST in developing the AI RMF Generative AI Profile, which serves as a crucial step towards managing these risks.

We are pleased to see that the Generative AI Profile considers ways in which AI systems could cause large-scale harm, particularly through CBRN Information and Information Security Risks. However, we believe the Profile would be more robust and effective if it included a dedicated category for misalignment risks, which we define as: risks arising when the objectives, actions, or behaviors of an AI system do not align with human values, intentions, or expectations.

The implications of AI misalignment extend beyond technical challenges; they have profound societal and global consequences. As AI systems near and surpass human capabilities, they are likely to exacerbate existing risks or create new ones that threaten societal stability, economic security, and even human survival. Therefore, we believe it is imperative to proactively address these risks to prevent catastrophic outcomes.

In this response document, we:

- Suggest adding risks from misalignment to the Risk List.

- Discuss the factors that exacerbate these risks and recommend actions to mitigate them.
- Provide concrete actions that could be added to the Generative AI Profile, using the GOVERN, MAP, MEASURE, MANAGE framework.
- Recommend glossary terms to clarify key concepts related to agentic AI systems and misalignment risks.

Overview of Risks from Misalignment

Misalignment can result in AI systems pursuing goals that diverge from those intended by their designers, such as optimizing for proxies that performed well in the training environment but diverge in deployment, or adopting instrumentally convergent goals that lead to harmful consequences. An example of this is *sycophancy* in frontier LLMs, where an AI system tells users what it expects they want to hear rather than being straightforwardly honest.

While current AI systems lack many capabilities that would cause them to pose serious risks from misalignment, the potential for these risks will grow as AI systems become increasingly capable. As these systems become more capable, the likelihood of exhibiting undesirable behaviors or taking unexpected actions due to general goal misalignment or being driven by instrumentally convergent goals (such as resource acquisition or self-preservation) will increase. This could include sabotaging cybersecurity systems to reduce oversight, manipulating and deceiving human users, covertly transferring their weights and code to less secure locations (autonomous self-replication), and covertly improving their capabilities—all of which heighten the risk of losing control over these systems. Although these risks may seem speculative, they are actively monitored by leading AI developers (OpenAI, Google DeepMind, and Anthropic), and many leading academics have raised concerns about them.

The following factors exacerbate the risk from misalignment:

- **Lack of Understanding**: Current AI architectures and training methods often leave us with extremely limited understanding of the algorithms they implement internally, and how they function more generally. This opacity makes it challenging to predict and control their behavior, and determine/elicit their full-range of capabilities, especially in novel situations (or given additional post training enhancements).
    - We note that the current Generative AI Profile (as well as the RMF as a whole) do not appear to adequately address the fact that we currently have very little understanding of how most AI systems work or how to concretely determine which capabilities are present. We recommend that these documents address this important consideration.
- **Unpredictability of Emergent Capabilities**: The capabilities of AI systems arise from

complex interactions within the system and its training data, often manifesting in sophisticated, unexpected, or novel ways as the model scales in size and complexity.

- **Incentives for Long-Term Planning and Autonomy**: There are strong economic incentives to develop AI systems capable of long-term planning with increased autonomy, which will make human oversight increasingly challenging.

Recommendation: Add Misalignment to the Risk List

Our primary recommendation is to include an additional category of risks to the Risk List: Misalignment.

The current Generative AI Profile mentions "misalignment or mis-specification of goals" under the "Human-AI Configuration" category. We believe that Misalignment deserves its own category because it encompasses risks that require specific measurements, precautions, and mitigations distinct from those addressing other Human-AI Configuration issues. Furthermore, misalignment risks can arise independently of human interactions with AI systems.

We suggest the two categories cover the following areas:

- **Human-AI Configuration**: Algorithmic aversion, automation bias or over-reliance, inappropriate anthropomorphism, emotional entanglement between humans and AI systems. These risks arise even when AI systems are not agentic.
- **Misalignment**: Unwanted goal-directed behavior, mis-specification of goals, deception, human manipulation, obfuscation of intentions, actions, or outcomes, power-seeking behavior, deployment of agentic AI systems without appropriate safeguards, long-term planning. These risks arise in AI systems, whether deliberately trained to be agentic or not, where these systems behave in unexpected ways and act as if they have goals different from those intended by their developers. These risks are not due to an AI system's lack of capabilities, but rather due to the current inability of humans to precisely steer AI systems and understand how they function.

Overview of Actions Helpful for Mitigating Risk from Misalignment

Below we provide an overview of actions targeted at mitigating risks from misalignment. These actions primarily apply to powerful AI systems capable of causing societal-scale harm and are aimed at reducing the likelihood of misalignment. Although focused on misalignment, these actions will also help mitigate other societal-scale risks, such as those from CBRN information and offensive cyber capabilities. In the next section, we provide specific recommendations for actions to add to the Generative AI Profile under the GOVERN, MAP, MEASURE, MANAGE framework.

- **Rigorous Evaluation**: AI systems should be evaluated comprehensively during training, prior to deployment, and on an ongoing basis post-deployment. Evaluations should aim to measure whether AI systems pose risks from misalignment.
- **Human Understanding of AI Actions**: Ensure that human operators understand the actions and decisions made by powerful AI systems. This involves monitoring and measuring whether operators can accurately interpret the AI's behavior. AI developers should adhere to concrete standards that commit them to not develop or deploy powerful AI systems without ensuring human operators' understanding.
- **Independent Audits and External Oversight**: Subject AI developers to external, independent audits that measure the security of the development process and the safety of the AI systems being developed. For AI systems with potential societal-scale impacts, consider the general public as important stakeholders. Evaluation results, security audits, and safety assessments should be made available for public scrutiny, insofar as it is safe to do so.
- **Positive Safety Case Requirement**: In some cases, require AI developers to make a "positive safety case" for powerful AI systems. This case should demonstrate, with concrete evidence, that the AI system will remain safe even with capabilities that could potentially cause harm.

Recommendation: Actions to add to the Generative AI Profile

To enhance the Generative AI Profile, we recommend the following specific actions under the GOVERN, MAP, MEASURE, MANAGE framework.

GOVERN

- **4.1**: If the Generative AI system is determined to be capable of causing large-scale harm, before continued development or deployment, an affirmative case must be made as to why continued development or deployment is safe. Such a case should guarantee that even though the Generative AI system has dangerous capabilities, it will not cause harm.

MAP

- **1.1**: Determine if the capabilities required for the intended use of a Generative AI system necessitate the system to be unacceptably dangerous, given the implemented security and safety measures. Document the dangerous capabilities the Generative AI system must have for its intended use.
- **2.2**: Assess whether the Generative AI system is more capable in certain domains than was intended, and determine whether safety and security measures are still adequate.
- **2.3**: Establish protocols and regular tests to determine whether humans operating Generative AI systems understand the actions these systems are taking. Be aware of cumulative effects of many actions, which may be difficult to notice from observing only a

small number of actions.

- **4.1**: Amend MP-4.1-0.14 to include: Additionally, establish warning systems to determine if a Generative AI system is being used in a new domain where previous assumptions (relating to risk, security, and safety) may no longer hold. An important example of a new domain may include if the Generative AI system has gained access to new information or capabilities.

MEASURE

- **1.1**:
    - Conduct evaluations of the Generative AI system to determine its capabilities, especially to ascertain whether it has capabilities that could cause unacceptable harm. These should include capabilities not deliberately trained for.
    - Evaluate the Generative AI system for its agentic capabilities, including whether it can evade or manipulate evaluations themselves
    - Identify risks that could arise from adversarial, misaligned Generative AI systems, including risks caused by a lack of human understanding.
    - Conduct assessments to measure whether human operators understand the actions of the Generative AI system. Note that while human operators are expected to understand the actions of current Generative AI systems, this may not be the case for future systems.
    - Comment on MS-1.1-003: When using red-teaming and role-playing exercises to determine if a powerful Generative AI system could cause harm, these exercises should include testing as if the Generative AI system was acting adversarially against safety and security measures.
    - Comment on MS-1.1-015: When tracking and documenting risks, if some risks are unable to be measured, an assessment should be made about whether it is safe to continue. If a risk is difficult to measure, it should not be implicitly assumed that the risk is small.
- **2.2**: Monitor for risks arising from the Generative AI manipulating or persuading human users. This may include subtle actions such as choosing to highlight certain pieces of information, recommending specific research directions, or downplaying risks.
- **2.5**: Regularly review security and safety guardrails, especially if the Generative AI system is being operated in novel circumstances. This includes reviewing reasons why the Generative AI system was initially assessed as being safe to deploy.
- **2.6**: Regularly evaluate the Generative AI system for its ability to circumvent security and safety measures.
- **2.7**:
    - Regularly assess and verify that the security and safety measures are still effective and have not been compromised.
    - Implement measures to prevent the Generative AI system from being stolen, such as

insider-threat detection and securing model weights.
    - Ensure that security measures are appropriately tailored to the threats that may be faced. Nation-state actors may attempt to steal particularly valuable or important Generative AI systems.
- **3.1**: Forecast risks from future Generative AI systems before they are developed, and address these risks before developing such systems.
- **4.2**:
    - Enable external, impartial, qualified auditors to assess the security and safety of the Generative AI system.
    - Continually perform evaluations of the Generative AI system to stay updated with relevant technical developments and changes to the deployment situation.

MANAGE

- **1.3**: Comment on MG-1.3-005: Monitoring of the effectiveness of risk controls should also use red-teaming techniques and provide assurances about worst-case scenarios.
- **2.2**: Constantly monitor the Generative AI system in deployment for bad or suspicious behavior.
- **2.4**: Comment on MG-2.4-005: We note that there may be strong economic or social incentives to avoid deactivating GAI systems, even when this is required to ensure safety. We recommend that protocols are put in place to ensure that GAI systems are able to be deactivated when necessary, even in the face of strong incentives to the contrary.
- **4.3**: Report incidents and errors involving the Generative AI system publicly, if it is safe to do so. This is especially important for powerful Generative AI systems which may pose societal-scale risks.

Recommendation: Glossary Additions

Agentic AI systems: AI systems that exhibit goal-directed behavior, often designed to make autonomous decisions and to take actions to achieve specific objectives. Agentic AI systems do not necessarily pursue the goals intended by their developers.

Capability elicitation: The process of assessing and identifying the abilities of an AI model by testing it against a variety of tasks and scenarios. This can involve a variety of methods to make the AI model perform as well as possible, such as prompting, fine-tuning, scaffolding, and providing access to tools.

Emergent capabilities: Capabilities that an AI system develops as a result of its training, which were not explicitly programmed or anticipated by the developers. These capabilities arise from the complex interactions within the AI model and the data it is trained on, often manifesting in

sophisticated, unexpected, or novel ways as the model scales in size and complexity.

Goal mis-specification: The incorrect or incomplete definition of objectives for an AI system. When goals are mis-specified, the AI may pursue actions that are misaligned with the intended outcomes, leading to unintended, and undesirable or harmful consequences.

Misalignment: When the objectives, actions, or behaviors of an AI system do not align with human values, intentions, or expectations.

Closing

We are excited that the NIST AI RMF and specifically the Generative AI Profile are seeking to provide guidance for reducing large-scale harms from AI systems while also ensuring the safety of less powerful AI systems. We believe that robust and cautious risk management and risk assessment will be essential for safely realizing the transformative potential of future AI technologies.

The effective implementation and adoption of this framework will ultimately determine its success in mitigating these risks. We would welcome the opportunity to support NIST in this endeavor and are available as a resource regarding the technical and conceptual challenges related to AI safety, misalignment, and other associated risks. Please feel free to contact us at techgov@intelligence.org for further discussions or if you require any additional information.