Comments of the Messaging Malware Mobile Anti-Abuse Working Group (M3AAWG) on NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

Consultation reference: NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

The Messaging Malware Mobile Anti-Abuse Working Group (M^{3}AAWG) appreciates the opportunity to submit comments in response to the above-referenced consultation. M^{3}AAWG is a technology-neutral global industry association. As a working body, we focus on operational issues of internet abuse including technology, industry collaboration, and public policy. With more than 200 institutional members worldwide, we bring together stakeholders in the online community in a confidential yet open forum, developing best practices and cooperative approaches for fighting online abuse.

With the growing importance of AI in society and the challenges of AI-related security and abuse issues, appropriate management of AI risk is becoming ever more pertinent.

M^{3}AAWG's comments address nine main areas:

1. Management of systemic risk
2. AI-related attacks and associated risks
3. Interactions of AI with other systems, including other AI systems
4. Organization and prioritization of recommendations
5. Importance of considering both open- and closed-source models
6. AI vendor management
7. Overexposure of training data
8. Use of references in the document
9. Assessment and removal of sensitive data or information in CBRNE-related training datasets.

1. The current version of the document pays relatively little attention to organizational, human, and national security risks when analyzing AI system risks. As these are key elements in understanding the risks associated with deploying AI systems in organizations, more in-depth guidance would be useful. Greater emphasis on end-to-end risk assessment, architecture, and process engineering, as well as threat modeling should be considered. (See also point 3.)

   Specifically, M^{3}AAWG recommends clearer guidance on how to assess technology, systems, and integration before and during the life cycle of AI systems in the context of the organization, and potentially relevant impacts beyond the organization.

a) These concerns should include AI overuse and underuse; i.e., risks emerging from not deploying AI where rationally useful, or over-relying on AI systems where they are not necessary.

b) They should also include AI “under-trust” and AI “over-trust”; i.e., risks emerging from trusting AI too much, or not enough. For example, if an AI system is believed to have greater capability and/or procedural integrity than it actually does, wrong decisions might be rendered that are not caught by human oversight, which is considered to be unnecessary. On the other hand, trusting a system too little could result in resources being expended without tangible benefit.

c) Guidance should be given for measuring the effectiveness and efficiency of AI solutions.

2. The document should address risks related to attacks on AI systems. While we understand that NIST AI 100-2 E2023 focuses on this aspect, we still believe that NIST AI 600-1 should provide detail on risks from and mitigations of adversarial attacks.
    a. For example, a clearer list or reference table of risks related to attacks on AI systems as part of NIST AI 600-1 would be useful for the reader, with references provided to the more in-depth NIST AI 100-2 E2023.
    b. If used in critical infrastructure, misclassification and non-detection can represent large risks of undesired behavior. For example, risks may emerge when AI inputs/outputs are used directly or indirectly to trigger actuators or read sensors. For example if a bad actor has control over one or more sensors they could manipulate the outputs of these sensors to influence the inference of the AI to trigger undesired output that would open or close an actuator. A real-world example of this could be a pipeline carrying oil where the bad actor causes pressure sensors to under-report the pressure in the pipe. The AI takes these readings and infers that the actuators do NOT need to be opened to relieve the pressure.

3. The document should provide more detail on systemic interactions and how to manage them, both with regard to AI-AI system interactions as well as AI interacting with other, non-AI systems.
    a. How should organizations analyze, assess, manage, and disclose the use of AI outputs as AI inputs in downstream systems or downstream actors, whether internal or between organizations? While the document touches on this with third-party plugins and prompt injection, this is a more general concern. As an example, probabilistic inputs (be they accidental or part of an attack) could negatively impact downstream AI actors, especially if that AI agent is directly taking action rather than just providing outputs to human agents.
    b. What security measures must be taken when the output of an AI system becomes an input to itself or another upstream system? An AI system trained on its output can result in not only poor accuracy and misinformation but also reinforced bias and the amplification of attacks such as data poisoning.

   c. How should organizations address the risks of payloads of malware and ransomware hidden within model weights? These risks have been established in literature.
   d. How should organizations address the risk of sensitive interaction history being exfiltrated, or leaked?
   e. How should organizations structure interactions between open-source and closed-source systems and models? As further outlined in point 4, the risk of undesired output resulting from interactions between open-source and closed-source systems and models should be considered in greater detail.

4. Many foundational governance and risk management recommendations contained in the document draft in part overlap with other relevant NIST documents. The large volume of recommendations calls for greater guidance in terms of framing or prioritization. If recommendations were more clearly prioritized and more obviously mapped to various GAI actors, the document would enable organizations to better understand their risk profiles, and select or implement actions.
   a. The summary of Actions to Manage GAI Risks makes clear that not all actions apply to all AI actors, citing the example that not all actions relevant to GAI developers may be relevant to GAI deployers. Yet, the document then provides little guidance regarding the relevant differences.
   b. Many recommended actions are not mapped to the 12 identified "Risks Unique to or Exacerbated by GAI." NIST should reference all items with an identified risk, or explain why an action is not applicable to the mapping.
   c. Extensive provisions in this report may create moderate or substantial new processes, documentation, and reporting requirements for provenance that might not be feasible.

5. The document at this point assumes the use of closed models. We believe that it also needs to address open-source models, as those are likely to become more relevant in the future, partly due to their open-source nature and resulting malleability.
   a. Thus, the RMF should include guidance for risk associated with developing open-source models, and should specifically call out which recommendations are applicable to open-source developers and open-source deployers – and which are not.
   b. The document should specifically address additional risks associated with third-party modifications to open-source models, including:
      - Removal of guardrails
      - Altering of model weights:
         - Altered inference
         - Hidden malware in floating point bits (https://arxiv.org/abs/2107.08590)
      - Fine-tuning for fraud, malware generation, phishing, etc.
   c. Thus, following from point 3, the systematic interplay between open-source and closed models should be considered in greater depth.

    d. Several recommended actions imply responsibility for open-source model developers for outputs created by third parties using their open-source models. The document should clarify to what extent open-source developers and providers should assume responsibility for third-party use, and what controls are necessary to mitigate this risk.

6. The document currently lacks detail on appropriate approaches to AI vendor management and how to frame such a program. We believe this to be extremely relevant, as most organizations will either employ third-party models and systems directly, or build functionality on top of existing systems by existing providers.
    a. As alluded to in point 4.a above, the draft report does not currently distinguish between actions that apply to GAI developers and actions that apply to GAI deployers. Those distinctions are critically important in order to protect the broader AI and AI open-source ecosystem, and to allow organizations to respond to AI risk appropriately.

7. While M^{3}AAWG supports transparency wherever feasible, it might be apposite to consider under which circumstances other approaches or methods might be more efficient and/or effective. Under certain conditions, an overexposure of training data could lead to possible risk and harm. For example, some training data like nuanced regional terms that refer to slurs, violent crimes, etc., may be sensitive. Such training data may then serve as “borderline prompts,” soliciting further prompts that lead to unexpected behavior. Opening such a training set for any actor to use could actually present more harms than benefits.
    a. Numerous references in the document to making data available for inspection or audit is generally appreciated. However, in some cases, similar outcomes may be better accomplished in other ways, like external benchmarking.

8. Many links in the present draft, especially those related to risk profiles in the introductory section, lead to resources that are either paywalled, non-stable, or both. NIST should consider how to make central references easily accessible to readers.

9. For CBRNE-related topics, the draft guidance prioritizes assessment and removal of sensitive data or information in training datasets. In order to do so without unduly reducing a model's capacity to produce scientific outputs, NIST should also provide greater clarity about what specific data must be removed; greater clarity about how to measure these risks; and the definition of thresholds which indicate the type of mitigation necessary and sufficient under varying circumstances. Further, the draft guidance should also consider the development of validated post-training mitigation strategies.

We appreciate the opportunity to submit these comments, and we welcome further opportunities to engage as needed to answer any questions during this process. Please address any inquiries to M^{3}AAWG Executive Director Amy Cadagin at comments@m3aawg.org.

Sincerely,
Amy Cadagin, Executive Director
Messaging Malware Mobile Anti-Abuse Working Group
comments@m3aawg.org
P.O. Box 9125 Brea, CA 92822