Holistic AI’s Response on NIST’s execution of its responsibilities under Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

2 February 2024

National Institute of Standards and Technology (NIST)
100 Bureau Drive
Gaithersburg, MD
20899

RE: Holistic AI’s Response to the Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)

Thank you for the opportunity to provide feedback on this important matter.

1. About Holistic AI

Holistic AI is an AI Governance, Risk and Compliance platform with a mission to empower enterprises to adopt and scale AI with confidence. We are a multidisciplinary team of AI and machine learning engineers, data scientists, ethicists, business psychologists, and legal and policy experts.

We have deep practical experience in auditing AI systems, having assured over 100 enterprise AI projects covering more than 20,000 different algorithms. Our clients and partners include Fortune 500 corporations, SMEs, governments, and regulators. We work with several companies to conduct independent AI Audits and offer a proprietary Software as a Service platform for AI Governance, Risk Management, and Regulatory Compliance.

We welcome the RFI on NIST’s mandate pursuant to the Biden-Harris Executive Order on AI, and are dedicated to assisting the Institute in achieving its objectives of gaining deeper insights into the ever-evolving AI Governance ecosystem and providing evidence-based insights to influence better policy outcomes ahead.

2. Key comments

Our Comments take the form of a series of considerations we believe the NIST may benefit as it actions its mandate on the Executive Order. As the NIST develops a companion resource to the NIST AI Risk Management Framework (AI RMF) for generative AI (NIST AI 100-1), it should ensure the framework’s development is cognisant of the many risks and harms emanating from the deployment of these models, but also how they manifest in different model modalities.

Furthermore, the AI governance discourse contains many nuances that current methods of model safety evaluation, such as benchmarking, red teaming, and model audits, do not fully capture. We believe that the NIST is well-positioned to address these nuances, and provide recommendations accordingly.

Finally, we believe that the NIST AI 100-1 can serve as a potent vehicle to guide the responsible development, deployment and access of generative models, and to that end provide insights from academic literature, operational recommendations and concluding remarks.

As such, our feedback focuses on providing information on Model Risks, Model Modalities, Modes of Model Safety Operationalisation, Benchmarking, Red-Teaming and finally, enabling the responsible development and deployment of generative models through the NIST AI 100-1.

2a. Academic and applied research on Model Risks of LLMs should be leveraged when developing NIST AI 100-1

Literature on model risks has steadily grown in recent months and years. However, the articulations of said risks are themselves dynamic, with new vectors of harm being discovered as a result of a concerted research impetus and increasing model deployment. Given this rapidly changing landscape, we find it beneficial to categorize emerging species of generative model-generated risks and harms based on taxonomies grounded in academic literature. Notably, we rely on the following:

a. A typology derived from Koshiyama et al. (2021) where algorithmic risks are categorized into five verticals, namely: *Robustness, Bias, Privacy, Explainability and Efficacy*. These risks are contextualized to generative models below:
    i. Robustness: Risks of models being susceptible to adversarial attacks, such as using particular techniques to reveal training data
    ii. Bias: Risks of the model generated biased outputs due to improper training data (training bias), inappropriate application context (transfer-context bias) and inadequate inferential capabilities (inference bias)
    iii. Privacy: Risks associated with models leaking sensitive information or personal data
    iv. Explainability: Risks of models generating arbitrary decisions, with its outputs not understandable to developers, deployers and users
    v. Efficacy: Risk of models underperforming relative to their use-cases, or not meeting safety and performance baselines, such as generating incorrect or misleading outputs

b. A taxonomy of risks posed by Large Language Models (LLMs) (Weidinger et al. 2022), listing the following:
    i. Discrimination: Risks of an LLM perpetuating societal stereotypes and biases, resulting in unfair discrimination and exclusion
    ii. Information Hazards: Risks of an LLM compromising user privacy by leaking sensitive and personal information
    iii. Misinformation Hazards: Risks of LLMs generating misleading and inaccurate information, used to proliferate mis/disinformation across text, video, and audio capabilities at a large scale
    iv. Malicious Use: Risk of an LLM being co-opted by Bad Actors (e.g. Violent Extremism/ Terrorist organisations to harmful propaganda)
    v. Human-Computer Interaction (HCI) Harms: Risk of users over-relying on LLM capabilities and using them in unsafe ways (e.g., assisted suicide)
    vi. Automation and Environmental Harms: Negative externalities associated with the use of LLMs, particularly due to their high environmental footprint

As the AI 100-1 is developed, we urge NIST to leverage such risk typologies to inform its approach. However, nuances depending on the context of deployment must also be taken into account. For example, operationalizing bias mitigation in the medical context (where individual differences based on protected characteristics are pivotal to providing effective and appropriate diagnoses and treatments) may run counter to bias mitigation strategies in non-medical fields, such as in the case of employment and job screening algorithms where the objective is largely to remove the influence of protected characteristics.

Further, when operationalizing algorithmic risk management and mitigation strategies, trade-offs between risk parameters may occur as, indeed, these are not mutually exclusive – they overlap and interact. Koshiyama et al. (2021) provide a helpful overview of some of these interactions, with imperatives on algorithmic explainability often conflicting with actions taken to preserve the model's privacy. However, different risk parameters may be mutually reinforcing, such as instances wherein improving a system's interpretability leads to enlarged apertures for bias mitigation.

Finally, risks can also be intersectional – with harmful model outputs often interweaving representational harms (based on protected characteristics like gender and ethnicity) with instances of toxicity and hallucinations. Erstwhile harm mitigation interventions, like model evaluations and red-teaming measures, do not adequately capture these – making proactive efforts to reflect these complexities imperative.

Considering these nuances, we believe it important for NIST to develop a repository of these risks with extensive stakeholder consultation from academia, civil society, industry and government. This repository should, in turn, aid in the development of a common risk typology that can be adapted to various use-cases and contexts, that is reflective of the many interplays between risk parameters.

2b. Nuances in Model Modalities and the state of existing model safety operationalization must be considered:

In addition to the many risk species that may emanate from the usage of generative models, it is important to critically assess how these may manifest in different model modalities. For example, instances of violent extremism may produce a greater "shock-value" in image, audio and video modalities as opposed to textual interfaces. On a related note, contextuality and temporality in multimodal contexts become crucial determinants in gauging the level of (potential) model-generated harm; an innocuous synthetic video of an army training exercise with a provocative audio clip issuing a clarion call to take up arms may be particularly misconstrued as harmful misinformation during times of conflict, when the prospect of war is more salient.

These challenges are further exacerbated due to the fact that there currently does not exist an adequate number of safety interventions for non-text model modalities. This is representative in the coverage of model risks as well, with recent research from Google DeepMind finding the number of safety interventions decreasing progressively for image, audio and video modalities – with almost no evaluations for risks of representational harms, privacy hazards and malicious use for audio, multimodal and video. Indeed, this is symptomatic of a larger problem – a relative lack of research on and investment in multimodal AI safety.

We note that NIST is establishing the US AI Safety Institute (AISI) to further its mandate on Generative AI Safety delineated by the Executive Order. As such, we recommend that proposed dedicated Working Groups (WGs) on generative AI risk management (WG#1) and capability evaluations (WG#3) be leveraged to investigate these issues and prioritize the development of such safety interventions.

2c. Existing academic and industry efforts on benchmarking should be furthered

As mentioned above, there are perceptible gaps in the state of capability benchmarks today. Not only do they towards certain model risks (representation and toxicity harms) and text modalities, but popular aggregate benchmarks may also be prone to memorisation, or instances where a model is inadvertently trained on the very dataset, reducing accuracy. As evidenced by Rauh et al. (2021), there exists a perceptible gap between current benchmarks and their efficacy in effectively capturing emerging risk species, and over-relying on them may lead to ignoring harmful model capabilities that may lay undetected.

That said, benchmarking remains a pivotal component of safety evaluations, and we recommend that NIST support pre-existing work in this arena by organizations like MLCommons and the Alignment Research Centre (ARC)/ Model Evaluation and Threat Research (METR), as well as academic contribution in this field, and convene rich stakeholder expertise to develop fit-for-purpose open-source benchmarks that models can be assessed against.

While there are a number of ways that NIST might further such work, such as through funding or research collaborations, one avenue that we propose is developing a comparable 'safety benchmark leaderboard' for the developers of generative models akin to the Open LLM leaderboard developed by HuggingFace. To support this, the inclusion and exclusion criteria for appropriate benchmarks for such a leaderboard should be determined after extensive stakeholder consultation with experts from multiple domains, and we recommend that appropriate WGs from the US AISI be convened for this purpose.

Moreover, as NIST develops its guidance on benchmarking, it should ensure that the following gaps are actively and iteratively addressed:

1. Gaps in benchmarking for particular model harms
2. Gaps in benchmarking for different modalities
3. Gaps in current benchmarks in assessing intersectional model harms
4. Gaps in determining *who* (developers, deployers, independent auditors, regulatory authorities, etc) should be responsible for which type of model evaluation

2d. Red Teaming should be standardised and systematised

Red-teaming or adversarial testing is an effective evaluation mechanism to help discover the unknown risks and vulnerabilities associated with a generative model. Current red-teaming and adversarial testing exercises are however largely an art rather than a science, signalling the need to systematise this intervention. This would involve establishing industry-wide standardised protocols that guide the entire red teaming process, ensuring consistency and thoroughness in assessments. Moreover, recognising the importance of diversity, and a participatory approach are paramount, and efforts should be made to embed a diverse range of perspectives and expertise in the red teaming methodology. For example, a red-teaming exercise for a systemic model risk like misinformation should seek to gather the pooled collective intelligence of experts from the medical and public health domain, journalism, media studies, sociology, behavioural science, computer science, machine learning, safety policy, hate-speech, among others. This not only fosters inclusivity but also enriches the evaluation process by considering a broader spectrum of potential risks and vulnerabilities.

Red teaming exercises should also incorporate a diversity of technical exercises like curated prompt attacks, training data extraction or extractive memorization, backdooring the model, adversarial prompting, data poisoning and exfiltration mechanisms, among others. Given the resource-intensity of red teaming, approaches such as that used by Perez et al. (2022) that seek to deploy language models to stress-test target models through adversarial prompt generation, curation and engineering, and modules like Anthropic's Constitutional AI (Bai et al. 2022) should also be leveraged. Striking the right balance between these human and automated approaches ensures comprehensive evaluations that reflect the multifaceted nature of the challenges associated with these exercises and may provide a more scalable solution than human evaluations alone.

We also consider it important that red-teaming practices of erstwhile large model providers be subject to comprehensive scrutiny. As such, we urge NIST to embed transparency and confidence-building measures such as mandating external researcher access, governance and process audits of red-teaming exercises, as well transparency reports that can publicly communicate the outcomes of the same in an accessible manner.

Finally, to enhance the efficacy of this emerging mechanism, we recommend that NIST establishes key exemplary outcomes that could be expected from a successful red-teaming exercise. These could include insights that can be used to generate model-level mitigations (like prompt refusals for violative content) and system-level mitigations (such as block-lists of certain keywords), as well as guidance on the development of effective policies, such as content policies for API developers, and the development of new evaluations for emerging and unknown risks.

In essence, a comprehensive approach to red teaming involves not only the meticulous execution of assessments, but also the integration of diverse perspectives, technical and human-centric interventions, and transparent communication to fortify organizational security measures.

2e. NIST AI 100-1 should provide guidance on the responsible development and release of generative models

In addition to effectively capturing the risk profiles associated with various generative AI models, we believe that NIST AI 100-1 is well-placed to provide comprehensive guidance on the extent of release and structured access of foundation models, in a way that is benchmarked to safety and ethics calibrations during model development.

To this end, we refer NIST to work by Solaiman (2023), which proposes a gradient of access that ranges from fully closed, followed by gradual/staged release, hosted access, cloud-based/API Access, downloadable models, to finally fully open models. Indeed, the more open a model is, the higher its auditability and feasibility to conduct research which can help improve safety outcomes. However, a trade-off is seen here, with progressive openness resulting in fewer outcomes for risk control and management – signalling the need for model development lifecycles to incorporate bespoke evaluations and auditing mechanisms to calibrate the model's safety levels.

A similar idea has been put forth by Shevlane et al. (2023), who posit that developers should gradually release a model's exposure to the external world as they accumulate evidence about a model's safety and extent of alignment through model evaluations, scaling external research access and audits. Other approaches include mandating Responsible Scaling Policies (RSPs), that help provide thresholds of acceptable deployment linked to a developer's current protective measures, as well as indicating zones of dangerous capability where model development should be paused until appropriate protective measures are improved and productionised.

Essentially, the development and deployment scope of powerful foundation models should be intricately linked to their safety calibrations, and the NIST AI 100-1 is well-positioned to provide guidance in this regard. We believe that leveraging the NIST AI 100-1 can serve as a valuable tool for guiding responsible development and deployment while also contributing to grounding the open-closed source AI debate on safety calibration, evaluations, audits and risk management.

3. Resources by Holistic AI

In lieu of the fact that identification, measurement, and mitigation of generative AI risks is relatively new, below we link some resources and references from our academic research.

- **Holistic AI Open Source**

- Towards Auditing Large Language Models: Improving Text-based Stereotype Detection
- LLM Auditing Guide
- Towards Algorithm Auditing
- AI Assurance Processes

4. Concluding statement

Holistic AI welcomes the opportunity to provide comments on this important matter. We appreciate the open and collaborative approach taken by NIST.

We support the important objectives of the Executive Order and the work that NIST is doing in the field of AI risk management. We stand ready to support NIST and other public authorities or agencies involved in the development of generative AI frameworks and resources.

Please contact publicpolicy@holisticai.com for any further information or follow-up on this submission.

Sincerely,
Holistic AI Inc.
publicpolicy@holisticai.com