FAS FEDERATION OF AMERICAN SCIENTISTS

February 2, 2024

National Institute of Standards and Technology
100 Bureau Drive
M/S 1070, Public Affairs Office
Gaithersburg, MD 20899

RE: 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); NIST-2023-0009-0001

To whom it may concern,

The Federation of American Scientists (FAS) is a catalytic, non-partisan, and nonprofit organization committed to using science and technology to benefit humanity by delivering on the promise of equitable and impactful policy. FAS believes that society benefits from a federal government that harnesses science, technology, and innovation to meet ambitious policy goals and deliver impact to the public. Today we are writing as the Emerging Technologies and National Security team at FAS to provide an opinion on the National Institute of Standards & Technology (NIST)'s assignments regarding generative AI risk management, AI evaluation, and red-teaming.

Specifically, our comments today will be in response to section 1(a)(2) of the Request for Information, concerning "Availability of, gap analysis of, and proposals for metrics, benchmarks, protocols, and methods for measuring AI systems' functionality, capabilities, limitations, safety, security, privacy, effectiveness, suitability, equity, and trustworthiness." The Request for Information highlights several potential risks and impacts which our comments will address, including: "Negative effects of system interaction and tool use…chemical, biological, radiological, and nuclear (CBRN) risks…[e]nhancing or otherwise affecting malign cyber actors' capabilities…[and i]mpacts to individuals and society."

## Summary

A new class of risk mitigation policies has recently come into vogue for frontier AI developers. Known alternately as Responsible Scaling Policies or Preparedness Frameworks, these policies outline commitments to risk mitigations that developers of the most advanced AI models will implement as their models display increasingly risky capabilities.[1] While the idea for these policies is less than six months old, already two of the most advanced AI developers, Anthropic

[1] METR. (2023, September 26). *Responsible Scaling Policies (RSPs)*. https://metr.org/blog/2023-09-26-rsp/

and OpenAI, have published initial versions of these policies. The U.K. AI Safety Institute asked frontier AI developers about their “Responsible Capability Scaling” policies ahead of the November 2023 UK AI Safety Summit.[2]

This comment will provide some background on this class of risk mitigation policies (we use the term Preparedness Framework, which captures the key idea of preparing for catastrophic risk, rather than Responsible Scaling Policy). We outline criteria for robust Preparedness Frameworks (PFs) and evaluate two key documents, Anthropic’s Responsible Scaling Policy[3] and OpenAI’s Preparedness Framework[4], against these criteria. We argue that these policies are net-positive and should be encouraged. At the same time, we identify shortcomings of current PFs, chiefly that they are underspecified and insufficiently conservative. Improvement in the state of the art of risk evaluation for frontier AI models is a prerequisite for a meaningfully binding PF. Most importantly, PFs, as unilateral commitments by private actors, cannot replace good policy.

**FAS Impact To Date**

The Federation of American Scientists (FAS) has a track record in AI standards and governance, and we have worked closely with NIST on related topics. We have advocated for a federal AI testbed that “will help AI researchers and developers better understand how to construct testing methods and ultimately build safer, more reliable AI models.”[5] In a public comment to the Office of Science and Technology Policy, we recommended a number of relevant policy measures, including developing a pre-deployment risk assessment protocol for frontier AI models.[6] In another public comment to the Ranking Member of the Senate Committee on Health, Education, Labor, & Pensions, we explained how industry leaders are starting to adopt the NIST AI Risk Management Framework and encouraged its continued uptake.[7] And we are excited to participate in the forthcoming U.S. AI Safety Institute Research Consortium.

FAS has worked on the NIST AI Risk Management Framework in a number of ways. We participated in the NIST GAI Public Working Group, providing feedback on the development of a

[2] UK Government. (2023). *Company Policies*. AI Safety Summit 2023. https://www.aisafetysummit.gov.uk/policy-updates/#company-policies

[3] Anthropic. (2023). *Anthropic’s Responsible Scaling Policy, Version 1.0*. https://www-cdn.anthropic.com/files/4zrzovbb/website/1adf000c8f675958c2ee23805d91aaade1cd4613.pdf

[4] OpenAI. (2023, December 18). *Preparedness Framework (Beta)*. https://cdn.openai.com/openai-preparedness-framework-beta.pdf

[5] Huang, T. (2022, January 19). *Creating an AI testbed for Government.* Federation of American Scientists. https://fas.org/publication/creating-an-ai-testbed-for-government/

[6] Kaushik, D., Titus, J., & Alexander, L. (2023, July 12). *Six Policy Ideas For The National AI Strategy.* Federation of American Scientists. https://fas.org/publication/six-ideas-for-national-ai-strategy/

[7] Titus, J. (2023, October 5). *AI In Action: Recommendations For AI Policy In Health, Education, And Labor*. Federation of American Scientists. https://fas.org/publication/ai-in-action-help/

generative AI profile.[8] We evaluated OpenAI's GPT-4 against the RMF and provided recommendations for better risk management practices.[9] And we co-organized a workshop in Washington, DC titled "Operationalizing the Measure Function of the NIST AI RMF."[10] These are just a few examples of the impact FAS has had thus far in the realm of AI policy.

**Motivation for Preparedness Frameworks**

As AI labs develop potentially dual-use foundation models[11] with capability, compute, and efficiency improvements, novel risks are likely to emerge, some of them potentially catastrophic.[12] Today's foundation models can already cause harm and pose some risks, especially as they are more broadly used.[13] Advanced, computationally intensive models at times display unpredictable behaviors.[14]

To this point, these harms have not risen to the level of posing catastrophic risks. The capabilities of models at the current state of the art simply do not imply levels of catastrophic risk above current non-AI related margins.[15] While it's true that the scaling of foundation models in terms of compute has historically been linked to improvements in their capabilities[16], it's important to recognize that the relationship between the number of model parameters and their capabilities is not fully understood. The field of machine learning is rapidly evolving, and many researchers acknowledge the possibility of significant advancements that could disrupt current

[8] NIST AIRC - NIST AI Public Working Groups. https://airc.nist.gov/generative_ai_wg

[9] Kaushik, D. & Alexander, L. (2023, May 11). *How Do OpenAI's Efforts To Make GPT-4 "Safer" Stack Up Against The NIST AI Risk Management Framework?* Federation of American Scientists. https://fas.org/publication/how-do-openais-efforts-to-make-gpt-4-safer-stack-up-against-the-nist-ai-risk-management-framework/

[10] https://casmi.northwestern.edu/news/articles/2023/workshop-to-explore-sociotechnical-standards-to-better-manage-ai-risks.html

[11] As defined by the AI EO: Exec. Order No. 14110, 88 Fed. Reg. 75191 (Oct 30, 2023). https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[12] Hendrycks, D., Mazeika, M., & Woodside, T. (2023). *An Overview of Catastrophic AI Risks*. https://arxiv.org/abs/2306.12001

[13] *Artificial Intelligence Incident Database*. (n.d.). Retrieved January 29, 2024, from https://incidentdatabase.ai/

[14] See for example, Roose, K. (2023, February 16). Why a conversation with Bing's chatbot left me deeply unsettled. *The New York Times*. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

[15] For further consideration of current frontier AI and catastrophic risks, one recent study, conducted at RAND, found that "biological weapon attack planning currently lies beyond the capability frontier of LLMs as assistive tools." Mouton, C., Lucas, C., & Guest, E. (2024). *The Operational Risks of AI in Large-Scale Biological Attacks*. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2977-2.html

[16] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020, January 23). Scaling laws for neural language models. https://arxiv.org/abs/2001.08361

trends. For instance, it's conceivable that future technical breakthroughs might lead to models that are equally effective with significantly fewer parameters.

Given the dynamic nature of AI development, it's prudent for policy-making to concentrate on the broader outcomes and characteristics that are most relevant, rather than fixating on specific technical aspects such as the number of parameters or the exact type of neural architecture used. This approach allows for more flexibility and adaptability in responding to the fast-paced changes and unforeseen advancements in the AI field.

As capabilities increase, risks are likely to increase, and new risks are likely to appear. Executive Order 14110 (the Executive Order on Artificial Intelligence, or the “AI EO”) detailed some novel risks of potentially dual-use foundation models, including risks associated with easing development of chemical, biological, radiological, or nuclear (CBRN) weapons and advanced cybersecurity risks. Other risks are more speculative, such as risks of model autonomy, loss of control of AI systems, or negative impacts on users including risks of persuasion.[17]

Developers must proactively manage the risks associated with advancements in AI capabilities. This involves implementing policies that address potential risks iteratively. Preparedness frameworks are crucial in this context. These frameworks evaluate risk levels in various categories and propose mitigation strategies. According to OpenAI's Preparedness Framework, their aim is to enhance our understanding of catastrophic risks and establish processes to ensure safe development. In the absence of such frameworks, the tendency to prioritize speed over safety concerns might prevail. While the exact consequences of failing to mitigate these risks are uncertain, they could potentially be significant.

### Weaknesses of Preparedness Frameworks

Preparedness frameworks (PFs) have limitations in addressing the risks of AI. They are voluntary and depend on developers' commitment to their principles. For example, the criteria for risk levels in Anthropic's RSP are vaguely defined, creating ambiguity and potential for misuse. Additionally, not all developers share the same level of concern about catastrophic risks.

[17] Deepfakes and disinformation can be considered an early “persuasion” risk, as they potentially cause humans who interact with model outputs to change their beliefs based on false pretenses. At the current frontier, the AI models themselves are not agentically using these outputs to attempt to persuade; rather, they are being misused by human prompters. Conceptually, though, some researchers believe it is possible that advanced AI models will develop agentic, goal-oriented behaviors that might include seeking to persuade human users for their own ends.

OpenAI's PF and Anthropic's commitments show a serious approach to these risks, but gaps remain between the spirit and the letter of these frameworks. More specific risk thresholds and transparency could help bridge this gap.

PFs, being unilateral, don't address broader societal risks effectively. They're similar to individual actions against climate change: significant in isolation but limited in addressing aggregate risks. The integration of AI into the economy and other systems could pose unforeseen structural risks.

Furthermore, PFs might reduce the urgency for government intervention. By appearing safety-conscious, developers could diminish the perceived need for regulatory measures. Policymakers might over-rely on self-regulation by AI developers, potentially compromising public interest for private gains.

Policy intervention is essential to fill the void left by PFs. Policy, aligned with public interest and enforceable, can address risks more comprehensively, especially structural risks associated with AI. In general, while PFs contribute to holding developers accountable, they are insufficient on their own. Government action is necessary to mitigate the broader, potentially catastrophic risks of advanced AI systems.

**Suggested Criteria for a Robust Preparedness Framework**

These criteria are adapted from the ARC Evals blog post, Anthropic’s RSP, and OpenAI’s PF. Broadly, they are aspirational, as no existing PF meets all or most of these criteria. To the extent possible, these criteria follow from the level of potential risk of advanced AI capabilities and the nature of PFs as voluntary commitments.

For each criterion, we will explain the key considerations for developers adopting PFs. We will then analyze OpenAI’s PF and Anthropic’s RSP to illustrate the strengths and shortcomings of their approaches. Again, these policies are net-positive and should be encouraged. They demonstrate costly unilateral commitments to measuring and addressing catastrophic risk from their models; they meaningfully improve on the status quo. At the same time, these initial PFs are underspecified and insufficiently conservative. Improvement in the state of the art of risk evaluation and mitigation, and subsequent updates to these PFs, would make them more robust.

- Robust preparedness frameworks should cover the **breadth of potential catastrophic risks** of developing frontier AI models. These risks may include, among others, CBRN risks, societal harms, and other potentials for misuse, including cybersecurity and critical infrastructure. These frameworks should be updated when new risk vectors emerge.

- Robust preparedness frameworks should **define the developer's acceptable risk level** ("risk appetite") in terms of likelihood and severity of risk in accordance with the NIST AI Risk Management Framework, section Map 1.5.[18,19] It is important to note that two of the industry leaders, viz. OpenAI and Anthropic, have not publicly declared their risk appetite. Given this is a nascent field of research, NIST and other standard-setting bodies will be crucial in developing AI risk metrology. For now, PFs should state developers' risk appetites as clearly as possible, and update them regularly with research advances.

- Effective preparedness frameworks should **clearly define capability levels and risk thresholds**. These thresholds need to be quantitatively robust to ensure developer accountability. While the exact likelihood and quantification of risks from frontier AI models are not well-understood, both OpenAI and Anthropic have attempted to outline qualitative risk categories. For example, OpenAI's High risk threshold in the CBRN category includes scenarios where models significantly aid in creating threats. However, terms like "meaningfully improved assistance" are vague and require more precise definition. As the science of AI risk measurement evolves, these frameworks should also improve in detailing risk thresholds. Over-reliance on quantification can lead to a checkbox approach, which might quickly become obsolete, underscoring the importance of judgment in assessing risks.

- Preparedness frameworks must **include detailed evaluation procedures for AI models, ensuring comprehensive risk assessment** within a developer's tolerance. For instance, Anthropic's RSP rigorously details model autonomy evaluations but is less specific about misuse risks, acknowledging the challenges and uncertainties in this area. OpenAI's PF provides a 'Model Scorecard' for hypothetical evaluations across risk categories, though it's not exhaustive. Measurement science expertise at institutions like NIST will be pivotal in developing robust evaluation methods.

- Effective preparedness frameworks should **clearly define risk thresholds and associate them with specific risk mitigations**. OpenAI's PF and Anthropic's RSP commit to various mitigations according to risk levels. For example, Anthropic's ASL-2

[18] Tabassi, E. (2023), Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST Trustworthy and Responsible AI, National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.AI.100-1, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225.

[19] The Berkeley Center for Long-Term Cybersecurity's foundation model profile includes resources for frontier AI developers as they seek to define their risk appetite. Barrett, A., Newman, J., Nonnecke, B., Hendrycks, D., Murphy, E., & Jackson, K. (2023). *AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models, Version 1.0*. UC Berkeley Center for Long-Term Cybersecurity. https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf

models, such as Claude, have certain safeguards like model cards and use policies, while higher-risk models require more stringent measures like restricted access to training details and enhanced security. Mitigation strategies differ for development versus deployment phases, reflecting the varying risks of internal possession versus external interaction with models. These strategies may include development restrictions, cybersecurity enhancements, and model interaction controls. Regular updates and some flexibility in judgment are crucial as the technology evolves.

- Risk mitigations in Preparedness Frameworks (PFs) **need to effectively address significant risks within a developer's risk appetite**. The challenge lies in confidently ensuring that these mitigations are adequate. For instance, OpenAI's approach to risk mitigation includes a commitment to reduce 'critical' pre-mitigation risk to at most 'high' level post-mitigation. However, their criteria, such as a model's ability to self-exfiltrate, suggest that more stringent development restrictions may be necessary. Similarly, Anthropic's risk mitigations for high-risk models focus on preventing model weight theft by non-state actors and making it costly for state actors. A more cautious approach would aim to secure models against all potential threats laid out by their developers.

- Effective preparedness frameworks should **combine credible risk mitigation commitments with governance structures** that ensure these commitments are fulfilled. While Anthropic's RSP includes proactive planning and transparency measures, its "White Knight" clause, allowing for loosened restrictions in extreme emergencies, could undermine its commitment's credibility. OpenAI's PF outlines a structured operation with dedicated teams and advisory groups for risk mitigation, enhancing its credibility. Balancing risk management with other incentives like profit is crucial in these frameworks.

- Preparedness frameworks like OpenAI’s PF and Anthropic’s RSP must **include a mechanism for regular updates** in response to ongoing research and technological advances in AI. Both frameworks acknowledge this need: Anthropic’s RSP is labeled “Version 1.0” and includes a board-approved "Update Process," while OpenAI’s PF, marked as “(Beta),” commits to regular updates of its model "Scorecards" and evaluations. These updates are crucial for adapting to new safety insights and unforeseen AI capabilities.

- For models with risk above the lowest level, **both pre- and post-mitigation evaluation results and methods should be public, including any performed mitigations**. Transparency in publishing these evaluations is vital for accountability. However, sensitivity is needed regarding the extent of disclosed information to prevent misuse by malicious actors. Anthropic's RSP commits to sharing results publicly where feasible, sometimes with delays for safety. OpenAI's PF, in contrast, does not commit to publishing its Model Scorecards.

**NIST Potential for Involvement**

As these preparedness frameworks are voluntary commitments, they are not themselves good candidates for public policy or regulatory adaptation (and of course NIST is not itself a regulatory body). However, there are a number of potential overlaps between developers' PFs and NIST's work on AI safety and red-teaming and NIST should be apprised of the existence, adoption, and relative strengths and weaknesses of companies' PFs.

These documents can potentially be seen as developers adopting the NIST AI Risk Management Framework. While Anthropic and OpenAI have not explicitly connected their PFs to the AI RMF, their risk management approaches overlap significantly. Further research, and working more closely with these developers, could help NIST understand the extent to which these PFs represent adoption of the AI RMF, and how they differ. There are also likely close connections, and key differences, between these PFs and the GPAIS and Foundation Model Profile cited above; NIST could evaluate these different risk management approaches and update or publish guidance accordingly.

Another reason for NIST to be apprised of PFs is that they connect very closely with NIST's requirements under section 4.1 of the AI EO. NIST is responsible for developing red-teaming guidance for developers of potentially dual-use foundation models. This will presumably encompass key elements of preparedness frameworks, especially capability evaluation, risk thresholds, and risk mitigations. These companies might even participate in the U.S. AI Safety Institute Research Consortium, using their expertise to assist the development of NIST's red-teaming guidance. This comment seeks to add context to these companies' PFs to provide another perspective and hopefully assist NIST in developing this crucial guidance.

**Conclusion**

In conclusion, preparedness frameworks represent a promising approach for AI developers to voluntarily commit to robust risk management practices. However, these policies are not panaceas. Preparedness frameworks have weaknesses, including potential gaps between spirit and text, and their nature as private commitments rather than public policy.

For preparedness frameworks to meaningfully contribute to risk mitigation, they must meet certain criteria of robustness. These include: covering a broad range of catastrophic risks, defining acceptable risk levels, detailing risk thresholds, outlining comprehensive evaluation procedures, specifying pre-determined risk mitigations, and including credible commitments and governance structures.

Current preparedness frameworks from Anthropic and OpenAI represent important first steps, but have shortcomings in specificity, evaluation methods, risk mitigations, among others. As the science of AI risk assessment advances, these frameworks should be recursively updated to more robustly address catastrophic risks. Ultimately, policy and regulation will also be crucial complements to further encourage and enforce responsible development of potentially high-risk AI systems.

We appreciate the opportunity to provide input on this important topic. FAS stands ready to continue assisting NIST in developing guidance for responsible AI development, including preparedness frameworks, model evaluations, and red-teaming. We believe close collaboration between government, industry, and civil society is key to ensuring AI's benefits are realized while catastrophic risks are avoided.

If you have any questions, please feel free to reach out to Divyansh Kaushik at dkaushik@fas.org.

Best wishes,

Divyansh Kaushik, Ph.D.
Associate Director for Emerging Technologies
and National Security
Federation of American Scientists

Jack Titus
AI Policy Fellow
Federation of American Scientists

Liam Alexander
Policy Associate
Federation of American Scientists