RECOMMENDATIONS: Generative AI Away from the Frontier

The National Artificial Intelligence Advisory Committee (NAIAC)

October 2023

RECOMMENDATIONS

Recommendation 1:
Encourage voluntary risk assessments of generative AI systems with more constrained access.

For systems with more constrained access, the Biden-Harris administration should encourage companies to extend voluntary commitments according to a risk-based assessment to include off-frontier generative AI systems, particularly with regards to independent testing, risk identification, and information sharing about risks.

Recommendation 2:
Collaborate with diverse stakeholders to conduct risk assessments of generative AI systems with unconstrained access.

For generative systems with unconstrained access (including open-source systems), NIST should work collaboratively with a wide range of stakeholders, including academia, civil society, advocacy organizations, and industry (where legally and technically feasible). These stakeholders should develop test and analysis environments (including sandboxes or other testing-specific environments), measurement systems, tools for testing generative AI systems, and appropriate methodologies to determine critical potential risks of these systems.[1]

CONTEXT

Generative AI systems have captured the public's attention, both for their novel capabilities and potential risks. Much of the focus for industry, government, and the public has been on "frontier" systems — that is, systems that are at the cutting-edge of hardware, software, and data.[2] While this focus is understandable, we ought not ignore off-frontier generative AI systems.[3] NAIAC believes that it is important to have a clear understanding of the potential risks posed by these more widely available

[1] We deliberately do not prescribe a specific scope for these analyses, as we believe it is critical for them to be able to flexibly adapt to changing risks to individuals, communities, and the nation. However, the scope should be established partly through public engagement and consultation.
[2] There is no clear line that can be used to distinguish "frontier" from "non-frontier" models, particularly since different organizations are building cutting-edge generative AI systems with different goals in mind (e.g., maximizing training data vs. minimizing model size). Nonetheless, we use this term to capture the idea of the "latest and greatest" models and systems.
[3] For example, the voluntary commitments from several technology companies are restricted to "generative models that are overall more powerful than the current industry frontier."

generative AI systems. In particular, these systems (whether proprietary or open-source) potentially pose at least three different types of risks when released in relatively uncontrolled settings, though we emphasize that the actual scope and magnitude of these risks are currently largely unknown:

- Individuals or organizations acquiring potentially harmful information (e.g., techniques to develop chemical weapons or develop a novel computer virus)

- Individuals or organizations acquiring private information (e.g., learning personal information such as a private address, or exposing corporate trade secrets)

- Rapid generation of potentially harmful content, driven by either deliberate human intent (e.g., mis- or disinformation) or shortcomings of the system (e.g., pervasive "hallucination" or misleading interface)

Each of these risks could have significant impacts from the individual up through national security concerns. (We do not take a position on the likelihood of the AI system itself posing a threat to humanity, rather than the AI system enabling humans to pose such threats to ourselves.) However, the exact nature, scope, and possibility of these risks from off-frontier systems remains largely a "known unknown" for policy-makers and others outside of a few select companies. Moreover, investigations into these risks cannot be merely technical, but must also bring insights from social sciences, behavioral sciences, ethics, and more disciplines. An understanding of the scope, scale, and likelihood of these risks is critical to support decisions about where regulation or other forms of governance might be needed.

The current limits on our understanding and knowledge of risks and benefits is particularly concerning for those off-frontier generative AI systems that are widely available, or available without constraints or oversight. One such group of systems — both on- and off-frontier — are those that have been released as open-source (e.g., LLaMa-2, Alpaca, HuggingChat, GPT-NeoX-20B). We emphasize that discussion of the potential risks from widespread or malicious uses of open-source systems should be balanced against the benefits of such systems. The democratization of access through open-source generative AI systems, or through increasingly open API access to (proprietary) previous-generation systems, holds the potential of significant positive impact, including spurring innovation and increasing creative expression. Individuals and small companies who cannot afford to build their own generative AI systems could especially benefit. Moreover, open-source systems have historically

been far more transparent, and thus often better understood, than proprietary systems, though open-source generative AI may be an exception to that trend.[4]

The challenges to understanding the risks of off-frontier systems are different if one has unconstrained access (including open-source) vs. more controlled access (for systems retained by a company). In the latter case, the company itself may have an understanding of the potential risks from their generative AI systems, whether through internal red-teaming, external bias audits, and other analyses. However, the results of those efforts have been largely held within each company, and so relevant information is often unavailable to outside actors, including policy-makers.

For systems where one has unconstrained access (primarily open-source), there are two interlocking challenges to understanding the potential risks: lack of a fixed target for assessment, and limitations on who can test and evaluate the system. One advantage of open-source systems is that they can be customized at the level of the source code, thereby providing more freedom than the fine-tuning currently provided by proprietary models (which is often limited to providing additional data or plug-ins).

However, the customizability of open-source systems also means that there is no single "target" for analyses of potential risks;[5] assessment of the potential risks requires a scalable testing environment and methodology that can be applied to multiple variants of a particular "base" system, but we currently lack a path forward to develop such a framework. The most extensive knowledge and expertise for adversarial testing of generative AI systems resides in private companies, but there can be significant legal and technical obstacles to working with these open-source systems in a corporate environment.

Universities, civil society, and government organizations are better able to work with open-source models, but do not typically have all of the necessary experience in designing and conducting informative tests of generative AI systems. As a result, there is currently a relative lack of understanding about exactly what risks are posed by presently available open-source generative AI systems, even though these systems are more accessible and transparent.

[4] Percy Liang, NAIAC briefing, August 3, 2023.
[5] For example, the guardrails on the initial release of an open-source generative AI system could potentially be removed through (malicious) customization.

ABOUT NAIAC

The National Artificial Intelligence Advisory Committee (NAIAC) advises the President and the White House National AI Initiative Office (NAIIO) on the intersection of AI and innovation, competition, societal issues, the economy, law, international relations, and other areas that can and will be impacted by AI in the near and long term. Their work guides the U.S. government in leveraging AI in a uniquely American way — one that prioritizes democratic values and civil liberties, while also increasing opportunity.

NAIAC was established in April 2022 by the William M. (Mac) Thornberry National Defense Authorization Act. It first convened in May 2022. It consists of leading experts in AI across a wide range of domains, from industry to academia to civil society.

https://www.ai.gov/naiac/

###