THE
FUTURE
SOCIETY

The Future Society's Response to OSTP's Request for Information on National Priorities for Artificial Intelligence

Prepared by The Future Society (TFS)

July 7, 2023

Stacy Murphy
Deputy Chief Operations Officer and Security Officer
Office of Science and Technology Policy (OSTP)
Executive Office of the President
Eisenhower Executive Office Building
1650 Pennsylvania Avenue
Washington, D.C. 20504

Subject: The Future Society's Response to OSTP's Request for Information on National Priorities for Artificial Intelligence (Docket (OSTP-TECH-2023-0007)

Dear Ms. Murphy,

The Future Society is pleased to submit comments in response to the Office of Science and Technology Policy (OSTP) Request for Information on National Priorities for Artificial Intelligence. We appreciate and welcome OSTP's effort to update U.S. national priorities and future actions on AI.

The Future Society (TFS) is a U.S. non-profit 501(c)(3) organization that develops, advocates for, and facilitates the implementation of AI governance mechanisms, ranging from laws and regulations to voluntary frameworks such as global principles, norms,

standards, and corporate policies. Through our activities, we hope to ensure that AI systems are safe and adhere to fundamental human values throughout their lifecycle.

Introduction

We are increasingly concerned by the rate of development and deployment of general-purpose AI systems (GPAIS): AI systems that can accomplish or be adapted to accomplish a broad range of tasks, including some for which they have not intentionally and specifically been trained [1]. For our intents and purposes, lexically, we treat GPAIS similar to "frontier models" or "foundation models" [2], terms describing the machine learning models underpinning GPAIS. Examples include but are not limited to OpenAI's GPT-4, Anthropic's Claude, Google's LaMDA, Google DeepMind's Chinchilla, Meta AI's LLaMa, Microsoft's GPT4-Prometheus, and Stability AI's Stable Diffusion 2.0 & XL.

Given the potential societal-scale impact of GPAIS, particularly in relation to national and global security, we focus our comments and recommendations on these systems, but we believe that, in principle, our recommendations serve to improve and safeguard the development of all AI systems. These comments build upon and complement our response to NTIA's AI Accountability Policy Request for Comment [3].

We urge OSTP to pay careful attention to the distinctive and broad-ranging national and global security challenges posed by GPAIS, as well as other societal-scale risks related to misinformation, bias, labor market transformation, and threats to democratic processes and human rights.

GPAIS advancements are progressing rapidly, often surpassing the pace of policy development. We lack conventional indicators that would allow us to clearly distinguish between benign and malicious applications, differentiate between intentional or unintentional misuse, or reliably attribute liability.

To address these risks, we suggest the OSTP adopts the following *high-level approach*:

- A holistic strategy that covers the entire AI lifecycle, from design to deployment, and aligns behaviors of developers and deployers with the public interest.
- A combination of technical and socio-technical mechanisms that promote responsible and safe innovation, and include complementary binding and voluntary governance frameworks.

We believe that such an approach will not stifle innovation but, on the contrary, produce a comprehensive AI governance regime that can ensure responsible innovation and secure U.S. leadership in a global landscape.

Our three recommendations propose *specific* AI governance measures that the OSTP can promptly take to promote safety and the protection of fundamental human values. We recognize that none of the recommendations are silver bullets, and stress that they are only a small part of a broader, much-needed governance regime composed of complementary binding and voluntary mechanisms.

Recommendation 1. Create and promote information, cyber, and physical security standards for the development of GPAIS.

Recommendation 2. Fund the development of measurement and evaluation frameworks for GPAIS.

Recommendation 3. Facilitate the adoption of a comprehensive industry-wide code of conduct that institutionalizes responsible behaviors and promotes a culture of safety among GPAIS developers.

Recommendations

Our recommendations directly address Question 1 but also touch upon issues explored in Questions 3 and 7.

- *Question* 1. What specific measures – such as standards, regulations, investments, and improved trust and safety practices – are needed to ensure that AI systems are designed, developed, and deployed in a manner that protects people's rights and safety? Which specific entities should develop and implement these measures?
- *Question 3.* Are there forms of voluntary or mandatory oversight of AI systems that would help mitigate risk? Can inspiration be drawn from analogous or instructive models of risk management in other sectors, such as laws and policies that promote oversight through registration, incentives, certification, or licensing?

- *Question 7.* What are the national security risks associated with AI? What can be done to mitigate these risks?

Recommendation 1. Create and promote information, cyber, and physical security standards for the development of GPAIS.

To address existing security vulnerabilities, we believe there is an urgent need for shared security standards to guide the development of GPAIS. These standards should be complemented by and ultimately enforced through federal regulation.

What are the security concerns related to GPAIS?

Currently, GPAIS are vulnerable to theft, hacks, leaks, and adversarial attacks throughout their lifecycles [4]. These threats exist from external actors, such as foreign adversaries, and can result in the use of GPAIS for malicious purposes, such as large-scale disinformation campaigns, phishing campaigns, cyber attacks, and other forms of terrorism.

Security threats also exist from internal actors, intentionally or unintentionally providing access to a model or its weights (e.g. in the form of leaks). The consequences in these scenarios can be just as detrimental for society and the environment.

For example, the infamous leak of Meta's LlaMa in February 2023 is just one case demonstrating how easily a model can be released into the world and adopted by spammers and those who engage in cybercrime to facilitate fraud or other obscene material. Additionally, this demonstrates how a leak, hack or theft can result in a foreign state getting access to millions of dollars worth of American R&D.

Considering the significant economic and geopolitical implications associated with GPAIS, we are concerned that adversaries, including state actors, will attempt to exploit them even more in the upcoming months and years.

How are GPAIS, specifically, vulnerable from a security perspective?

During the development stage, systems can be subject to data poisoning through different methods. An example is a trojan attack, when an external actor introduces a change to the learning environment, causing the system to produce erroneous or malicious outputs later on [5]. Another attack at the development stage is model inversion, which involves using a separate ("inversion") model to attempt to

reverse-engineer the training process to elicit data—which may include sensitive, private information—used to train the original model.

Adversarial attacks also occur downstream, against deployed AI systems. They may involve exploiting the system's processing methods through techniques such as adversarial examples, which trick AI systems into misclassifying data, such as mistaking an image of a tank for a school bus. Another type of attack against systems is prompt injection via the application interface of large language models, an existing vulnerability for which there are no known solutions [6].

Studies have also shown how GPAIS—having been trained with datasets that include information about dual-use biotechnology (e.g. synthetic DNA)—present a security threat by democratizing hazardous information. Last month, a group of MIT students demonstrated how, within one hour, they were capable of using a GPAIS to procure extensive instructions for making pandemic-class agents [7]. This included suggesting potential pathogens, explaining how they could be generated from synthetic DNA, identifying detailed protocols and how to troubleshoot them, and even supplying the names of DNA synthesis companies that would be unlikely to screen orders.

The complexity of the GPAIS lifecycle and the involvement of various stakeholders at different stages of their data processing, model training, and model supply chains only exacerbates their exposure to vulnerabilities. A GPAIS can contain multiple models, one of which may be fine-tuned with data from one source, using a base model from a specific vendor that claims data is used from a range of sources, where the data from each of those sources may also be labeled by different vendors.

Given their generality and potential to be adopted for a wide range of use cases, GPAIS also present single-point-of-failure risks [8], in which an exploit of a vulnerability of a single GPAIS could lead to far-reaching disruption across a range of applications and spheres (e.g. digital or physical infrastructure, financial services, and national and global security) [2].

How can we securitize GPAIS? What can we learn from other industries and sectors?

Shared security standards for the development of GPAIS can protect society from these vulnerabilities. Information and cyber security standards (as well as certification schemes, risk-management procedures, oversight mechanisms, and capacity-building programs) can leverage existing practices in nuclear facilities and banking infrastructure

[9], as well as best practices outlined in NIST Special Publication (SP) 800-53 [10]. Research facilities, as well as model testing and training environments could bolster their physical security by adapting relevant practices developed for biological laboratories by the CDC and NIH, outlined (and continuously updated) in Biosafety in Microbiological and Biomedical Laboratories (BMBL) [11]. Data centers involved in large training runs should be designed and operated in adherence to a combination of recognized security and infrastructure standards, such as NIST SP 800-53 [10], ISO/IEC 27001 [12], TIA-942 [13], and the Uptime Institute's Tier Classification System [14]. When handling sensitive data, the principles laid out in, for example, the Health Insurance Portability and Accountability Act (HIPAA) Security Rule [15] and Gramm-Leach-Bliley Act's (GLBA) Safeguards Rule [16] should also be taken into account to ensure stringent data security and privacy.

Recommendation 2. Fund the development of measurement and evaluation frameworks for GPAIS.

Measurement and evaluation frameworks can provide empirical data on the performance of AI systems, thereby improving our ability to govern them effectively. They can improve conformity assessments, support the development of normative instruments such as standards, and facilitate technology transfer [16].

Additionally, because establishing a comprehensive AI governance regime will be a prolonged process, measurement and evaluation frameworks could help establish some immediate guardrails. The OECD's framework for the classification of AI systems, to which TFS contributed, is one such example [17]. There have been numerous ad hoc efforts to develop measurement and evaluation tools, however, there is a unique opportunity for the U.S. to take leadership in this space: by publicly funding the development of GPAIS-focused measurement and evaluation frameworks, such as benchmarks, to ensure they are built in a manner that protects and promotes democratic values.

Benchmarks, for example, have a great potential to democratize access to reliable evaluation tools, due to costs of adoption that are comparatively lower than other policy instruments (whereas auditing, for example, can require domain expertise and processes that can be expensive for SMEs to implement). In addition to lower costs, with an appropriate structure in place—including capacity-building—benchmarks can

democratize the AI landscape by allowing end-users and impacted people to assess systems' performance and limitations.

We recommend that the Administration invest not only in the creation of dynamic measurement and evaluation tools that can assess GPAIS throughout their lifecycle today, but also in appropriate maintenance, including regular updates that will be necessary to account for new technological capabilities. Beyond stakeholders with technical assessment competence, the creation of these tools should meaningfully involve domain-relevant and social science experts to mitigate techno-solutionism or safety-washing.

The federal government can direct funds to the creation of evaluation systems in a collaborative, multi-stakeholder environment, which would increase their robustness and incentivize their widespread adoption. Specifically:

- U.S. Congress could increase funding for NIST Appropriations, specifically of Scientific and Technical Research and Services (STRS), so that they can focus on the development of measurement and evaluation tools of GPAIS.
- NSF, while clarifying their safety and ethical criteria which guide how their grants are allocated (42 U.S.C. §19052), could prioritize funds towards academic and civil society efforts which focus on GPAIS measurements (specifically, metrology) and evaluations.
- The Administration could increase staffing at the US Mission to the OECD (and other multilateral fora), to support measurement and evaluation efforts. A standalone AI advisor would ensure that the U.S. is adequately equipped to take leadership at the OECD, while improving purpose clarity and reducing the burden across the Mission.

Recommendation 3. Facilitate the adoption of a comprehensive industry-wide code of conduct that institutionalizes responsible behaviors and promotes a culture of safety among GPAIS developers.

We welcome the announcement by the U.S.-EU Transatlantic Trade and Technology Council of the development of an industry-wide code of conduct on AI [18]. Such a code of conduct should not undermine ongoing regulatory efforts, but rather support regulatory efforts and accelerate conformity to best practices across the industry. We

believe that a code of conduct could serve as a framework for further international cooperation and coordination, and that the U.S., serving as the physical headquarters for many frontier labs, has both an obligation and an opportunity to lead this effort.

The code of conduct can be effective if GPAIS developers commit to a series of safe and trustworthy practices including rigorous risk management protocols, internal model testing and evaluations, third-party auditing, accident prevention policies, staged release strategies, pre-registering large training runs, know-your-customer policies, and post-deployment assessments. To be operationalized, GPAIS developers must also dedicate adequate financial and human resources for the implementation of the commitments under the code of conduct, and there must also be a national or international monitoring and oversight mechanism. Such a mechanism could be institutionalized within the National Institute of Standards and Technology (NIST).

We believe it is vital that the code of conduct includes specific clauses applicable in the research and development (R&D) stages of GPAIS, in order to enable safe and trustworthy model design, development and testing. We are actively engaging industry professionals from frontier labs to compile a list of such clauses, such as commitments to document vulnerabilities and weaknesses, implement quality and risk-management frameworks, and promote testing protocols, and we plan to publish our recommended commitments in fall 2023.

Other civil society organizations are also making progress toward this area of research. For example, a survey published by the Centre for the Governance of AI has identified measures that have received near unanimous support from leading experts from AI labs, academia, and civil society [19]. The Partnership on AI is also working on a set of best practices and guidelines for downstream stages (i.e. deployment and monitoring) of the AI lifecycle [20]. Additionally, UC Berkeley's Center for Long-Term Cybersecurity has developed a risk-management standards profile for increasingly multi- or general-purpose AI [21]. We encourage the OSTP to leverage existing research and initiatives to facilitate the production and adoption of a holistic, industry-wide code of conduct.

Furthermore, we advocate for international coordination and collaboration in the development of a code of conduct that extends beyond the United States. This can be achieved through platforms like the U.S.-EU Transatlantic Trade and Technology Council, ensuring that the clauses are relevant and enforceable for developers

worldwide. By coordinating efforts internationally, the U.S. can work together with other countries to establish common guidelines for AI development and deployment. In the long-term, this harmonization will help avoid fragmented and conflicting regulations that could hinder innovation, create barriers to cross-border collaboration, and serve as obstacles for safe and trustworthy AI.

Conclusion

We, at The Future Society, appreciate this opportunity to comment on OSTP's request for information. If you have any questions regarding these comments and recommendations, please contact Niki Iliadis at niki.iliadis@thefuturesociety.org (cc:info@thefuturesociety.org).

Sincerely,

Niki Iliadis,
Director, AI and the Rule of Law, The Future Society

Amanda Leal,
Associate, The Future Society

Samuel Curtis,
Associate, The Future Society

The Future Society
867 Boylston Street, 5th Floor
Boston, MA, 02116

THE
FUTURE
SOCIETY

References

[1] C. I. Gutierrez, A. Aguirre, R. Uuk, C. Boine, and M. Franklin, "A Proposal for a Definition of General Purpose Artificial Intelligence Systems." Rochester, NY, Oct. 05, 2022. doi: 10.2139/ssrn.4238951.

[2] R. Bommasani *et al.*, "On the Opportunities and Risks of Foundation Models." arXiv, Jul. 12, 2022. doi: 10.48550/arXiv.2108.07258.

[3] The Future Society, "TFS Response to NTIA Request for Comment on AI Accountability Measures and Policies (NTIA-2023-0005-0001)," Jun. 15, 2023. https://www.regulations.gov/comment/NTIA-2023-0005-1185 (accessed Jul. 05, 2023).

[4] M. Musser *et al.*, "Adversarial Machine Learning and Cybersecurity," *Center for Security and Emerging Technology*. https://cset.georgetown.edu/publication/adversarial-machine-learning-and-cybersecurity/ (accessed Jul. 05, 2023).

[5] N. Strout, "The three major security threats to AI," *Center for Security and Emerging Technology*, Sep. 10, 2019. https://cset.georgetown.edu/article/the-three-major-security-threats-to-ai/ (accessed Jul. 05, 2023).

[6] Y. Liu *et al.*, "Prompt Injection attack against LLM-integrated Applications." arXiv, Jun. 08, 2023. Accessed: Jul. 05, 2023. [Online]. Available: http://arxiv.org/abs/2306.05499

[7] E. H. Soice, R. Rocha, K. Cordova, M. Specter, and K. M. Esvelt, "Can large language models democratize access to dual-use biotechnology?" arXiv, Jun. 06, 2023. Accessed: Jun. 10, 2023. [Online]. Available: http://arxiv.org/abs/2306.03809

[8] R. Bommasani and P. Liang, "Reflections on Foundation Models," *Stanford HAI*, Oct. 18, 2021. https://hai.stanford.edu/news/reflections-foundation-models (accessed Jul. 05, 2023).

[9] A. R. Wasil, "A Regulatory Framework for Advanced Artificial Intelligence."

[10] NIST Computer Security Resource Center, "Security and Privacy Controls for Information Systems and Organizations," National Institute of Standards and Technology, NIST Special Publication (SP) 800-53 Rev. 5, Dec. 2020. doi: 10.6028/NIST.SP.800-53r5.

[11] Centers for Disease Control and Prevention, National Institutes of Health, "Biosafety in Microbiological and Biomedical Laboratories (BMBL) 6th Edition | CDC Laboratory Portal | CDC," Feb. 03, 2021. https://www.cdc.gov/labs/BMBL.html (accessed Jul. 05, 2023).

[12] 14:00-17:00, "ISO/IEC 27001 Standard – Information Security Management Systems," *ISO*. https://www.iso.org/standard/27001 (accessed Jul. 05, 2023).

[13] "TIA-942 Certification," *TIA Online*. https://tiaonline.org/products-and-services/tia942certification/ (accessed Jul. 05,

2023).
[14] A. S. George, "Tier Classification System," *Uptime Institute*. https://uptimeinstitute.com/tiers (accessed Jul. 05, 2023).
[15] Office for Civil Rights (OCR), "The Security Rule," *U.S. Department of Health & Human Services*, Sep. 10, 2009. https://www.hhs.gov/hipaa/for-professionals/security/index.html (accessed Jul. 05, 2023).
[16] Federal Student Aid, "Updates to the Gramm-Leach-Bliley Act Cybersecurity Requirements | Knowledge Center," Feb. 09, 2023. https://fsapartners.ed.gov/knowledge-center/library/electronic-announcements/2023-02-09/updates-gramm-leach-bliley-act-cybersecurity-requirements (accessed Jul. 05, 2023).
[17] "OECD Framework for the Classification of AI systems | en | OECD." https://www.oecd.org/publications/oecd-framework-for-the-classification-of-ai-systems-cb6d9eca-en.htm (accessed Jul. 05, 2023).
[18] N. Lomas, "EU and US lawmakers move to draft AI Code of Conduct fast," *TechCrunch*, May 31, 2023. https://techcrunch.com/2023/05/31/ai-code-of-conduct-us-eu-ttc/ (accessed Jul. 05, 2023).
[19] J. Schuett *et al.*, "Towards best practices in AGI safety and governance: A survey of expert opinion".
[20] L. Baldwin, "PAI Is Collaboratively Developing Shared Protocols for Large-Scale AI Model Safety," *Partnership on AI*, Apr. 06, 2023. https://partnershiponai.org/pai-is-collaboratively-developing-shared-protocols-for-large-scale-ai-model-safety/ (accessed Jul. 05, 2023).
[21] A. M. Barrett, D. Hendrycks, J. Newman, and B. Nonnecke, "AI Risk-Management Standards Profile for Increasingly Multi- or General-Purpose AI: First Full Draft," *Google Docs*, May 10, 2023. https://docs.google.com/document/d/1Q98T2GHmyXoAKGihTZJKpKN72wl2Jf_Wuzo49Dr0zXA/edit?usp=embed_facebook (accessed Jul. 05, 2023).