National Priorities for Artificial Intelligence

Response to the OSTP Request for Information by the Centre for the Governance of AI

Jonas Schuett
Research Fellow
Centre for the Governance of AI
jonas.schuett@governance.ai

Markus Anderljung
Head of Policy
Centre for the Governance of AI
markus.anderljung@governance.ai

Lennart Heim
Research Fellow
Centre for the Governance of AI
lennart.heim@governance.ai

Elizabeth Seger
Research Scholar
Centre for the Governance of AI
elizabeth.seger@governance.ai

7 July 2023

About the Centre for the Governance of AI (GovAI)

The Centre for the Governance of AI (GovAI) is a nonprofit based in Oxford with a presence in the US. It was founded in 2018 as part of the Future of Humanity Institute at the University of Oxford, before becoming an independent research organization in 2021. GovAI's mission is to build a global research community, dedicated to helping humanity navigate the transition to a world with advanced AI.

About the authors

Jonas Schuett is a Research Fellow at GovAI and a Research Affiliate at the Legal Priorities Project. His research focuses on the governance of frontier AI developers, with a focus on risk management, as well as frontier AI regulation. Before joining GovAI, he advised the UK government on AI regulation and interned at Google DeepMind's Policy Team. He has a background in law.

Markus Anderljung is Head of Policy at GovAI, an Adjunct Fellow at the Center for a New American Security (CNAS), and a member of the OECD AI Policy Observatory's Expert Group on AI Futures. His work aims to identify new and improve upon existing AI governance policy recommendations. He was previously seconded to the UK Cabinet Office as a Senior Policy Specialist.

Lennart Heim is a Research Fellow at GovAI and a member of the OECD.AI Expert Group on AI Compute and Climate. His work focuses on compute governance, including the role of compute in the AI production function, the compute supply chain, forecasting emerging technologies, and AI system security. He has a background in computer engineering.

Elizabeth Seger is a Research Scholar at GovAI and a Research Affiliate with the Centre for the Study of Existential Risk (CSER) at Cambridge University. Her research focuses on the democratization of AI and epistemic security. She has a background in philosophy of science.

The views expressed in this submission are those of the authors and do not represent the views of GovAI.

Overview

We welcome the opportunity to respond to the [OSTP Request for Information on National Priorities for Artificial Intelligence] and look forward to future opportunities to provide additional input. We offer the following submission for your consideration.

Frontier AI regulation [more]	**Compute governance** [more]	**AI and democracy** [more]
● We need specific regulation for frontier AI models. ● Defining the scope of frontier AI regulation is challenging. ● Regulators need more visibility into frontier AI development. ● Frontier AI developers should be required to: ○ Conduct thorough risk assessments informed by evaluations of dangerous capabilities and controllability ○ Engage external experts to scrutinize frontier AI models ○ Follow shared guidelines for how frontier AI models should be deployed based on their assessed risk ○ Monitor and respond to new information on model capabilities ○ Comply with cybersecurity standards. ● In the future, the deployment and potentially even the development of frontier AI models may require a license. ● The US Government should support the creation of standards for the development and deployment of frontier AI models.	● Compute is a particularly promising node to govern frontier AI models. ● The US Government should grant the Bureau of Industry and Security (BIS) a larger budget and empower it with the tools to effectively enforce the October 7th export controls. ● Frontier AI developers should be required to report training runs above a certain threshold. ● Compute providers should be required to have "Know Your Customer (KYC)" processes for compute purchases above some very large size. ● If companies want access to more compute, they should be subject to additional review requirements ("more compute, more responsibility").	● AI might threaten democracy. ● "Democratizing AI" does not mean that frontier AI developers should open-source models. ● "Democratizing AI" is ultimately about ensuring benefits of AI are distributed widely and fairly.

Table 1. Overview of our recommendations

1. Risks from frontier AI models

It is important that the US Government has an accurate understanding of the risks from frontier AI models. By "frontier AI models", we mean highly capable foundation models[1] that could have dangerous capabilities sufficient to cause severe harm to public safety and global security.[2] We think the next generation of state-of-the-art foundation models – in particular, those models trained using substantially greater computational resources than any model trained to date – are likely enough to have these capabilities to warrant regulation.

Foundation models already cause significant harm. For example, language models like GPT-4 can produce racist, sexist, and homophobic outputs,[3] or be used for disinformation campaigns[4] and cyberattacks,[5] while image generation models like Stable Diffusion or Midjourney can be used to create harmful content,[6] such as non-consensual deepfake pornography.[7]

Further integrating foundation models into society might lead to systemic risks. Since many AI-based applications and services are built on top of frontier AI models, flaws in the base model can quickly affect the entire economy. The increasing reliance on frontier AI models therefore introduces new vulnerabilities and systemic risks, especially if such models are used in critical infrastructure.

As foundation models become more capable, more extreme risks might emerge. Training models with more compute, bigger datasets, and more parameters predictably lead to more capable models.[8] This trend has been the driver of recent progress in AI research and development, but it also has concerning implications. As models are scaled up, new capabilities can emerge unintentionally and unpredictably,[9] some of which might be dangerous.[10] For example, models might become able to manipulate people, discover cyber vulnerabilities, or develop novel biological weapons.[11] These capabilities may be misused by malicious actors or used inadvertently by the system itself, with potentially catastrophic consequences.[12] Some even think that certain combinations of capabilities could lead to human extinction.[13]

[1] "Foundation models" are models trained on broad data that can be adapted to a wide range of downstream tasks, see Bommasani et al., On the opportunities and risks of foundation models, 2021.

[2] Anderljung et al., Frontier AI regulation: Managing emerging risks to public safety, forthcoming.

[3] Bolukbasi et al., Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, 2016; Bender et al., On the dangers of stochastic parrots: Can language models be too big?, 2021; Weidinger et al., Ethical and social risks of harm from language models, 2021.

[4] Buchanan et al., Truth, lies, and automation: How language models could change disinformation, 2021.

[5] Brundage et al., The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, 2018; Hazell, Large language models can be used to effectively scale spear phishing campaigns, 2023.

[6] Horvitz, On the horizon: Interactive and compositional deepfakes, 2022.

[7] Westerlund, The emergence of deepfake technology: A review, 2019.

[8] This phenomenon is commonly known as "scaling laws" (Kaplan et al., Scaling laws for neural language models, 2020) and the claim that this trend will continue as the "scaling hypothesis" (Gwern, The scaling hypothesis, 2020). But note that it has been argued that the current rate of scaling may be unsustainable (Lohn & Musser, AI and compute: How much longer can computing power drive artificial intelligence progress? 2022; Heim, This can't go on(?) - AI training compute costs, 2023).

[9] Ganguli et al., Predictability and surprise in large generative models, 2022; Wei et al., Emergent abilities of large language models, 2022.

[10] Shevlane et al., Model evaluation for extreme risks, 2023.

[11] Urbina et al., Dual use of artificial-intelligence-powered drug discovery, 2022; Sandbrink, Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools, 2023.

[12] Shevlane et al., Model evaluation for extreme risks, 2023.

[13] Carlsmith, Is power-seeking AI an existential risk? 2022; Ngo, Chan, & Mindermann, The alignment problem from a deep learning perspective, 2022.

2. Frontier AI regulation

Regulating frontier AI models should be a key priority for the US Government. In a recent paper, we argue that such a regulatory regime is needed and describe its necessary building blocks.[14]

We need specific regulation for frontier AI models. While other AI systems also need to be regulated, frontier AI models warrant targeted attention. Self-regulation and civil liability are important, but will not be sufficient. They should be seen as a complement to regulation, not a substitute.

Defining the scope of frontier AI regulation is challenging. We define "frontier AI models" as highly capable foundation models[15] that could have dangerous capabilities sufficient to cause severe harm to public safety and global security.[16] However, any binding regulation of frontier AI models would require a much more precise definition.[17] The definition would also be an important building block for the creation and dissemination of voluntary standards. It is worth noting that what qualifies as a frontier AI model changes over time. The scope definition needs to be able to account for this.

Regulators need more visibility into frontier AI development. They need information to address the appropriate regulatory targets and design effective tools for governing frontier AI models. Mechanisms to give regulators visibility into frontier AI development might include disclosure regimes, monitoring processes, and whistleblower protections.

Requirements for frontier AI developers. While it is still unclear which specific requirements a regulatory regime for frontier AI development should contain, the following seem particularly important.[18]

- **Conducting thorough risk assessments informed by evaluations of dangerous capabilities and controllability.**[19] This would reduce the risk that deployed models possess unknown dangerous capabilities or behave unpredictably and unreliably.

- **Engaging external experts to scrutinize frontier AI models.**[20] External scrutiny of the safety and risk profile of models would both improve assessment rigor and foster accountability to the public interest.

- **Following shared guidelines for how frontier AI models should be deployed based on their assessed risk.** The results from risk assessments should determine whether and how a model is deployed, and what safeguards are put in place. Options could range from deploying the model

[14] Anderljung, Barnhart, Leung, Korinek, O'Keefe, & Whittlestone et al., Frontier AI regulation: Managing emerging risks to public safety, forthcoming.
[15] "Foundation models" are models trained on broad data that can be adapted to a wide range of downstream tasks (Bommasani et al., On the opportunities and risks of foundation models, 2021).
[16] This definition is taken from Anderljung, Barnhart, Leung, Korinek, O'Keefe, & Whittlestone et al., Frontier AI regulation: Managing emerging risks to public safety, forthcoming.
[17] Schuett, Defining the scope of AI regulations, 2023.
[18] For more information, see Anderljung, Barnhart, Leung, Korinek, O'Keefe, & Whittlestone et al., Frontier AI regulation: Managing emerging risks to public safety, forthcoming.
[19] See Shevlane et al., Model evaluation for extreme risks, 2023.
[20] Anderljung et al., Public accountability via external scrutiny of foundation models: Audits, red teaming, and researcher access, forthcoming; Mökander et al., Auditing large language models: A three-layered approach, 2023; Thornton et al., Response to the NTIA AI Accountability Policy Request for Comment, 2023.

without restriction to not deploying it at all until risks are reduced. In many cases, an intermediate option—deployment with appropriate safeguards, such as restrictions on the ability of the model to respond to risky instructions—will be appropriate.

- **Monitoring and responding to new information on model capabilities.** The assessed risk of deployed frontier AI models may change over time due to new information and new post-deployment enhancement techniques. If significant information on model capabilities is discovered post-deployment, risk assessments should be repeated, and deployment safeguards updated.

- **Comply with cybersecurity standards**, such as ISO/IEC 27001 or the NIST Cybersecurity Framework. These standards need to be tailored to the context of frontier AI developers.

In the future, the deployment and potentially even the development of frontier AI models may require a license. Although licensing is a promising regulatory instrument, we are uncertain if it will be warranted for the next generation of frontier AI models. It might also be the case that the requirements mentioned above are sufficient. In general, designing a well-balanced licensing regime will be challenging. On the one hand, we should be sensitive to the risks of overregulation and stifling innovation. On the other hand, we need to keep up with the pace of AI progress and emergent risks. In any case, policymakers should seriously consider this type of instrument.

The US Government should support the creation of standards for the development and deployment of frontier AI models. For example, the NIST AI Risk Management Framework[21] could be tailored to frontier AI developers.[22] The Partnership on AI (PAI) has initiated a multi-stakeholder dialogue to develop shared protocols for the safety of large-scale AI models.[23] We recently conducted an expert survey on best practices in AI safety and governance.[24] The US Government should also collaborate with its allies. Existing efforts like the US-EU Trade and Technology Council (TTC) are a promising starting point.[25]

3. Compute governance

The US Government should take additional measures to control computing power ("compute") or use it as a governance node.

Compute is a particularly promising node to govern frontier AI models.[26] There are at least three reasons for this. First, compute can improve our understanding of how actors use, develop, and deploy AI, as well as which actors are relevant. This knowledge is crucial, as it allows us to make more accurate decisions, anticipate problems, and track key outcomes. Second, compute can be used for strategic resource allocation, allowing us to influence who has access to what AI capabilities. A

[21] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.
[22] Barrett et al., AI risk management-standards profile for increasingly multi- or general-purpose AI, 2023; Barrett et al., Actionable guidance for high-consequence AI risk management: Towards standards addressing AI catastrophic risks, 2022.
[23] Partnership on AI, PAI Is collaboratively developing shared protocols for large-scale AI model safety, 2023.
[24] Schuett et al., Towards best practices in AGI safety and governance: A survey of expert opinion, 2023.
[25] US-EU TTC, Joint statement of the Trade and Technology Council, 2023; US-EU TTC, Joint roadmap on evaluation and measurement tools for trustworthy AI and risk management, 2022.
[26] Brundage et al., Computing power and the governance of artificial intelligence, forthcoming; Heim, Introduction to Compute Governance, 2023.

special case of distribution is differential technological development,[27] incentivizing the development of more beneficial AI systems. Third, compute can be a tool to respond to violations, like an actor training an excessively risky AI system. Enforcement could be achieved via norms, laws, or other procedures, and the details will differ substantially based on the particular context.

The US Government should grant the Bureau of Industry and Security (BIS) a larger budget and empower it with the tools to effectively enforce the October 7th export controls.[28] PRC-based AI developers could access restricted chips via illicit procurement networks, i.e., chip smuggling. This could be addressed via creating a chip registry including randomized end-use checks.[29] [30]

Frontier AI developers should be required to report training runs above a certain threshold. As a first step, the US Government could set up a voluntary information-sharing pilot program with frontier AI developers (e.g. OpenAI, Google DeepMind, and Anthropic), focusing on compute usage and model capability evaluations.[31] Such information-sharing would only be encouraged for a small number of frontier AI models, which are particularly compute-intensive or have especially general capabilities. This information could be shared both before and throughout model training and deployment processes. Shortly before release, developers could also grant a state actor direct access to their models.

Compute providers should be required to have "Know Your Customer (KYC)" processes for compute purchases above some very large size. Such requirements should apply to cloud providers in the US and actors who use chips with US-originating technology. These actors should identify who is using large amounts of their computing power. This is analogous to requirements imposed on banks to know who their customers are. A first step could be to include non-invasive checks on whether their compute is being used for suspicious activities (e.g. unusually large transfers of data to Chinese locations).

If companies want access to more compute, they should be subject to additional review requirements ("more compute, more responsibility"). In general, we think that regulatory burdens for frontier AI developers should scale with the capabilities of the relevant models. More capable models will tend to offer higher potential benefits, but also pose larger risks. As the amount of compute used to train a model is a useful proxy for its performance, we expect that the amount of compute could become a particularly important factor that determines the regulatory burdens imposed on future systems.

4. AI and democracy

The US Government should take the effects that frontier AI models might have on democracy seriously. They should also have a nuanced understanding of what "democratizing AI" means.

[27] Sandbrink et al., Differential technology development: A responsible innovation principle for navigating technology risks, 2022.
[28] Allen et al., Improved export controls enforcement technology needed for U.S. national security, 2022.
[29] Fist, Heim & Schneider, Chinese firms are evading chip controls, 2023.
[30] Heim & Anderljung, Comment on October 7 advanced computing and semiconductor manufacturing equipment rule, 2023 (Confidential version available on request).
[31] See Mulani & Whittlestone, Proposing a foundation model information-sharing regime for the UK, 2023.

AI might threaten democracy. Generative AI models can be used to produce false or misleading synthetic media which can then be targeted at individuals to sway public opinion and drive polarization.[32] These capabilities could allow malicious actors to influence election outcomes or to undermine well informed public deliberation, the cornerstone of well-functioning democracies.[33] It has been proposed that future capabilities might also be used to construct intricate but false histories for public leaders making it difficult to contextualize and verify disinformation about them.[34] Such capabilities could severely undermine public trust in democratically elected leaders and democratic systems.

"Democratizing AI" does not mean that frontier AI developers should open-source models. Open-sourcing frontier AI models has a number of benefits. For example, it facilitates external evaluation of AI models by the wider AI community and helps to distribute influence over the future of AI out from under the sole purview of big tech. However, we wish to emphasize that open-sourcing frontier AI models is not always desirable. Most notably, open-sourcing AI models allows anyone with the requisite technical background (e.g. computer science graduate students) to bypass safety restrictions and to optimize models for malicious use. Once a decision to open-source is made, it can not be rolled back if major safety or misuse issues emerge.[35]

"Democratizing AI" is ultimately about ensuring benefits of AI are distributed widely and fairly.[36] This is partly about facilitating easy access to safe and beneficial AI tools. It is also about redistributing profits so that the massive value produced by AI does not accrue only to a few leading AI labs. Diverse stakeholder interests should also be represented in important decision-making about AI - e.g. about how profits are redistributed, about acceptable levels of risk for model release, and about how and by whom AI should be governed.

[32] Goldstein et al., Generative language models and automated influence operations: Emerging threats and potential mitigations, 2023.
[33] Seger et al., Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world, 2020.
[34] Horvitz, On the horizon: Interactive and compositional deepfakes, 2022.
[35] Seger et al., Don't rush to open-source foundation models, forthcoming.
[36] Seger et al., Democratising AI: Multiple meanings, goals, and methods, 2023.

Further resources

GovAI researchers have published several pieces relevant to this Request for Information.

Frontier AI regulation

- Anderljung, Barnhart, Leung, Korinek, O'Keefe, & Whittlestone et al., Frontier AI regulation: Managing emerging risks to public safety, forthcoming
- Shevlane et al., Model evaluation for extreme risks, 2023
- Schuett et al., Towards best practices in AGI safety and governance: A survey of expert opinion, 2023
- Schuett, Defining the scope of AI regulations, 2023
- Schuett, Risk management in the Artificial Intelligence Act, 2023
- Mökander et al., Auditing large language models: A three-layered approach, 2023
- Thornton et al., Response to the NTIA AI Accountability Policy Request for Comment, 2023
- Schuett & Anderljung, Comments on the Initial Draft of the NIST AI Risk Management Framework, 2022
- Anderljung et al., Public accountability via external scrutiny of foundation models: Audits, red teaming, and researcher access, forthcoming

Compute governance

- Heim & Anderljung, Future of compute review: Call for evidence, 2022
- Heim & Anderljung, Submission to the Request for Information (RFI) on Implementing Initial Findings and Recommendations of the NAIRR Task Force, 2022
- Mulani & Whittlestone, Proposing a foundation model information-sharing regime for the UK, 2023
- Whittlestone et al., Response to the UK's Future of Compute Review: A missed opportunity to lead in compute governance, 2023
- Brundage et al., Computing power and the governance of artificial intelligence, forthcoming.

AI and democracy

- Seger et al., Democratising AI: Multiple meanings, goals, and methods, 2023
- Seger et al., Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world, 2020
- Seger et al., Don't rush to open-source foundation models, forthcoming