ANTHROP\C

548 Market St., PMB 90375
San Francisco, CA 94104-5401

412-837-9797
anthropic.com

March 27, 2024

Bertram Lee
National Telecommunications and Information Administration
U.S. Department of Commerce
1401 Constitution Avenue NW
Washington, D.C. 20230

RE: National Telecommunications and Information Administration's (NTIA) Request for Comment (RFC) on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights (Docket Number NTIA–2023–0009)

Submitted via: Regulations.gov

Thank you for the opportunity to respond to the National Telecommunications and Information Administration's (NTIA) Request for Comment (RFC) on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights. We understand that the purpose of this request is to "conduct a public consultation process and issue a report on the potential risks, benefits, other implications, and appropriate policy and regulatory approaches to dual-use foundation models for which the model weights are widely available."[1]

About Anthropic

Anthropic is an AI safety and research company working to build reliable, interpretable, and steerable AI systems. Our mission is to develop and deploy advanced AI systems that are helpful to people. Core to our mission is the development of cutting-edge, or "frontier" large language models—we use these to conduct empirical safety research and as the main ingredient in the systems we deploy commercially.

Introduction

The vast majority of science has advanced over the years due to a culture of openness and transparency around research. This is especially true in the field of artificial intelligence, where many advancements have been made possible by the open publication of research. These openly shared innovations have served as building blocks for further technological developments.

[1] 89 Fed. Reg. 14059 (Feb. 26, 2024).

At Anthropic, we recognize the importance of fostering the societal benefits that have stemmed from this open and transparent approach to research. Our core belief is that while the vast majority of today's AI models are safe to release openly, that may not always be true in the future. As models become increasingly more capable,[2] they may have the potential to lead to major misuses or catastrophic accidents. We believe that AI-driven accidents or misuses could lead regulators to take sudden or extreme actions, which we want to avoid. Therefore, we believe it is necessary for us to collectively design tests that effectively assess the potential for accidents and misuse of AI systems—whether proprietary or openly disseminated—and use these tests to guide policy solutions.

We believe it is crucial that the government strike a balance between mitigating the potential safety risks of open source models, while also preserving robust innovation. Indeed, constraining the open dissemination of AI research and AI systems could have a chilling effect on competition in AI markets and personal liberties, which should be viewed as an extreme cost. We should only undertake such costly actions if we have direct and serious evidence that AI systems could cause accidents or enable misuse that have critical national security implications or substantial societal costs that cannot be reasonably mitigated.

For this reason, our comment recommends that powerful general purpose AI models—those that are open source and proprietary—undergo standardized safety testing. We believe that developing and standardizing these safety tests will enable stakeholders across government, industry, academia, and civil society to gather information about the misuse and accident potential of such systems and use this to inform a (no doubt robust) debate about how to approach openly disseminated AI systems. We believe such a safety testing regime will incentivize model safety and create a safety race-to-the-top amongst industry actors who will strive to have the safest models on the market. It can also serve as the template for potential regulatory approaches—if, and only if, we arrive at safety measures that are viewed as legitimate indicators of serious accident or misuse.

Mitigating Safety Risks of Powerful AI Models Through a Standardized Pre-Release Testing Regime

As government, industry, academic, and civil society stakeholders work to balance responsible AI development and innovation in America's rapidly evolving AI landscape, we must also understand the potential safety risks associated with frontier models (and mitigate any consequences therefrom). At Anthropic, we believe it is possible to mitigate these risks and level the playing field by requiring that general purpose open source *and* proprietary AI models go through a shared set of tests to generate information about their misuse and accident

---

[2] *See* Written Testimony of Jack Clark, Co-Founder and Head of Policy, Anthropic, Hearing on "*Federal Science Agencies and the Promise of AI in Driving Scientific Discoveries*," Committee on Science, Space, and Technology United States House of Representatives (Feb. 6, 2024), *available at* https://www.congress.gov/118/meeting/house/116790/witnesses/HHRG-118-SY15-Wstate-ClarkJ-20240206.pdf (accessed Mar. 27, 2024).

potential. We believe such tests are an essential prerequisite to any regulatory conversation about these systems.

A core tenet of AI safety is the ability to accurately describe and measure the capabilities and safety characteristics of AI systems. This ability is both an enabler and a prerequisite to effective regulation, as measurement tools allow us to objectively assess the capabilities of AI systems and ensure they meet appropriate safety thresholds. To achieve this, Anthropic strongly advocates for the establishment of a standardized safety testing regime conducted by an independent third party.

We recently encouraged NIST, as the leading U.S. federal agency on measurement science and standards development, to prioritize the creation, operation, maintenance, verification, and sharing of results from authoritative benchmarks for AI systems.[3] These benchmarks, developed and administered by an independent third party, will help establish norms and standardize practices for the evaluation and reporting of AI systems, ultimately driving transparency and trust in AI technology. It will also create a natural point of coordination for the AI sector: questions about the safety of AI systems are naturally complex and at times controversial, and actors like ourselves are not going to be viewed as impartial. Therefore, having a third-party take on this challenge has the best chance of generating the best information that we can use to discuss the safety of AI systems.

This recommendation aligns with our June 2023 submission to NTIA, where we emphasized the crucial role of governments in supporting the development of rigorous capability and safety evaluations targeted at critical risks from advanced AI, such as deception and autonomy.[4] By investing in third parties to carry out AI measurements, we can lay the ground for a comprehensive, independently administered testing regime that we can use to generate better information about AI systems.

Ultimately, to achieve safety across a range of large-scale AI models, a testing regime must be easy to understand, have a low barrier for participation, and not pose an onerous cost to relatively small organizations seeking to openly and broadly disseminate models. Such a regime would ensure that all AI developers, regardless of their size or resources, can contribute to the advancement of AI technology while prioritizing safety and responsibility. This testing regime would also serve as a kind of regulatory laboratory, enabling the government, industry, and other stakeholders to develop different measures, run them against AI systems, then come together to discuss whether the results of these measurements are meaningful for policymakers. This would generate the evidence needed to develop potential regulatory approaches that tie these testing approaches to regulations about the ability to broadly deploy or release AI systems.

---

[3] Anthropic Comment on FR Doc # 2023-28232 (Feb. 2, 2024), *available at* https://www.regulations.gov/comment/NIST-2023-0009-0100 (accessed Mar. 27, 2024).
[4] Anthropic Comment on FR Doc # 2023-07776 (June 15, 2023), *available at* https://www.regulations.gov/comment/NTIA-2023-0005-0640 (accessed Mar. 27, 2024).

Anthropic would welcome the opportunity to contribute to the creation of such safety standards via our work on model capability and evaluation and, in particular, our work on frontier red teaming for national security threats.

Conclusion

It is critical to drive transparency and trust in AI technology while ensuring appropriate safety thresholds are met before release. We believe the establishment of a standardized pre-release safety testing regime for powerful general-purpose AI models will best advance this goal in the near term and we strongly encourage the Department of Commerce (through NIST) to create benchmarks, guidelines, and best practices for evaluating and reporting on AI systems. In addition to further research on model evaluations, we would also suggest that governments, academia, and industry developers invest in research to explore opportunities to make models more resistant to fine tuning, given that it is currently easy to remove built-in safety guardrails from openly-released models. Such research, if successful, could mitigate many of the safety concerns unique to releasing powerful models openly. Thank you again for the opportunity to submit this comment and for NTIA's leadership on this important issue.