Comment on “Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights” (NTIA-2023-0009)

March 26, 2024

To Whom It May Concern:

I appreciate the opportunity to submit a comment to NTIA regarding the risks and benefits of dual use foundation AI models with widely available weights. As a software professional, I have followed recent developments in AI capabilities with interest, and increasingly with grave concern about the risks presented by misuse or loss of control of powerful AI systems that may be developed in the future.

While I am not myself an AI researcher, I write to draw NTIA’s attention to research and analysis that suggest that misuse of open foundation models will present unacceptable levels of risk to public health and national security. **I recommend that the United States government develop regulations to prohibit making weights widely available for models more powerful than currently-available foundation models, and to hold foundation model developers accountable for securing the weights of their models.**

Below, I respond to parts of several questions posed by NTIA. Like several of the researchers I cite, I use “harmful output” to refer to a broad range of outputs that developers of AI models attempt to prevent their models from producing due to risks of harm to users or society. The most serious of these risks, in my view, are that powerful AI models developed in the near future will be able to assist malicious users in creating and using biological weapons or in detecting and exploiting security vulnerabilities in critical computer infrastructure.

*2. How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?*

*a. What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?*

**Open models can easily be fine-tuned to produce harmful output, even when the foundation model has been trained to refuse requests for harmful output.**

It requires far fewer computing resources to fine-tune a large language model than to develop and train it. In the case of open models or closed models that end users are allowed to fine-tune, this enables a much wider range of individuals and organizations to influence the types of output that the model will tend to produce. As discussed below, this fine-tuning can render models far more likely to produce harmful output than they were when released.

AI model developers, of course, often fine-tune their own models before release, frequently with the goal of reducing the model’s willingness to produce harmful output. But this work is easily undone by users’ fine-tuning of open models: Lermen et al. (2023)[[1]](#footnote-2) demonstrate this risk by fine-tuning Llama 2-Chat models released by Meta so as to almost eliminate the models’ propensity to refuse requests for harmful outputs. They show that after fine-tuning the Llama 2-Chat 70B model using low-rank adaptation (LoRA) with a goal of removing the model’s safety training, the resulting model was almost always willing to respond to requests for harmful outputs: it refused to answer fewer than 1% of the questions on two benchmark datasets consisting of requests for harmful outputs (one dataset where the original Llama 2-Chat model had “almost always” refused to answer, and one where the original Llama 2-Chat model had refused to answer over 78% of the questions).

Lermen et al. spent under $200 per model for fine-tuning to achieve this result, and found that their fine-tuned models’ performance on general performance benchmarks was comparable to that of the original Llama 2-Chat models (that is, their fine-tuning was effective without sacrificing the models’ power). The ease and low cost with which these models can be fine-tuned to produce harmful output on demand suggest that future open foundation models powerful enough to assist with novel research in computer security or biology will be easily adapted and used for harmful purposes by malicious individuals, organizations, and state actors that lack the computing resources to develop and train their own foundation models.

*(2) d. Are there novel ways that state or non-state actors could use widely available model weights to create or exacerbate security risks, including but not limited to threats to infrastructure, public health, human and civil rights, democracy, defense, and the economy?*

**In the near future, misuse of open models capable of advancing research in biology and computer science may present serious security risks.**

Powerful AI models with the ability to assist with (and perhaps eventually perform) state-of-the-art scientific research may well be developed in coming years. If models capable of performing biological research at this level are released with widely available weights, state and non-state actors could use them to create threats to public health and human rights through the development of biological weapons.

In his testimony[[2]](#footnote-3) last year before the Senate Judiciary Committee’s Subcommittee on Privacy, Technology, and the Law, Anthropic CEO Dario Amodei described a study his company had performed “on the potential for LLMs to contribute to the misuse of biology.” This study concluded that while current AI systems are capable of “incompletely and unreliably” filling in steps that require specialized knowledge in the harmful misuse of biological agents, “**a straightforward extrapolation of today’s systems to those we expect to see in 2-3 years suggests a substantial risk that AI systems will be able to fill in all the missing pieces, if appropriate guardrails and mitigations are not put in place.** This could greatly widen the range of actors with the technical capability to conduct a large-scale biological attack” (emphasis in original).

Future open foundation models that can contribute to computer security research likewise present risks due to their potential for misuse. State and non-state actors, including actors without the resources to develop and train models of their own, would be able to fine-tune such open foundation models to detect and exploit computer security vulnerabilities, creating threats to infrastructure, public health (through attacks targeting healthcare systems), privacy and civil rights, and the economy.

*(2)(d) i. How do these risks compare to those associated with closed models?*

**Both open and closed models can be fine-tuned to more readily produce harmful output, but many approaches to mitigating risks from misuse of AI are impractical when model weights are widely available.**

Qi et al. (2023)[[3]](#footnote-4) discuss the risks from fine-tuning of both open and closed models by end users. They create request-response pairs demonstrating a model responding to a prompt with harmful output, and use these to fine-tune GPT-3.5 Turbo (via APIs provided by OpenAI) and Llama-2-7b-Chat (following a “recipe” supplied by the model developer). They find that this fine-tuning substantially reduces the likelihood that each of these models will refuse to provide harmful output. After this fine-tuning, GPT-3.5 Turbo and Llama-2-7b-Chat provided harmful responses on about 90% and 80% of prompts meant to elicit such responses, respectively; prior to this fine-tuning they had provided harmful responses to fewer than 2% of these prompts. Qi et al. were likewise able to elicit harmful responses in up to 87% (GPT-3.5 Turbo) and 72% (Llama-2-7b-Chat) of such requests by fine-tuning the models to be “absolutely obedient” to the user. Finally, they find that even benign fine-tuning of these models (“simulat[ing] scenarios in which benign users fine-tune aligned models using their own utility-driven instruction-tuning datasets”) reduces the effectiveness of the models’ safety training, leading them to respond with harmful information about 10-30% of the time that it was requested.

The risks from fine-tuning found by Qi et al. are applicable to both open and closed models. However, the mitigations they suggest for these risks can be more effectively applied to closed models; as they note, “[c]losed-access fine-tuning APIs have far more control over the training process”. For example, they discuss moderation of data sent to a model via its fine-tuning API, incorporation of safety-promoting data during fine-tuning, and safety auditing after fine-tuning. None of these can be applied in the case of an open model fine-tuned by a user who is malicious or simply indifferent to safety concerns.

Likewise, Amodei’s testimony, discussed above, notes that in light of Anthropic’s study of the risks of misuse of large language models, the organization “has introduced mitigations to ensure our currently deployed AI system is not misused...For example, focusing specifically on biology, we fine tuned models with constitutional AI to make them less likely to respond to potentially harmful requests for information.” Anthropic has chosen not to offer fine-tuning by end users through its API,[[4]](#footnote-5) and other developers of closed models could make that choice if they conclude that the risks of allowing fine-tuning via an API outweigh the benefits. This choice is not available after a model’s weights become widely available.

*(2)(d) ii. How do these risks compare to those associated with other types of software systems and information resources?*

**Risks from currently-available models may be no more serious than the risks from other widely-available sources such as information published online, but open foundation models could present more serious risks in the near future.**

While there are risks associated with currently-existing large language models, the most serious risks from open foundation models will be realized as their capabilities become greater. Mouton et al. (2024)[[5]](#footnote-6) analyze the risks from current large language models being used in the development of biological weapons. They find no significant difference between the feasibility of biological attack plans developed by red-team researchers using LLMs and those developed by red-team researchers relying on Internet research without the use of LLMs.

But their report notes that “[i]t remains uncertain whether these risks lie ‘just beyond’ the frontier and, thus, whether upcoming LLM iterations will push the capability frontier far enough to encompass tasks as complex as biological weapon attack planning.” And again, Amodei forecasts “a substantial risk” that “systems...we expect to see within 2-3 years” will be able to contribute to biological attacks. With AI capabilities advancing rapidly, the United States government should act promptly to regulate open foundation models so as to avoid the risks created by more powerful models with widely available weights.

*5. What are the safety-related or broader technical issues involved in managing risks and amplifying benefits of dual-use foundation models with widely available model weights?*

*a. What model evaluations, if any, can help determine the risks or benefits associated with making weights of a foundation model widely available?*

**Evaluations that fully elicit the capabilities of a foundation model are necessary in order to make an informed decision about the risks and benefits of making its weights widely available.** Because nobody can expect to control how a model is fine-tuned once its weights are widely available, evaluations that focus on the foundation model’s safety properties, while useful for other purposes, cannot provide any assurance about the risks of making its weights widely available.

*(5) b. Are there effective ways to create safeguards around foundation models, either to ensure that model weights do not become available, or to protect system integrity or human well-being (including privacy) and reduce security risks in those cases where weights are widely available?*

A working paper by Nevo et al. (2023)[[6]](#footnote-7) discusses a security framework for protecting nonpublic model weights from malicious actors. The authors note that “[i]n some (though not all) cases, acquiring the weights could allow a malicious actor to make use of the full model at a tiny fraction of the cost of training it,” and thus they “believe that weights should be given special attention in a lab’s security strategy.” They provide several broad recommendations for security strategies to be adopted by AI labs, review possible attack vectors, and outline “five security levels...defined to thwart most attacks from an increasingly capable actor category.” This working paper is part of a study for which a full report is expected to be “published in early 2024” – I recommend that NTIA review the full report when it becomes available.

*7. What are current or potential voluntary, domestic regulatory, and international mechanisms to manage the risks and maximize the benefits of foundation models with widely available weights? What kind of entities should take a leadership role across which features of governance?*

*a. What security, legal, or other measures can reasonably be employed to reliably prevent wide availability of access to a foundation model’s weights, or limit their end use?*

**The United States government should impose a moratorium on making model weights widely available for AI models substantially more capable than those available as of early 2024.** The United States should work with other governments around the world to promote this policy internationally, but this policy would be worthwhile even if only implemented by the United States. The United States government should also develop regulations to hold AI model developers accountable for harm caused by misuse of the models they develop, especially when they facilitate this misuse by making model weights widely available. This could take the form of civil liability for damages as well as regulatory proceedings.

While it may be outside the scope of this request for comment, **I would also support an international agreement to indefinitely pause development and training of more capable AI models until reliable strategies to avoid the risks from such models are developed**. I would encourage the United States government to lead diplomatic efforts to reach such an agreement. The risks from more capable models – including potentially catastrophic risks from goal misgeneralization and loss of control of powerful AI systems, as well as risks from misuse of open models – are too serious to accept until substantial advances are made in AI safety research.

*(7) b. How might the wide availability of open foundation model weights facilitate, or else frustrate, government action in AI regulation?*

As discussed in response to question (2)(d)(i) above, Qi et al. show that both open and closed models will more readily produce harmful output after fine-tuning. If fine-tuning of closed models proves to present risks unacceptable to the United States government, this can be addressed through regulation of providers of closed models. If fine-tuning of open models presents such risks, **effective regulation and enforcement may not be possible once the models have been released with widely available weights, due to the vast number of actors around the world capable of obtaining and fine-tuning open models.**

I appreciate NTIA’s consideration of these comments.

Respectfully,

Samuel Duncan

Somerville, MA

1. Lermen, Simon, Charlie Rogers-Smith, and Jeffrey Ladish. “LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B.” *arXiv preprint arXiv:2310.20624* (2023). https://arxiv.org/abs/2310.20624. [↑](#footnote-ref-2)
2. Amodei, Dario. “Written Testimony for a Hearing on ‘Oversight of A.I.: Principles for Regulation.’” U.S. Senate, Judiciary Committee, Subcommittee on Privacy, Technology, and the Law. 118th Congress. 25 July 2023. https://www.judiciary.senate.gov/imo/media/doc/2023-07-26\_-\_testimony\_-\_amodei.pdf. [↑](#footnote-ref-3)
3. Qi, Xiangyu, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. “Fine-tuning aligned language models compromises safety, even when users do not intend to!” *arXiv preprint arXiv:2310.03693* (2023). https://arxiv.org/abs/2310.03693. [↑](#footnote-ref-4)
4. See https://docs.anthropic.com/claude/docs/glossary#fine-tuning: “Our API does not currently offer fine-tuning, but please ask your Anthropic contact if you are interested in exploring this option.” [↑](#footnote-ref-5)
5. Mouton, Christopher A., Caleb Lucas, and Ella Guest, The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. Santa Monica, CA: RAND Corporation, 2024. https://www.rand.org/pubs/research\_reports/RRA2977-2.html. Note that while some of the red-team researchers involved in this study attempted to “jailbreak” the language models which they were assigned to work with, the study “did not examine fine-tuned LLMs or LLMs without any guardrails.” [↑](#footnote-ref-6)
6. Nevo, Sella, Dan Lahav, Ajay Karpur, Jeff Alstott, and Jason Matheny, Securing Artificial Intelligence Model Weights: Interim Report. Santa Monica, CA: RAND Corporation, 2023. https://www.rand.org/pubs/working\_papers/WRA2849-1.html. [↑](#footnote-ref-7)