**Before the**
**NATIONAL TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION**
**Washington, DC 20554**

| | | |
|---|---|---|
| In the matter of | ] | |
| Dual Use Foundation Artificial Intelligence Models | ] | |
| With Widely Available Model Weights | ] | NTIA–2023–0009 |

**COMMENTS OF THE CENTER FOR LAW & AI RISK AND LEGAL SCHOLARS**

We, the directors of the Center for Law & AI Risk, write in support of sensible restrictions on the open-sourcing of powerful frontier generative AI systems. While we generally favor openness in technological development, we believe that frontier AI is different. Near-future AI systems threaten to cause—via intentional misuse, accidents, or autonomous action—large-scale harms to human life, limb, and freedom. We believe that there is a compelling case for systemic regulation of AI systems.[1] The dissemination of open source models being one part of this composite regulatory approach.

Some of these risks are already evident in the capabilities of current-generation non-open-sourced systems. Consider, for example, a recent paper from researchers at the University of Illinois showing that GPT-4 "can autonomously hack websites, performing complex tasks," including independently "finding vulnerabilities in real-world websites" and performing "SQL union attacks."[2] All "without human feedback." If such systems were used to access and disable vital infrastructure, it would cause significant harm.

Such dangerous capabilities in AI systems often emerge suddenly, with successive generations of systems displaying discontinuous abilities.[3] Consider: while GPT-4 could be used to hack 73% of the secure systems the aforementioned study presented it with, GPT-3.5 could hack just 6.7%. Current open-source models are less capable; *none* could autonomously hack any of the test systems.

---

[1] *See, e.g.,* Yonathan Arbel et al., *Systemic Regulation of Artificial Intelligence*, Az. St. L.J. (forthcoming 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4666854.
[2] Richard Fang et al., LLM Agents can Autonomously Hack Websites, arxiv (Feb. 16, 2024), https://arxiv.org/pdf/2402.06664.pdf.
[3] Jason Wei et al., Emergent Abilities of Large Language Models, Transactions on Machine Learning Research (Aug. 2022), https://openreview.net/pdf?id=yzkSU5zdwD.

Cybersecurity is not the only avenue of threat from advancing AI capabilities. There are also, for example, risks from bioterrorism and chemical attacks. Systems based on current-generation language models can readily coach non-experts in the step-by-step process of producing explosives and poisons.[4] They can identify known pandemic viruses and instruct non-experts on the best ways to obtain live samples, including by identifying synthesis labs with lax safety protocols.[5]

True, most of this information about chemistry and virology is already available online, meaning that current-generation models are likely not adding much risk. However, other kinds of currently-existing AIs provide strong evidence that this could change in the near future. For example, certain "narrow" AI systems already have superhuman abilities[6] to generate novel proteins with specific,[7] and potentially harmful, biological functions. Others have a superhuman ability to invent new harmful molecules,[8] including ones much more deadly than the chemical weapon, VX. Likewise, recent advances in embodied AI suggest that soon, such systems will be able to directly control the machines required to produce such harmful substances, making them without the need for any human intervention or expertise.[9]

Open sourcing powerful systems raises these risks significantly. In closed AI systems, dangerous behavior can be mitigated via a suite of tools. Among them are "alignment" methods, like reinforcement learning from human feedback,[10] which train models to refuse to comply with dangerous requests. Closed systems' behavior can also be influenced via system prompts,[11] which are appended before a user's prompt and instruct an AI to behave only in safe ways. And if a closed AI system nevertheless attempts to

[4] Andres M. Bran et al., ChemCrow: Augmenting large-language models with chemistry tools, arxiv (Apr. 11, 2023), https://arxiv.org/abs/2304.05376.
[5] Emily H. Soice et al., Can large language models democratize access to dual-use biotechnology?, arxiv (Jun. 6, 2023), https://arxiv.org/abs/2306.03809.
[6] John Jumper et al., Highly accurate protein structure prediction with AlphaFold, nature (July 15, 2021), https://www.nature.com/articles/s41586-021-03819-2.
[7] Ali Madani et al., Large language models generate functional protein sequences across diverse families, nature (Jan. 26, 2023), https://www.nature.com/articles/s41587-022-01618-2
[8] Fabio Urbina et al., Dual use of artificial intelligence powered drug discovery, nature (Mar. 7, 2022), https://www.nature.com/articles/s42256-022-00465-9
[9] Benjamin Burger et al., A mobile robotic chemist, nature (July 8, 2020), https://www.nature.com/articles/s41586-020-2442-2
[10] Yuntao Bai et al., Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, arxiv (Apr. 12, 2022), https://arxiv.org/abs/2204.05862v1
[11] System prompts, Anthropic (last accessed Mar. 15, 2024), https://docs.anthropic.com/claude/docs/system-prompts

engage in unsafe behavior, its outputs can be automatically filtered or the user's session can be automatically ended. Perhaps the most critical point is that open-source systems have, in theory at least, a centralized shutdown button, with identifiable actors subject to domestic regulation.

None of this is possible with open-sourced systems. Open-sourced systems can have their alignment stripped away at trivial cost.[12] They can be merged and fine-tuned to produce novel behaviors and capabilities. They are disseminated widely in remote locations, effectively insulated from any enforcement action. They can be used to research adversarial attacks and vulnerabilities in deployed systems by state actors. They can be improved over time using novel breakthroughs and outputs from other models. And there is simply no way to undo the dissemination of a model discovered to be dangerous.

To be clear, we do not oppose the open sourcing of all generative AI models. Open sourcing models is a force for equity in a technology that is built on the joint contributions of human society. We also do not deem current generation open-source models to be risky, and it may be the case that even today's frontier models–like GPT-4 or Claude Opus–are not sufficiently dangerous. Moreover, open sourcing weaker models has proven helpful in research into safety and interpretability. Open sourcing can do much good.

What seems utterly clear to us, however, is that nothing guarantees safety going forward. Judgments about how new, powerful frontier models may be used must be made on a model-by-model basis, and only after carefully understanding what each system can do or be adapted to do, if it falls into the wrong hands. This, we think, is an anodyne observation about powerful technologies more generally: The fact that one nuclear power plant design is certified as safe does not imply that all of them will be. Nor is everyone in the world sufficiently responsible to independently operate a nuclear plant, even one with a certified safe design. Open sourcing poorly understood frontier models imposes a risk, and within a

[12] Xiangyu Qi et al., Fine-Tuning Aligned Language Models Comprises Safety, Even When Users Do Not Intend To!, arxiv (Oct. 5, 2023), https://arxiv.org/pdf/2310.03693.pdf

comprehensive system of regulation of frontier AI systems, limitations on open-sourcing is a critical component.

**Respectfully submitted,**

**Peter N. Salib**
Assistant Professor
University of Houston Law Center

Director
Center for Law & AI Risk

**Yonathan Arbel**
Silver Associate Professor
University of Alabama

Executive Director
Center for Law & AI Risk

**Kevin Frazier**
Assistant Professor
St. Thomas University College of Law

Director
Center for Law and AI Risk