Comment to the National Telecommunications and Information Administration

Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights

March 26, 2024

Regulations.gov Docket No. NTIA-2023-0009

NTIA Docket No. 240216-0052

Max Gulker, Ph.D.[1]

Spence Purnell[2]

Richard Sill[3]

[1] Senior Policy Analyst, Reason Foundation
[2] Director of Technology Policy, Reason Foundation
[3] Technology Policy Intern, Reason Foundation

Introduction

On behalf of Reason Foundation[4], we respectfully respond to the National Telecommunications and Information Administration's ("NTIA") request for comments on the risks and benefits of "dual-use foundation artificial intelligence models with widely available weights," or open foundation AI models.[5] Reason Foundation is a national 501(c)(3) public policy research and education organization with expertise across a range of policy areas, including emerging technology.

These generative AI models entered widespread public awareness following the release of OpenAI's first version of ChatGPT in late 2022.[6] Open-foundation AI models with publicly available weights, a particular category of generative AI model explained further below, allows greater or full access to the inputs of models so that others may customize or create applications with them.

The NTIA is not a policymaking body but will issue an advisory report to the president on potential policy in this area. In the thus far limited public debate some have proposed preemptive measures to limit the openness of AI foundation models or their proliferation.[7]

Our comment discusses several of the questions posed by the NTIA, particularly Question #3 on the potential benefits of open foundation models, and Question #2 on the potential risks. We argue that allowing model developers to freely choose and innovate along dimensions of openness may be indispensable in realizing many of the technology's benefits, without any evidence of specific risks over and above those of more closed approaches.

The NTIA should not at this time recommend a policy aimed at restricting open foundation AI models with publicly available weights.

What are open foundation AI models?

OpenAI publicly released ChatGPT in late 2022.[8] It is therefore not surprising that the issue of open versus closed foundation AI models remains very much under development and has not yet led to widespread public debate.

For AI foundation or large language models ("LLMs," of which the ChatGPT releases are examples), "open" versus "closed" hinges on the degree of access allowed by model developers to the computer code of the model itself, the data the model was trained on, and the numerical weights assigned to the data when producing output such as text or pictures. Building on earlier work,

[4] See About Reason Foundation, https://reason.org/about-reason-foundation/

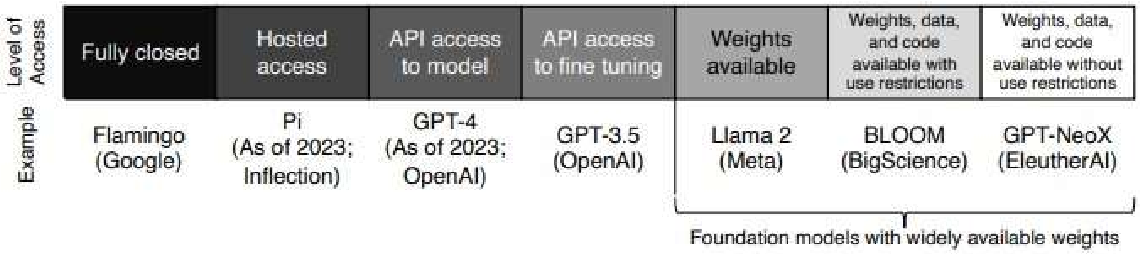
[5] National Telecommunications and Information Association, "Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights," *Request for Comment,* 26 Feb 2024. https://www.federalregister.gov/documents/2024/02/26/2024-03763/dual-use-foundation-artificial-intelligence-models-with-widely-available-model-weights

[6] "Chat-GPT," Wikipedia page accessed 26 March 2024. https://en.wikipedia.org/wiki/ChatGPT

[7] Sayash Kapoor et al., "On the Societal Impact of Open Foundation Models," arXiv:2403.07918 [cs.CY], 27 Feb 2024. https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf

[8] "Chat-GPT," Wikipedia page accessed 26 March 2024. https://en.wikipedia.org/wiki/ChatGPT

researchers in late 2023 categorized current and anticipated generative AI models on an approximate open/closed continuum:[9]



Level of Access	Fully closed	Hosted access	API access to model	API access to fine tuning	Weights available	Weights, data, and code available with use restrictions	Weights, data, and code available without use restrictions
Example	Flamingo (Google)	Pi (As of 2023; Inflection)	GPT-4 (As of 2023; OpenAI)	GPT-3.5 (OpenAI)	Llama 2 (Meta)	BLOOM (BigScience)	GPT-NeoX (EleutherAI)

The figure shows that different foundation model developers have already experimented with an array of degrees of openness. ChatGPT, for example, has allowed public access to some model features but not model weights.

Full access to model weights in addition to other inputs would allow end users to customize the data on which LLMs are trained or adjust the weights of already-existing data. Anyone lacking the means to create their own specialized generative AI models from the ground up could customize open-foundation models. Third-party developers would likely produce an array of applications of interest to end users. As of March 2024, we have yet to see this rapidly developing aspect of the technology fully deployed.

However, those less familiar with the technology need not view these dimensions of openness for AI models as unusual relative to other uses of computer software and applications. There is now a long history of closed and proprietary computer software existing in the same ecosystem as a vibrant open-source movement.[10] Consumers now understand the potential of third-party developers through the case of smartphone apps, where Apple and the Android ecosystem have been two examples of intermediate cases on different points of a similar open/closed continuum.[11]

Benefits of open foundation AI models (NTIA question #3)

“What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?”

Consumers, end-users, and third-party developers are not merely the beneficiaries of innovation for novel technologies like generative AI, but contribute indispensably to the ongoing process of innovation itself. They provide feedback for new ideas through the market mechanism, informing

[9] Rishi Bommasani et al., “Considerations for Governing Open Foundation Models,” HAI Policy and Society Issue Brief, December 2023. https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf

[10] Tozzi, Christopher, For Fun and Profit: A History of the Free and Open Source Software Revolution, MIT Press, 2017.

[11] Purnell, Spence and Grayce Burns, “The pitfalls of regulating app stores,” Reason Foundation. https://reason.org/commentary/the-pitfalls-of-regulating-app-stores/

innovators in a continuous process of many small interactions that generally steers technology toward beneficial uses that a room full of the greatest minds could never anticipate.[12]

The reliably unpredictable development of internet applications provides numerous examples of consumers helping guide innovations to places few people foresaw. In the early days of mp3 downloads, both legal and through illicit file-sharing, companies assumed consumers would place a high value on owning files on their hard drives. But through the process of more reliable internet technology and growing consumer comfort with the model, streaming emerged as the truly disruptive force in the music industry.[13] Social media platforms present similar examples. Facebook was created for students at a single university before widespread adoption.

The role of consumers and end-users is especially important in the development of foundational technologies—advances beneficial through a wide array of applications rather than a single-use case. Like the internet technology noted above, generative AI will almost certainly derive its benefits by helping people perform tasks, access information, and communicate ideas.

Third-party application developers add another link in the chain between AI model and end-user but do not alter the fundamental truth that the most useful innovative ideas emerge from an evolutionary give-and-take process between developers and users of various types. In fact, such third parties would likely increase the speed and efficacy of the market process in fueling innovation.

Open-foundation AI models would allow for niche specialization by various groups who may lack the resources and know-how to create their own generative AI model. The NTIA request for comment makes frequent mention, for example, of medical and academic researchers.

Consider how a small group of cutting-edge researchers in a field might make use of an open-foundation AI model. Broadly speaking, generative AI models are trained over very large sets of language from the internet. Researchers in any number of fields might want to place different weights on different portions of academic literature, customizing what the generative AI model might do for them. The true benefits of open approaches to technology often lie in the process that ensues as adjustments and interactions take place up and down the chain between model developers, intermediaries such as application developers, and consumers. This is when millions of minds lead us to use cases that a few very smart people in a room could never anticipate.[14]

A scenario with only closed generative AI models and end users would likely lead to some benefits of this creative process going unrealized. When some AI models are open to varying degrees, it allows developers and end users to unpack and understand what makes such models work and provides the original model builders with an extra layer of feedback from sophisticated third-party

[12] Ridley, Matt, How Innovation Works, Harper-Collins, 2021.

[13] Ganz, Jacob, "How Streaming is Changing Music," National Public Radio, 1 June 2015. https://www.npr.org/sections/therecord/2015/06/01/411119372/how-streaming-is-changing-music

[14] Gulker, Max. "Calls to regulate AI ignore how consumers help shape innovationhttps://reason.org/commentary/calls-to-regulate-ai-ignore-how-consumers-help-shape-innovation/

developers. This makes many more people than the model's proprietary creators able to experiment, debug, and innovate.

This does not mean that regulators should require certain types of open access or otherwise tilt the playing field in favor of more openness in generative AI models. Allowing developers to experiment with different degrees of openness and means of providing data is a critical part of a robust market-based ecosystem for generative AI. As mentioned earlier, the Apple "closed" software system has been very succesful at incorporating user feedback and iterating successful versions. Requiring all models to be open access puts this model at risk. The feedback they receive from a well-functioning market will better allocate resources to models of varying openness than premature regulatory guesswork.

In a June 2023 public comment to the NTIA, Neil Chilson and Will Rinehart of the Center for Growth and Opportunity emphasize the indispensability of a market-based system of governance and accountability for generative AI more broadly:[15]

> "An accountability ecosystem for software already exists and has proven highly effective. It is polycentric in that it is layered and is comprised of business-to-business and business-to-consumer markets, reputational markets, and financial markets, all backed by generally applicable laws and norms."

This position does not preclude regulation or a role for the public sector entirely but comes with a warning that regulators should only step in when clear and well-established market failures are observed to take place.

This is especially important advice with highly novel technology like generative AI. Because the underlying technology and its use cases are still extremely early in their processes of development, creating preemptive rules, licensing regimes, or prohibitions—without indication of specific risks and market failures—would not be advisable.

Risks of open foundation AI models (NTIA question #2)

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

Though AI model developers have begun experimenting along dimensions of openness, we have not yet seen a fully operational open-foundation AI model with modifications and applications created by third parties. Thus far, computer scientists have not found clear or compelling theoretical evidence that points to open foundation AI models having new or greater risks because of their openness.

Writing in February 2024, a team of computer scientists and experts from related disciplines spanning academia and industry examined the theoretical foundations of open foundation AI models. They considered several commonly discussed categories of threats, including

[15] Chilson, Neil and Will Rinehart, "Public Interest Comment on the National Telecommunications and Information Administration (NTIA) AI Accountability Policy," The Center for Growth and Opportunity at Utah State University, 12 June 2023. https://www.thecgo.org/wp-content/uploads/2023/06/NTIA-comments-on-AI-accountability_03.pdf

disinformation, cyberattacks, and scams directed at individuals. In this newly burgeoning field, they find that “current research is insufficient to effectively characterize the marginal risk of open foundation models relative to pre-existing technologies,” but stress the need for more empirical research as those data become available.[16]

In a December 2023 issue brief published jointly by AI labs at institutions including Stanford University and Princeton University, a team of researchers stated that:[17]

> “While open foundation AI models are conjectured to contribute to malicious uses of AI, the weakness of evidence is striking. More research is necessary to assess the marginal risk of open foundation models.
>
> Policymakers should also consider the potential for AI regulation to have unintended consequences on the vibrant innovation ecosystem around open foundation models.”

These findings should give the NTIA particular pause in recommending specific regulatory restrictions on open foundation models before the technology can appropriately develop in the market.

Issues of equity in open foundation AI models (NTIA questions #2b and #3c)

“Could open foundation models reduce equity in rights and safety-impacting AI systems (e.g., healthcare, education, criminal justice, housing, online platforms, etc.)?”

“Could open model weights, and in particular the ability to retrain models, help advance equity in rights and safety-impacting AI systems (e.g., healthcare, education, criminal justice, housing, online platforms etc.)?”

The distinction between open and closed will likely not make a material difference to how AI impacts equity and rights.

Both models present a low risk of discrimination because these behaviors are already illegal whether carried out by humans or AI and can be enforced the same way. If an AI system discriminates against a protected class during a hiring process, citizens would still have grounding to bring suit. Whether the model is open or closed, the violation can only occur once that system has made a determination. Using a data collection or reporting system would help detect illegal patterns in AI decision-making.

Just like a human operator, an AI can be trained to follow certain rules but still needs to be monitored for accuracy. However, this doesn’t necessarily imply that regular AI source code audits are necessary, as current human-based hiring practices can discriminate but are not regularly audited by the government. An AI system could operate under the current complaint-based system

[16] Sayash Kapoor et al., “On the Societal Impact of Open Foundation Models,” arXiv:2403.07918 [cs.CY], 27 Feb 2024. https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf

[17] Rishi Bommasani et al., “Considerations for Governing Open Foundation Models,” HAI Policy and Society Issue Brief, December 2023. https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf

where citizens can file a complaint with the government if they believe discrimination is occurring. If enough complaints and evidence pile up, an investigation is initiated and the legal/court process handles the rest. These types of complaint-based systems exist in all the areas raised by the request (healthcare, education, hiring, etc.) and would continue to function normally without any new regulations whether and AI system was open or closed.

In addition to outcome-based evaluation, approximating the source code to both and closed source is similar. While open-source models have source code that can be more easily inspected, even closed systems can be interacted with enough to "reverse engineer" the source code to near exact approximation. In fact, ChatGPT has already been reverse-engineered and released onto the web several times, demonstrating this capability.[18] In the case of image and video generative AI, the recent rise of "deepfakes" has already spawned a counter industry of "deepfake detectors," which tell users whether images and videos have been generated or heavily edited by AI.[19]

Public policy should focus on working with industry to standardize and deploy AI detection and evaluation systems in appropriate areas. Developing technologies that evaluate AI decision-making will be critical to ensuring appropriate use.

Instead of seeking out specific regulations or trying to prevent broadly defined negative outcomes, policy should help cultivate and develop industry standards such as monitoring and reporting of AI systems for things that are already illegal.

Conclusion

Generative AI—including the features related to openness—is the most recent in a now long list of advances in information and communication technology that have sparked concern and debate about a similar set of risks. These risks are fundamentally tied to the technologies' vast benefits: enabling even individual users to access information, communicate, and create content in unprecedented ways. With new tools, individual users gain greater capacities to misinform, victimize others, or commit crimes.

We do not yet have any reason to believe open foundation generative AI models represent more than continued progress in what such technology enables us to do. We do have reason to believe that hasty regulation on this topic has a high chance of preventing us from realizing some of the benefits of this novel technology.

These considerations suggest that tight regulation, especially this early in the development of open foundation AI models, is not needed and would likely be counterproductive. Instead, the focus should be on markets, innovation, and free enterprise, where consumers and builders of AI models alike learn to set standards and mitigate risks through time.

[18] Y Combinator, accessed 26 March 2024. https://news.ycombinator.com/item?id=35742685

[19] MIT Media Lab, "Detect DeepFakes: How to counteract misinformation created by AI," accessed 26 March 2024. https://www.media.mit.edu/projects/detect-fakes/overview/