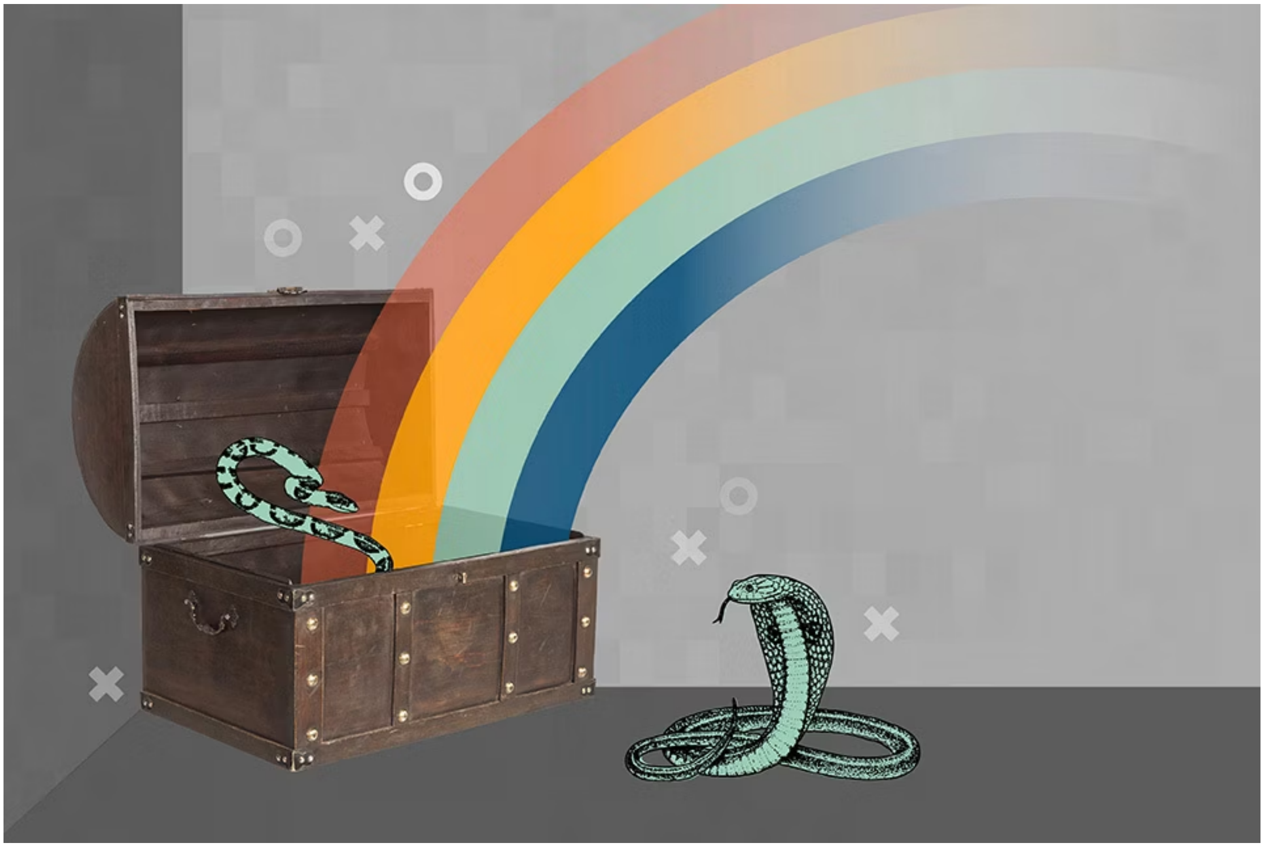
THIRD WAY

MEMO *Published March 7, 2024 • 9 minute read*

What is "Open" AI?



Mike Sexton
Senior Policy Advisor for Cyber and
Artificial Intelligence
@MikeESexton

The risks and benefits of "open-source" or "open-foundation" AI are a key issue in the AI safety debate. However, understanding these terms and risks is deceptively tricky. So what do "open-source" and "open-foundation" AI mean, and why do experts disagree about the opportunities and dangers they pose?

Takeaways

- “Open-foundation AI” refers to AI models like Meta’s Llama 2 or Google’s Gemma that the public can download and use without direct control from the original developer.
- An AI model’s openness is a gradient, not a binary. A model can be open-foundation without being transparent about what data it is trained on, for example.
- To those who think advanced AI models are inherently extremely dangerous, it follows intuitively that making them open-foundation is even worse.
- Other experts disagree. Open-foundation AI, they argue, democratizes the gains of AI development whereas closed-foundation AI concentrates them at the top.

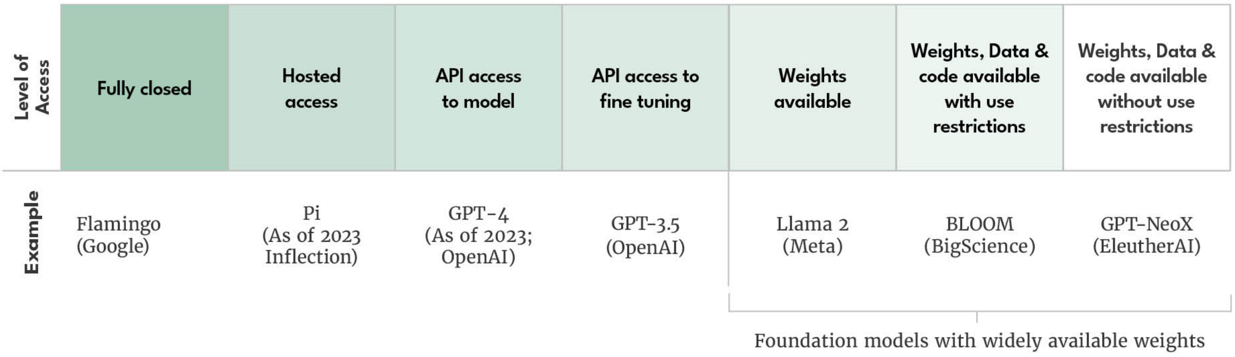
What is “Open” AI?

To begin unpacking this, let’s first consider the term “open-source.” When we say Mozilla Firefox is an “open-source” web browser, what we mean is its source code is freely available for anyone to download, tinker with, and redistribute (subject to the terms of its open-source license). Mozilla began when Netscape open-sourced its browser’s codebase in 1998, [1] and Netscape went on to build its browsers atop Firefox source code beginning in 2004. [2]

While OpenAI encourages users to customize their own bespoke AI programs on top of ChatGPT as a foundation model, [3] ChatGPT is not open-source because OpenAI still controls how it gets used downstream. Meta’s large language model

(LLM) Llama 2, on the other hand, is free to download and run independently, [4] and Mark Zuckerberg has promised its successor Llama 3 will be, too. [5] Similarly Google has recently released Gemma, a lightweight open AI model based on the same code and research as Google's chatbot Gemini. [6]

AI experts in the private sector [7] and academia [8] classify "open" AI not as a binary but a gradient:



Level of Access	Fully closed	Hosted access	API access to model	API access to fine tuning	Weights available	Weights, Data & code available with use restrictions	Weights, Data & 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)

Foundation models with widely available weights

Source: Bommasani, Rishi, et al. "Issue Brief Considerations for Governing Open Foundation Models | Stanford HAI." Stanford University Human-Centered Artificial Intelligence, 13 Dec. 2023, https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models. Accessed 23 January 2024.

THIRD WAY

When we say an AI foundation model is "open," we are mostly referring to its "weights," also known as "parameters." LLMs are neural networks, meaning they have neurons analogous to those in human brains, and the weights represent the relationships between the neurons. Developers set the LLMs' weights by exposing them to human language until they can predict and generate it as intuitively as people can. So if you tell an LLM "lions and tigers and bears," its uses its trained weights to predict as well as any person what comes next: "oh my!"

While ChatGPT is open to the public, its weights are not. A ChatGPT-based AI like Dean.Bot, which was made to imitate longshot presidential candidate Dean Phillips, works through something called an application programming interface or API. This means it runs through OpenAI's servers—not the developer's own—and OpenAI can still shut it down if it violates its terms of use (which it did). [9]

But an LLM's weights still do not tell us everything. In fact, because the parameters typically number in the billions or trillions, they are almost as inscrutable as the neurons in a human brain. Additional context can be provided by the LLM's training data: what books, websites, and other texts has it read to understand human language? This can provide insights into the relationships between data inputs and model outputs, but it also creates a competitive disadvantage for the developer.

Because of this conceptual ambiguity, experts more often refer to "open/closed-foundation LLMs" rather than "open/closed-source LLMs." "Foundation" underscores that ChatGPT, Llama, and Gemma are foundations for more specialized AIs, while "open/closed" signifies whether the developer (Meta/OpenAI/Google) retains control over how the specialized AI is used downstream.

Is Open-Foundation AI Bad?

AI developers and experts disagree over whether open-foundation AI is harmful or beneficial. Ilya Sutskever, cofounder and Chief Scientist of OpenAI, is perhaps more intimately familiar with this issue than anyone. OpenAI was founded as a nonprofit in 2015 to make artificial general intelligence (AGI) a freely accessible public good, but today it is better known for its secrecy than its transparency. [10]

Sutskever has said, "We were wrong. Flat out, we were wrong. If you believe, as we do, that at some point, AI—AGI— is going to be extremely, unbelievably potent, then it just does not make sense to open-source. It is a bad idea... I fully expect that in a few years it's going to be completely obvious to everyone that open-sourcing AI is just not wise." [11] In other words, if we're worried a rogue AGI could wreak havoc on society, it is absolutely critical that we build any newly advanced AIs to be as contained as possible.

The demise of the Dean.Bot, the ChatGPT-based chatbot of Dean Phillips, can be considered a low-stakes cautionary tale. [12] Matt Krisiloff, one of OpenAI's first employees, designed Dean.Bot with fellow entrepreneur Jed Somers to speak directly to voters as the Democratic presidential candidate. However, OpenAI determined Dean.Bot violated its usage policy against political campaigning and shut it down.

This story would have unfolded differently if Dean.Bot were built on Gemma or Llama 2 instead. If Krisiloff and Somers had used Llama 2 as a foundation model, Meta could suppress it on its platforms, but it could not shut it down. Dean.Bot needs two things to exist—its code and the servers to run it—and because Meta open-sourced Llama 2, it could control neither.

Dean.Bot is a far cry from the Terminator, but it captures the essence of concerns about open-foundation AI: if our new superpowered AI programs are so dangerous, shouldn't the developers of the foundation model retain the ability to pull the plug if their customers put it to bad use? What if someone used an open-foundation LLM to build a chatbot with a comprehensive knowledge of how to build chemical weapons?

Is Open-Foundation AI Good?

Researchers at the Stanford Institute for Human-Centered Artificial Intelligence (HAI) are skeptical of this risk. Rather than imagining the dangers of open-foundation AI in a vacuum, they say we must consider them relative to the dangers of closed-foundation AIs and other existing tools like web search. An open-foundation LLM tailored by a user to help build chemical weapons is dangerous, yes, but the danger is marginal because it would largely derive its knowledge from websites people can already access with web search.

Just like open-source software, open-foundation AI models have some important advantages that observers fixated on risk may overlook. Stanford HAI's scholars note three:

1. **Distributing power.** Whether they're monopolies or not, the roles Microsoft, Meta, Google, Apple, and Amazon play in our daily lives are undeniably enormous. When we use tools built on closed-source ChatGPT, we tacitly cede more control to OpenAI and Microsoft by extension. When we use tools based on Llama 2, Meta may get nominal credit, but it reaps nothing material.

2. **Catalyzing innovation.** When Netscape open-sourced the code for its browser, it allowed not just the creation of Firefox, but of countless other browsers that have been built *with* Firefox. Open-source software and open-foundation AI can be thought of as rising tides, lifting all the boats (that is, programs) built on top of them.

3. **Ensuring transparency.** Closed-foundation AI models may be safer in some ways, but they are also inherently more opaque. [13] That is, in fact, its own form of risk. AI's harms are insidious, and the less transparent developers are about how foundation models are built, the harder it will be to root out problems like racial and gender bias.

The Stanford HAI researchers recognize certain areas where open-foundation AIs are indeed riskier than closed-foundation models, such as disinformation, bioweapons, hacking, and various forms of illicit pornography. In most of these areas, however, they argue that the risks can and should be addressed through other chokepoints. For example, if Dean.Bot had been built on Llama 2 and survived, Meta could still classify it as a user policy violation and suppress it on Facebook, Instagram, and WhatsApp, and flag other networks like X to do the same.

AI-generated illicit pornography, specifically nonconsensual intimate imagery (NCII) and child sexual abuse material (CSAM), is an exception. There is little governments or companies could do to spot and stop someone using an open-foundation AI on their own servers to make NCII and CSAM if they strictly distributed it via encrypted apps. Strong counter-illicit finance mechanisms could, at most, eliminate the profit motive for this criminal activity, but it would remain a challenge for open-foundation AI developers.

Conclusion

AI risk is critically important but inherently nebulous, and nowhere is this truer than with respect to open-foundation AI models. Some leading AI developers consider open-foundation AI extremely dangerous, but this position also happens to be competitively advantageous for their own closed-foundation AI models. Proponents of open-foundation AI cite the massive public benefits of open-source software, but recognize some of its dangers may be all but impossible to contain. It is essential that we take a hard-nosed approach to assess all forms of AI risk, and while it may make sense to err on the side of caution, we must always remain willing to consider countervailing evidence and challenge our own assumptions and judgments.

Glossary

AI: artificial intelligence, in particular LLMs in the context of this paper

Open-source: software whose source code is publicly accessible to modify and redistribute

Open-foundation: AI models whose weights are publicly accessible to modify and redistribute

Weights (aka parameters): the relationships between the neurons in a neural network, represented as numbers

Neural network: an AI modeled after the human brain with neurons that “learn” through exposure to text, images, or other media

Llama: Meta’s open-foundation AI models

Gemma: Google’s open-foundation AI models

Gemini: Google’s closed-foundation AI model

Large language model (LLM): a kind of neural network trained on a large amount of text

Application programming interface (API): a mechanism for two computer programs to communicate with each other, including across the web with remote servers

Nonconsensual intimate imagery (NCII): this can refer either to authentic images (revenge porn) or AI-generated images (deepfake porn)

Child sexual abuse material (CSAM): preferred term to “child pornography”

TOPICS

TECHNOLOGY & SECURITY 11

ENDNOTES

[1] "Browser History: Epic Power Struggles That Brought Us Modern Browsers." *Mozilla*, https://www.mozilla.org/en-US/firefox/browsers/browser-history/. Accessed 23 Jan. 2024.

[2] Bishop, Alex. "First Look at Firefox-Based Netscape - MozillaZine." MozillaZine, 30 Nov. 2004, https://www.mozillazine.org/articles/article5691.html. Accessed 23 January 2024.

[3] Heath, Alex. "All the News from OpenAI's First Developer Conference." *The Verge*, 6 Nov. 2023, https://www.theverge.com/2023/11/6/23948619/openai-chatgpt-devday-developer-conference-news. Accessed 23 January 2024.

[4] Llama 2." *Meta AI*, https://ai.meta.com/llama-project. Accessed 23 Jan. 2024.

[5] Fried, Ina. *Meta Begins Training Llama 3, Reshuffles AI Responsibilities*. 18 Jan. 2024, https://www.axios.com/2024/01/18/zuckerberg-meta-llama-3-ai. Accessed 23 January 2024.

[6] Banks, Jeanine, and Tris Warkentin. "Gemma: Introducing New State-of-the-Art Open Models." Google, 21 Feb. 2024, https://blog.google/technology/developers/gemma-open-models/. Accessed 21 February 2024.

[7] Solaiman, Irene. *The Gradient of Generative AI Release: Methods and Considerations*. arXiv:2302.04844, arXiv, 5 Feb. 2023. *arXiv.org*, http://arxiv.org/abs/2302.04844. Accessed 23 January 2024.

[8] Bommasani, Rishi, et al. "Issue Brief Considerations for Governing Open Foundation Models | Stanford HAI." *Stanford University Human-Centered Artificial Intelligence*, 13 Dec. 2023, https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models. Accessed 23 January 2024.

9. Dwoskin, Elizabeth. “OpenAI Bans Developer of Dean Phillips Bot - The Washington Post.” *The Washington Post*, 22 Jan. 2024, https://www.washingtonpost.com/technology/2024/01/20/openai-dean-phillips-ban-chatgpt/. Accessed 23 January 2024.

10. Dave, Paresh. “OpenAI Quietly Scrapped a Promise to Disclose Key Documents to the Public.” *Wired*. *www.wired.com*, https://www.wired.com/story/openai-scrapped-promise-disclose-key-documents/. Accessed 24 Jan. 2024.

11. Vincent, James. “OpenAI Co-Founder on Company’s Past Approach to Openly Sharing Research: ‘We Were Wrong.’” *The Verge*, 15 Mar. 2023, https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview. Accessed 24 January 2024.

12. Miranda, Shauneen. “OpenAI Suspends Developer behind Dean Phillips Bot.” *Axios*, 21 Jan. 2024, https://www.axios.com/2024/01/21/dean-phillips-chat-gpt-ai-bot-suspension. Accessed 24 January 2024.

13. Bommasani, Rishi, et al. *The Foundation Model Transparency Index*. arXiv:2310.12941, arXiv, 19 Oct. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2310.12941. Accessed 24 January 2024.