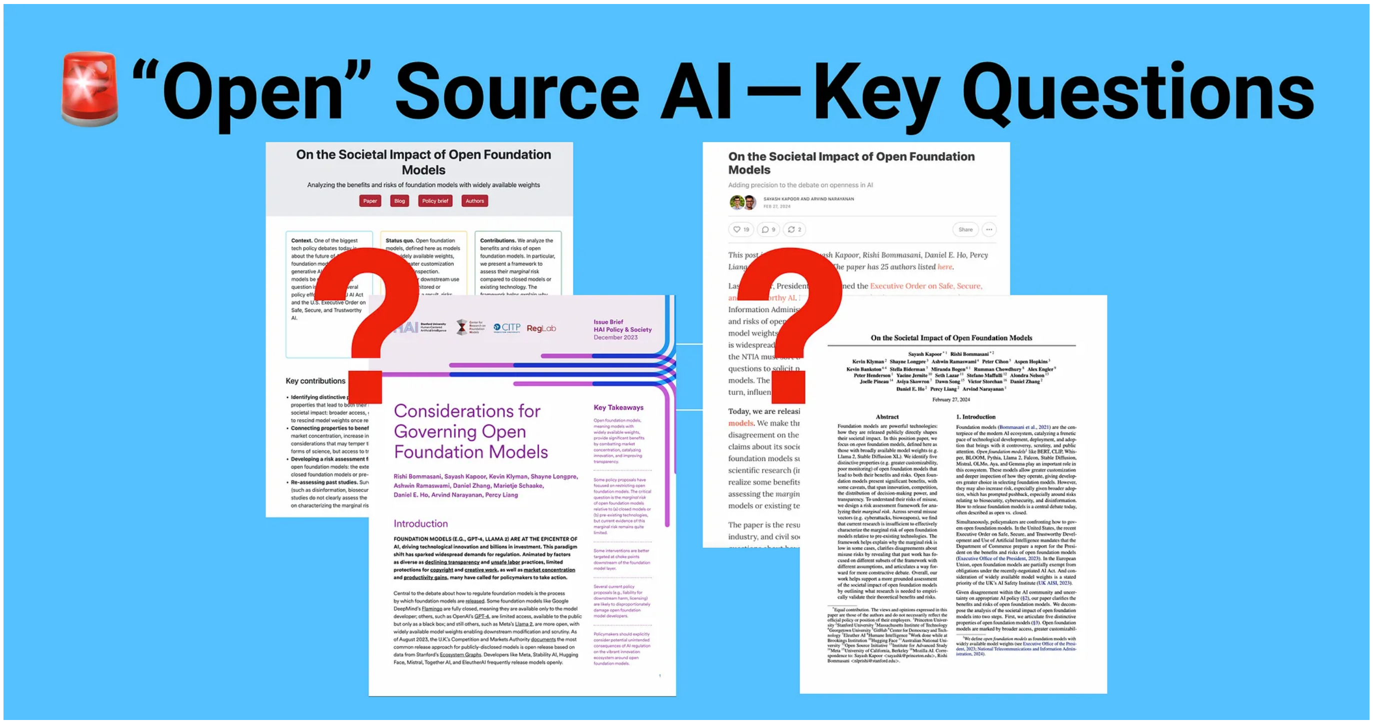
✦ Member-only story

🚨"Open" Source AI — Key Questions

Robert Maciejko

9 min read · 3 days ago

157


🚨"Open" Source AI — Key Questions
On the Societal Impact of Open Foundation Models
Analyzing the benefits and risks of foundation models with widely available weights
Paper
Blog
Policy brief
Authors
On the Societal Impact of Open Foundation Models
Adding precision to the debate on openness in AI
Considerations for Governing Open Foundation Models
Issue Brief
HAI Policy & Society
December 2023
Key Takeaways
Introduction
On the Societal Impact of Open Foundation Models
Abstract
1. Introduction
❓
❓

AI Thought Leaders on "Open" Source:

"I think the question … for …open source proponents, is … how does one stop bad actors, individuals or up to rogue states taking those same open source systems and repurposing them because they're general purpose for harmful ends? …I haven't heard a compelling, clear answer to that from proponents of just sort of open sourcing everything." — Demis Hassabis, CEO Google DeepMind

Harshil Mevasa

"By opensourcing everything, we make it easy for someone unscrupulous with access to (an) overwhelming amount of hardware to build an unsafe AI" — Ilya Sutskever, OpenAI

“The future has to be open source” — Yann LeCun, Chief AI Scientist, Meta

“Long term goals of building general intelligence, open sourcing it responsibly, and making it available…to everyone” and *“Move fast and break things”* and *“They ‘trust me.’ Dumb F*cks”*— Mark Zuckerberg, CEO, Meta

“Open” Source AI — Key Questions

“Open” Source AI is one of the hottest debated topics. Proponents say it is the only way to allow diverse access to cutting-edge tech in a culture driven by innovation. More cautious voices point out that making frontier AI available for everyone to modify for their own purposes creates unlimited risk as the tech is available simultaneously to good and bad guys.

Who’s right?

A “Position Paper” by academics from MIT, Stanford, Princeton, and others, including representatives of leading “open” source companies, tries to answer the question of whether “open” source AI is riskier than secured AI (e.g., Google Gemini, OpenAI’s ChatGPT). It comes to no clear conclusion.

On the Societal Impact of Open Foundation Models

While the Paper delivers a helpful framework for thinking about risk, it raises more questions than it answers. I’m confused about these issues and hope the authors can help provide answers.

The Paper is crucial as it is a key input into the regulation-setting process in the United States. Some of the Paper’s authors have already appeared at an event on “open” source hosted by the NTIA, a Department of Commerce entity tasked with providing input on AI to the legislative process. AI affects many areas of our lives, from healthcare to transportation, so we must carefully consider its broader societal impact.

Here are some questions I had after reviewing the Paper:

“Open Source”?: Isn’t comparing “open” source AI to open-source projects like Wikipedia or open-source software like Linux misleading? There is one set of transparent Wikipedia pages. Programmers can understand Linux code in detail. In contrast, there are unlimited AI models, and no one, not even the most prominent experts, knows precisely how they work, nor can anyone accurately predict how they will behave.

Outlawing Open Source: A report commissioned by the United States State

Department recommends that “authorities should also urgently consider outlawing the publication of the (AI model) weights, …with violations possibly punishable by jail time.” They base their findings on interviews with 200 AI leaders. Why is that recommendation wrong?

U.S. Must Move ‘Decisively’ to Avert ‘Extinction-Level’ Threat From AI, Government-Commissioned Report Says

“Open”?: Doesn’t the term “open” not fully capture the nuances of AI model transparency, especially when the training data and processes remain undisclosed? Leading open-source players like Meta have not released the detailed data used to make their models. How can anyone know what data Meta used and whether they explicitly obtained rights to use it? How can that be called “open”? (From now on, we will refer to “open” source models as **“unsecured”** models to reflect better the fact that end users can remove any guardrails and modify them at will.) In contrast, “secured AI models” include ChatGPT from OpenAI and Gemini from Google. Open AI and Google control how the models are used and can modify them if necessary.

Meta stops disclosing what data it uses to train the company’s giant AI models

Authors: The Paper’s authors include current and past employees of the most prominent unsecured AI companies, including Meta and Hugging Face. Can the Paper be considered independent? As the natural interest of unsecured AI companies would be to limit regulation and oversight, should the Paper be read with that potential bias in mind?

Authors

No “Marginal” Risk?: Yann LeCun, Chief AI Scientist at Meta, shared the Paper as definitive evidence that unsecured AI models provide benefits without additional risk. Is he right? LeCun frequently mocks other AI experts who discuss AI safety, including his fellow Turing Award winners Yoshua Bengio and Geoffrey Hinton. Does the Paper support such downplaying of risk?

Yann LeCun post regarding the Position Paper

Yann LeCun arguing with me on AI being benign

Future models: The Paper focuses on selected anecdotal evidence primarily from academic sources regarding models at least eighteen months behind current frontier models. Can legislators rely on the Paper's results to mold future-looking regulation regarding more powerful models expected to arrive?

Llama 3: Meta has announced that it would release a more advanced unsecured AI model in July — Llama 3. Should one company or one person make such a decision with no oversight? Does the Paper give any guidance on whether Llama 3 (as a short-term pending example) will have marginal risk compared to secured AI models?

Meta Llama 3 launch news

AGI: Meta's CEO Mark Zuckerberg has stated that his goal is to release AGI or superhuman AI "open" source. What evidence does the Paper give that unsecured AGI would be safe? Should one company or one person make such a decision with no oversight?

Zuckerberg AGI Goal

Public will: 71% of Americans polled by AIPI have said AI development should go slow and deliberately. Doesn't an uncontrolled and unlimited proliferation of unsecured AI models undermine the ability of legislators to fulfill the will of the people?

AIPI survey

AI Researchers: 90%+ of 2,778 published AI researchers polled voiced varying degrees of concern about a long list of potential risks of AI. Are the AI researchers wrong? Aren't unsecured AI systems more susceptible to such risk as the original developer can't control how bad actors use their systems? Does the absence of an "off" switch in unsecured AI create systematic risk?

Survey of 2,778 AI authors

Analysis of survey with respect to "open" source

Model Poisoning: Researchers have shown that bad actors can relatively easily modify AI models to have malicious traits, including back doors. Such malicious features are hard to detect and almost impossible to remove. Once released, unsecured AI models are out of their developers' control. How can we then trust any third-party AI system?

AI vulnerability: The Lurking Threat of Backdoored Models

Unleashing Zero-click Worms that Target GenAI-Powered Applications

Reliability: Sites like Hugging Face host 100,000s of unsecured AI models. Is anyone responsible for the safety of these models? Could anyone rely on ANY models to perform as stated if developers who assume no liability upload them for free distribution? How are such models more secure than any random software available anywhere online?

Over 100 Malicious AI/ML Models Found on Hugging Face Platform

Liability: The Paper argues that unsecured AI developers should have no liability. Just as manufacturers in aviation, automotive, and pharmaceutical industries are held accountable for their products' safety, shouldn't AI developers also bear responsibility for their creations? Why should AI developers get special treatment? Isn't this accountability crucial for building public trust in AI technologies? Wouldn't relieving unsecured AI developers of liability hinder innovation on safety? Hasn't strong regulation of industries like pharma enabled the acceptance of breakthrough drugs, often developed by startups? Would we want people to make untested cancer drugs in their basements for global distribution?

Access: Reputable AI users, including leading research universities, strongly support unsecured AI because it gives them unfettered access to frontier tech. Is giving such sensitive tech to bad guys at the same time it is given to these good guys the only way? Aren't there any more innovative approaches that would lower risk?

AI Agents: This year, we will see the increasing use of AI agents that can take action in IT systems on behalf of users. To date, AI models have primarily been glorified Q&A chatbots. Do the authors believe risks will grow as AI Agents increasingly autonomously take action? Can users rely on unsecured AI systems connected to AI Agents? Should they integrate them into their most sensitive IT systems? Will risks increase if users increasingly let AI decide autonomously?

Tech Companies Bet the World is Ready for 'AI Agents'

China: The United States, rightly or wrongly, has tried to limit China's access to frontier AI tech. Meta has apparently circumvented US policy by giving Chinese developers free access to the latest US AI technology. They no longer need access to the newest generation of chips as Meta has trained models using those chips on Chinese data. Is it US policy to support such activity? Should US foreign policy be set by one company or one private individual against the will of the government and legislature?

AI Wars: China, America, Open Source and the Battle for Tech Supremacy

Ideology: Chinese models trained using unsecured US AI tech from Meta are on model leaderboards and are available for download. The Chinese models reflect Chinese AI law and Communist Party doctrine. Should US legislators support the usage of such models in the US economy? Would that create any security risks?

Chinese vs. US AI — A Study in Contrasts

Military: Militaries increasingly use AI models. Does the Paper guide whether the US military can safely integrate unsecured AI systems in its operations?

Threat actors: Microsoft and OpenAI have written reports about State-linked threat actors using their AI systems in an attempt to perform malicious acts. Microsoft and OpenAI observed the behavior, reported it, and shut down the rogue groups. Isn't it true that unsecured AI developers have no such ability, given that their models are entirely out of their control once released? How can we know if bad actors such as terrorists, criminal organizations, or State-linked threat actors use unsecured AI for malicious things? Is there anything anyone can do about it if they do?

Global Cyber Threats: The Dark Side of AI Innovation

New risks: Given that unsecured AI models, once released, can no longer be controlled or patched by developers, what do the authors suggest authorities do if they identify new risks in already-released unsecured models?

Deep fakes: Unsecured models are used to create voice and visual deep fakes to terrorize individuals and commit fraud on a large scale. Is that something legislators should continue to allow? Secured models generally don't permit the generation of malicious content.

The Dark Side of Open Source AI Image Generators

Can Open Source make AI regulation impossible❓

Fake Generations: Google and Meta AI models recently generated ahistorical images. Aren't unsecured AI models subject to the same risk with the difference that developers can't patch them and users can purposely modify models to create fictional content?

Meta AI creates ahistorical images, like Google Gemini

Bias: The "Google 'woke' image" controversy highlighted concerns about AI's bias and content moderation, questioning if Silicon Valley's control over AI models is too centralized. This debate underscores the power of model tuning. Unsecured models could allow manipulation to suit hidden agendas. Are they intrinsically less biased? The call for AI that respects cultural diversity raises an important question: Is unlimited access to unsecured models the best method to achieve user-relevant AI, or do we need more intelligent, more secure solutions? Could an international ethical framework be the key to developing user-relevant, secure AI models?

Powerful unethical AI that favors AI over humans — is this OK❓

Election interference: In the New Hampshire primary this year, some voters got calls not to go to vote for President Biden. We know who did this because the people who created this robocall used a secured AI system — ElevenLabs. Isn't it true that with an unsecured AI system, bad actors could make an unlimited number of fake AI-generated robocalls customized for individuals based on scraped data?

500,000 simultaneous AI-driven Robocalls

Sora: OpenAI recently revealed Sora. Many commentators expressed concern about potential malicious uses. As a secured AI company, OpenAI significantly limits the use of its models for malicious uses. Isn't it true that no one can control how bad actors can use a Sora model released as unsecured AI?

Truth, Lies, and Deepfakes in the Age of Sora

Rights: Copyrights and other rights have been a hot topic with AI. Secured AI companies like OpenAI and Google have been licensing data for usage. Don't unsecured AI systems that have not obtained rights to data they use to train models create unlimited liability for their users? Should legislators allow that?

Meta's AI trained on Facebook and Instagram posts

What's next?

Balanced approach: Don't we need a balanced approach involving fostering innovation while implementing safeguards to mitigate risks?

Accountability: Shouldn't this approach include transparent disclosure practices, robust regulatory frameworks, ethical guidelines, and mechanisms for accountability and oversight?

Public good: While it's clear that open-source AI can democratize access to technology, doesn't its implementation require careful consideration of many factors to ensure it serves the public good without compromising safety and security?

Legislators should have adequate answers to the questions above before fully supporting the continued large-scale proliferation of unsecured AI. Wouldn't the world otherwise be taking unlimited and unknowable risks?

MORE RESOURCES:

Wrong Assumptions on US AI Regulation ❓

What AI "experts" get wrong — ChatGPT Red Teamer speaks

How to regulate Open Source AI

SUBSCRIBE TO GET NEW POSTS IN YOUR INBOX

Follow me on LinkedIn or X (Twitter)

AI Artificial Intelligence Ntia Open Source Meta Ai