R Street
Free markets. Real solutions.

R Street Policy Study
No. 319
April 2025

Mapping the Open-Source AI Debate: Cybersecurity Implications and Policy Priorities

By Haiman Wong

This study examines the ongoing debate between open- and closed-source AI, assessing the trade-offs between openness, security, and innovation while evaluating emerging "hybrid" solutions that aim to balance these competing priorities.

Executive Summary

Open-source artificial intelligence (AI) has emerged as a defining force in AI development, offering the potential to scale innovation while also complicating existing cybersecurity, market, and governance challenges. While open-source AI may seem like a novel paradigm, the principles of collaboration, transparency, and shared development have long been foundational to advancing technological progress, shaping everything from early software development to modern cloud infrastructure.

Yet even as open-source AI expands accessibility and competition, concerns over its cybersecurity risks, investment implications, and legal ambiguities continue to rise. This study examines the ongoing debate between open- and closed-source AI, assessing the trade-offs between openness, security, and innovation while evaluating emerging "hybrid" solutions that aim to balance these competing priorities.

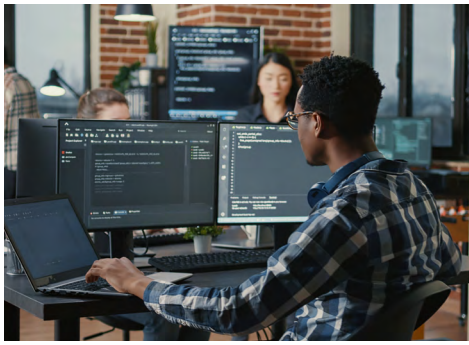
Over the past two years, federal and state legislative efforts, along with industry-led initiatives, have already sought to establish clearer governance frameworks for responsible open-source AI development. However, uncertainty remains over how best to govern these systems without undermining their role as drivers of U.S. innovation. As bipartisan recognition of open-source AI's strategic

Table of Contents

importance grows, the challenge ahead will be to craft adaptable policies that mitigate potential risks while harnessing its benefits. This study identifies key policy priorities, emerging technological solutions, and best practices to ensure that open-source AI remains a force for our economic growth, global AI competitiveness, and national security.

Introduction

Amid the ongoing advances in artificial intelligence (AI), open-source AI is emerging as a distinct branch within AI development. Unlike closed-source AI, where companies retain full control over the model data, development, and deployment, open-source AI allows developers to freely access, modify, and distribute models.[1] Rooted in principles of collaboration, transparency, and accessibility, open-source AI has the potential to reshape innovation, competition, and technological progress. However, its growing adoption also raises concerns, including those related to cybersecurity threats like algorithmic jailbreaking, sensitive data leaks, and model manipulation.[2] Alongside these concerns, the widespread availability of open-source models could shift investment priorities, potentially reducing incentives for proprietary AI development.[3]



Notably, DeepSeek's R-1 model raised major concern after researchers discovered cybersecurity flaws, including the exposure of sensitive data, such as chat histories, API keys, and directory structures.

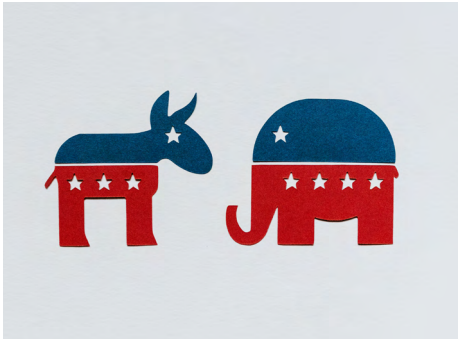
Meta's release of Llama 2 in 2023, which is a suite of large language models, exemplifies open-source AI's growing influence and industry-driven momentum.[4] Llama 2 allows developers to freely access, modify, and deploy AI models, marking a shift in Meta's strategy toward fostering AI collaboration.[5] Specifically, Mark Zuckerberg positioned open-source AI as key to accelerating model development, integration across cloud platforms, and customization.[6] Yet Llama 2 also highlighted unresolved challenges within the open-source community. Critics argue that open-source AI falls short of full openness because of restricted training-data access and commercial-use limitations.[7] Similar debates surround other AI releases, including Google's Gemma models, Microsoft's Phi-4, IBM's Granite, DeepSeek's V3 model, and Advanced AI's Sky-T1 reasoning model.[8] Notably, in January 2025, DeepSeek's latest R-1 model raised major concerns after researchers discovered cybersecurity flaws, including the exposure of sensitive data, such as chat histories, application programming interface (API) keys, and directory structures.[9] These examples illustrate both the potential and complexities of open-source AI, particularly in defining what constitutes "openness" across different models and governance approaches.

1. George Lawton, "Attributes of open vs. closed AI explained," TechTarget, July 8, 2024. https://www.techtarget.com/searchenterpriseai/feature/Attributes-of-open-vs-closed-AI-explained.
2. Gal Nagli, "Wiz Research Uncovers Exposed DeepSeek Database Leaking Sensitive Information, Including Chat History," Wiz, Jan. 29, 2025. https://www.wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak.
3. Kolawole Samuel Adebayo, "The Biggest Winner In The DeepSeek Disruption Story Is Open Source AI," *Forbes*, Jan. 28, 2025. https://www.forbes.com/sites/kolawolesamueladebayo/2025/01/28/the-biggest-winner-in-the-deepseek-disruption-story-is-open-source-ai.
4. Dave Bergmann, "What is Llama 2?," IBM, Dec. 19, 2023. https://www.ibm.com/think/topics/llama-2.
5. Ibid.
6. Mark Zuckerberg, "Open Source AI is the Path Forward," Meta, July 23, 2024. https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward.
7. Steven J. Vaughan-Nichols, "Meta can call Llama 2 open source as much as it likes, but that doesn't mean it is," The Register, July 24, 2023. https://www.theregister.com/2023/07/21/llama_is_not_open_source.
8. "Gemma Open Models," Google AI for Developers, last accessed Nov. 20, 2024. https://ai.google.dev/gemma; Carl Franzen, "Microsoft makes powerful Phi-4 model fully open-source on Hugging Face," VentureBeat, Jan. 8, 2025. https://venturebeat.com/ai/microsoft-makes-powerful-phi-4-model-fully-open-source-on-hugging-face; "IBM Granite," IBM, last accessed Nov. 20, 2024. https://www.ibm.com/granite; Shubham Sharma, "DeepSeek-V3, ultra-large open-source AI, outperforms Llama and Qwen on launch," VentureBeat, Dec. 26, 2024. https://venturebeat.com/ai/deepseek-v3-ultra-large-open-source-ai-outperforms-llama-and-qwen-on-launch.
9. Lily Hay Newman and Matt Burgess, "Exposed DeepSeek Database Revealed Chat Prompts and Internal Data," *Wired*, Jan. 29, 2025. https://www.wired.com/story/exposed-deepseek-database-revealed-chat-prompts-and-internal-data.

As the open-source AI debate unfolds, policymakers, developers, and researchers have competing priorities. Policymakers must balance innovation incentives with regulatory safeguards against cybersecurity threats and ethical risks. On the other hand, AI developers and researchers are exploring "hybrid" approaches, such as tiered or controlled access, to reconcile openness with security concerns. Meanwhile, end users face challenges in verifying trustworthiness, protecting data privacy, and addressing skills gaps that could lead to misuse or unintended vulnerabilities.

Although the precise direction of U.S. policy remains uncertain under the second Trump administration and the 119th Congress, the bipartisan recognition of open-source AI's potential is essential. A balanced regulatory approach that mitigates risks while fostering innovation is crucial for ensuring its continued progress.

This policy study maps the evolving open-source AI landscape, examining its benefits, limitations, and policy considerations. It also presents actionable best practices for industry leaders, policymakers, and the public. Recommendations include establishing federal guidelines to clarify legal ambiguities, fostering public–private partnerships for AI validation, implementing risk-tiered liability shields, and promoting community-driven accountability mechanisms. By grounding the open-source AI debate within the broader history of technological development, this study provides a foundation for informed policymaking that strengthens our national security, technological leadership, and economic competitiveness.



Bipartisan recognition of open-source AI's potential is essential.

A Brief History of Open Source

Although open-source AI may seem to some like uncharted territory, the open-source movement itself has a rich history that continues to influence today's development culture and technological landscape. While the evolution of the open-source movement is marked by many milestones, three defining benchmarks stand out as critical in establishing its principles, growth, and impact: the GNU Project led by Richard Stallman, which laid the ideological foundation for free software; the development of Linux, which showcased the scalability and success of open-source collaboration; and OpenAI's founding mission, which aimed to apply open-source principles into early efforts to advance AI.

The term "open source" was originally derived from the term "free software."[10] At a high level, "free software" was a term coined in the early 1980s to describe software that "respects users' freedom and community."[11] Richard Stallman, a Massachusetts Institute of Technology programmer, popularized "free software," emphasizing users' rights to modify and share software."[12] In the earlier days of software and computer development, distributing free software was common practice for companies and programmers.[13] Even as debates over access and control continued, software development shifted toward proprietary models by the 1980s, driven by growing commercial incentives and the increasing value of intellectual property.[14]

KEY DATES

EARLY 1980s:
Emergence of "Free Software"

1980s:
Shift Toward Proprietary Models

10. "What is Free Software?," Free Software Foundation, last accessed Nov. 20, 2024. https://www.gnu.org/philosophy/free-sw.html.
11. Ibid.
12. Ibid.
13. "GNU," Stanford University, last accessed Nov. 20, 2024. https://cs.stanford.edu/people/eroberts/courses/cs181/projects/2000-01/open-source/gnu.htm#.
14. Ibid.

In response to these shifts, Stallman started the "GNU's Not Unix!" (GNU) Project in 1983.[15] Driven by his belief that end users should be empowered to participate in the development and refinement of the software they use, Stallman's GNU Project aimed to make "cooperation possible once again by removing restrictions to set up proprietary software vendors."[16] Moreover, Stallman's 1985 "GNU Manifesto" laid the foundation for modern open-source principles: free use, modification, and sharing.[17]

Another major accomplishment of the GNU Project was the creation of the GNU General Public License (GPL). Developed in 1989, GPL was a "single license that could be used for all free software in place of all the individual licenses that were being written for individual programs."[18] This license played a critical role in codifying the freedom to run a software program for any purpose; study how the program works and customize it; redistribute copies of the software program to help others; and improve the software program and share individual changes that were made.[19] Most importantly, the GPL's "copyleft" provision ensured that derivative works remained open source, reinforcing GNU's mission.[20] Finally, the GNU Project produced critical tools, including the GNU operating system that would later be combined with the Linux kernel to create the pillar of many modern computing systems, such as desktops, Android phones, and routers, among others.[21]

However, despite the GNU Project's significant achievements and seminal contributions to the open-source movement, Stallman's rigid ideological stance often limited its impact.[22] Specifically, the GNU Project struggled to attract broader audiences, particularly in business and government contexts, leaving gaps in how open-source principles could align with practical, commercial needs.[23] These challenges would later be addressed in subsequent milestones in the development of open-source software and projects, such as the release of Linux, the formation of the Open Source Initiative, the development of GitHub, and the eventual founding of OpenAI.

Less than 10 years after Stallman founded the GNU Project, Linux Torvalds, developed the Linux operating system in 1991 after experiencing frustrations with the limitations of existing systems at the time like Minix.[24] Torvalds' Linux project, initially personal, became transformative after its GPL release.[25] By adopting the GPL, Torvalds ensured that Linux would remain freely accessible and open for modification, accelerating its expansion and refinement.[26] Torvalds' decision not only invited worldwide collaboration but also established Linux as one of the most significant open-source projects to

KEY DATES

1983
Launch of the GNU Project

1985
The GNU Manifesto

1989
Creation of the GNU General Public License (GPL)

1991
Development of Linux

15. Ibid.
16. Ibid.
17. Richard Stallman, "The GNU Manifesto," Free Software Foundation, March 1985. https://www.gnu.org/gnu/manifesto.en.html.
18. "GNU Definition," Linux Information Project, April 2, 2004. https://www.linfo.org/gnu.html.
19. Ibid.
20. "What is Copyleft?," Free Software Foundation, last accessed Nov. 20, 2024. https://www.gnu.org/licenses/copyleft.en.html.
21. Richard Stallman, "Linux and the GNU System," Free Software Foundation, last accessed Nov. 20, 2024. https://www.gnu.org/gnu/linux-and-gnu.en.html.
22. Steven J. Vaughan-Nichols, "GNU Project developers object to Richard M Stallman's continued leadership," ZDNet, Oct. 9, 2019. https://www.zdnet.com/article/gnu-project-developers-object-to-richard-m-stallmans-continued-leadership.
23. Glyn Moody, "Rebel Code: Linus Torvalds, Open Source, and the War for the Soul of Software," *The Guardian*, May 8, 2001. https://www.theguardian.com/education/2001/apr/10/highereducation.mathematics4.
24. Michael Calore, "Aug. 25, 1991: Kid From Helsinki Foments Linux Revolution," *Wired*, Aug. 25, 2009. https://www.wired.com/2009/08/0825-torvalds-starts-linux; Glyn Moody, "How Linux was born, as told by Linus Torvalds himself," Ars Technica, Aug. 25, 2015. https://arstechnica.com/information-technology/2015/08/how-linux-was-born-as-told-by-linus-torvalds-himself.
25. Christopher Tozzi, "Linus Torvalds on Early Linux History, GPL License and Money," Data Center Knowledge, Aug. 23, 2016. https://www.datacenterknowledge.com/business/linus-torvalds-on-early-linux-history-gpl-license-and-money.
26. Ibid.

date. Since its release, Linux has come to underpin cloud computing, drones, supercomputers, and more.[27]

The success of Linux demonstrated how copyleft agreements could scale effectively, fostering a global community of developers committed to improving and expanding a given software while safeguarding its openness.[28] By proving that open-source collaboration could produce secure, reliable, scalable, and innovative solutions, Linux successfully bridged the gap between the ideals of the GNU Project and the practical needs of developers, businesses, and even governments.

About 25 years later, OpenAI was founded as a nonprofit, aimed at democratizing AI access and mitigating existential risks.[29] Although OpenAI is perhaps most recognized today for its ChatGPT model, one of its earliest achievements was the 2016 release of OpenAI Gym—an open-source toolkit for developing and benchmarking reinforcement learning algorithms.[30]

As OpenAI's models and initiatives became more sophisticated and expensive to develop, the company began to impose restrictions on access to and use of those models, citing the need to attract and retain top talent, along with growing ethical and legal concerns.[31] By 2019, OpenAI transitioned to a "capped" for-profit model, sparking criticism that it had abandoned its open-access mission in favor of profit.[32]

Despite OpenAI's transition from full open-source accessibility, its early contributions and founding mission illustrate how open-source principles could be applied in AI innovation. OpenAI's changing business model also underscores an ongoing challenge in the open-source movement: navigating the tension between fostering accessibility and maintaining responsible oversight in rapidly changing technological ecosystems and markets. This tension is especially relevant as policymakers consider the similar balance that must be struck between security and innovation.

The Role of Open Source in AI Development

Currently, open source plays three major roles in advancing AI development: facilitating the creation of foundational datasets that fuel AI model training, providing the digital infrastructure necessary for collaborating on the refinement of AI systems, and democratizing access to AI resources that enable prototyping.[33]

Representative datasets are essential for training effective AI models.[34] Initiatives rooted in open-source principles not only ensure the availability and accessibility of these datasets but also promote their continuous expansion and enhancement,

KEY DATES

2015:
Founding of OpenAI

2019:
Transition to a "Capped" For-Profit Model

Open source plays three major roles in advancing AI development:

Facilitating the creation of foundational datasets that fuel AI model training

Providing the digital infrastructure necessary for collaborating on the refinement of AI systems

Democratizing access to AI resources that enable prototyping

27. Klint Finley, "Linux Took Over the Web. Now, It's Taking Over the World," *Wired*, Aug. 25, 2016. https://www.wired.com/2016/08/linux-took-web-now-taking-world.
28. Joe Casad, "The Story of the GPL," *Linux Magazine* 200 (2017). https://www.linux-magazine.com/Issues/2017/200/The-GPL-and-the-birth-of-a-revolution.
29. Kenneth Niemeyer, "OpenAI's mission to develop AI that 'benefits all of humanity' is at risk as investors flood the company with cash," Business Insider, Sept. 15, 2024. https://www.businessinsider.com/sam-altman-openai-mission-drift-for-profit-nonprofit-structure-investment-2024-9; Karen Hao, "The messy, secretive reality behind OpenAI's bid to save the world," MIT Technology Review, Feb. 17, 2020. https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality.
30. Greg Brockman, "OpenAI Gym Beta," OpenAI, April 27, 2016. https://openai.com/index/openai-gym-beta.
31. Hao. https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality.
32. "Our structure," OpenAI, March 2019. https://openai.com/our-structure; Carl Franzen, "OpenAI's former superalignment leader blasts company: 'safety culture and processes have taken a backseat,'" VentureBeat, May 17, 2024. https://venturebeat.com/ai/openais-former-superalignment-leader-blasts-company-safety-culture-and-processes-have-taken-a-backseat; Taylor Herzlich, "'Godfather of artificial intelligence' Geoffrey Hinton backs Elon Musk's OpenAI legal battle,'" New York Post, Dec. 31, 2024. https://nypost.com/2024/12/31/business/geoffrey-hinton-backs-elon-musks-legal-battle-against-openai.
33. Rahul Roy-Chowdhury, "Why open-source is crucial for responsible AI development," World Economic Forum, Dec. 22, 2023. https://www.weforum.org/stories/2023/12/ai-regulation-open-source.
34. Katharine Miller, "Data-Centric AI: AI Models Are Only as Good as Their Data Pipeline," Stanford University, Jan. 25, 2022. https://hai.stanford.edu/news/data-centric-ai-ai-models-are-only-good-their-data-pipeline.

creating a robust foundation for continued progress across AI subfields.[35] In natural language processing, for example, now-recognized state-of-the-art language models like GPT-3 (Generative Pre-trained Transformer) were made possible thanks to open repositories of web data, such as Common Crawl.[36] By adopting an "open" approach to its data, Common Crawl empowers researchers, particularly those who are unaffiliated or in resource-constrained institutions, to use its web archive to train and fine-tune AI models that would otherwise be prohibitively expensive to develop.

Beyond datasets, open-source frameworks have been instrumental in expanding the infrastructure needed to train, validate, and refine AI models.[37] Tools like TensorFlow and PyTorch have become essential in projects across both industry and academia, offering scalable and flexible platforms for developing AI systems. TensorFlow—an open-source, machine learning framework—supports large-scale production environments, while PyTorch—an open-source, deep learning framework—has quickly gained traction among researchers for its user-friendly design and adaptability to experimental needs.[38]

These frameworks have been complemented by repositories like Hugging Face, which provides an online hub of pre-trained AI models that developers can refine for targeted tasks, saving time and computational resources.[39] Over the past two years, Hugging Face has also become one of the most widely used platforms in the AI community, having more than doubled its valuation from roughly $2 billion in April 2022 to $4.5 billion in August 2023.[40] This growth underscores how open-source infrastructure not only facilitates technical AI development but also fosters a culture of shared innovation where progress and expertise are continuously exchanged to accelerate advancements across the AI landscape.

Over the past two years, Hugging Face has become one of the most widely used platforms in the AI community, having more than doubled its valuation from roughly $2 billion in April 2022 to $4.5 billion in August 2023.

Finally, the guiding principles of an open approach to AI development have substantially democratized access to AI resources, enabling increased prototyping and iterative experimentation that are key to technological innovation.[41] Open-source platforms like Jupyter Notebooks and OpenML lower barriers to entry by providing free tools, pre-trained models, and efficient workflows for hands-on experimentation. For instance, Jupyter Notebook provides an interactive environment where users can write and run code, visualize outputs, and collaborate in real time. This accessibility has made it an essential tool for prototyping and experimentation, learning, and sharing insights across the AI developer community.[42] Similarly, OpenML expands opportunities for collaborative development by allowing researchers and practitioners to benchmark models, refine algorithms, and contribute to shared datasets.[43]

35. "Open data and AI: A symbiotic relationship for progress," European Union, June 9, 2023. https://data.europa.eu/en/publications/datastories/open-data-and-ai-symbiotic-relationship-progress.
36. Nick Barney, "What is GPT-3? Everything you need to know," *TechTarget*, January 2025. https://www.techtarget.com/searchenterpriseai/definition/GPT-3; "Overview," Common Crawl, last accessed Nov. 20, 2024. https://commoncrawl.org.
37. Yuliya Melnik, "The Top 16 Frameworks and Libraries: A Beginner's Guide," DataCamp, Sept. 29, 2023. https://www.datacamp.com/blog/top-ai-frameworks-and-libraries.
38. Chris Tozzi, "Compare PyTorch vs. TensorFlow for AI and machine learning," *TechTarget*, Dec. 11, 2024. https://www.techtarget.com/searchenterpriseai/tip/Compare-PyTorch-vs-TensorFlow-for-AI-and-machine-learning.
39. Ben Lutkevich, "Hugging Face," *TechTarget*, September 2023. https://www.techtarget.com/whatis/definition/Hugging-Face.
40. "AI startup Hugging Face valued at $4.5 bln in latest round of funding," *Reuters*, Aug. 24, 2023. https://www.reuters.com/technology/ai-startup-hugging-face-valued-45-bln-latest-round-funding-2023-08-24.
41. Parth Nobel et al., "Open-Access AI: Lessons From Open-Source Software," Lawfare, Oct. 25, 2024. https://www.lawfaremedia.org/article/open-access-ai--lessons-from-open-source-software.
42. Chris Tozzi, "How to use and run Jupyter Notebook: A beginner's guide," *TechTarget*, Aug. 14, 2024. https://www.techtarget.com/searchenterpriseai/tutorial/How-to-use-and-run-Jupyter-Notebook-A-beginners-guide.
43. "Overview," OpenML, last accessed Nov. 20, 2024. https://openml.org.

The prominent role that open-source principles play in AI development today represents a marked paradigm shift from the proprietary strategies that leading technology firms have relied upon in recent decades. This change stems, in part, from the increasing complexity and resource demands (i.e., energy and compute power) of developing cutting-edge AI systems, which no single organization can easily or fully sustain independently.[44] By participating in open-source initiatives, technology firms like Meta, Google, and Microsoft benefit from the collective innovation and feedback of a global developer community, accelerating advances in AI while reducing the costs of research and development.[45] Signaling their commitment to the open-source approach also helps these firms influence industry standards, build goodwill with the public, and expand their ecosystem.[46]

By participating in open-source initiatives, technology firms like Meta, Google, and Microsoft benefit from the collective innovation and feedback of a global developer community, accelerating advances in AI while reducing the costs of research and development.

However, the sustainability of this open-source enthusiasm remains uncertain, as companies must balance the tension between fostering openness and protecting competitive advantages.[47] Whether the open-source approach endures in AI development will depend on the ability and willingness of stakeholders to address challenges like licensing ambiguities, cybersecurity threats, and ethical use across varying degrees of openness.

The Debate Between Open-Source and Closed-Source AI

At its core, the open-source AI debate hinges on two distinct approaches to AI development: open source and closed source. Each offers distinct benefits and limitations, with far-reaching implications for cybersecurity, governance, and AI innovation.

The "Open" Approach to AI Development

As described earlier, open-source AI fosters transparency, collaboration, and accessibility, allowing stakeholders from the public and across industry, government, and academia to share resources; test prototypes and emerging ideas; and refine AI development practices.[48] By making datasets, algorithms, and models freely available, open-source initiatives enable broader participation in AI development, particularly from under-resourced institutions and nontechnical end users.[49] This democratization fosters a culture of continuous learning and experimentation, which is vital for AI advancement and innovation.

However, the "open" approach to AI development also introduces pronounced challenges. Cybersecurity vulnerabilities represent one of the most pressing concerns, as unrestricted access to AI training data, code scripts, AI models, and AI systems can be exploited by malicious threat actors. In March 2024, for example, thousands of companies, including Uber, Amazon, and OpenAI, that use Ray—a popular open-source AI framework used to develop and deploy large-scale Python applications

44. Thomas Claburn, "To save the energy grid from AI, use open source AI, says open source body," The Register, Jan. 9, 2025. https://www.theregister.com/2025/01/09/linux_foundation_ai_energy_report.
45. Cailean Osborne, "Why Companies 'Democratise' Artificial Intelligence: The Case of Open Source Software Donations," arXiv, Sept. 26, 2024. https://arxiv.org/html/2409.17876v1.
46. Ibid.
47. Will Douglas Heaven, "The open-source AI boom is built on Big Tech's handouts. How long will it last?," MIT Technology Review, May 12, 2023. https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta.
48. Ben Brooks, "Open-Source AI Is Good for Us," IEEE Spectrum, Feb. 8, 2024. https://spectrum.ieee.org/open-source-ai-good.
49. Ibid.

for data processing and machine learning—were exposed to cyber attackers.[50] The vulnerability CVE-2023-48022 within Ray allowed cyber threat actors to steal credentials, remotely control servers, and corrupt AI models.[51] This attack, dubbed the "ShadowRay" campaign, highlights how open-source AI can exacerbate existing cybersecurity risks and threat vectors through the unrestricted access that open-source AI offers.[52]

Open-source AI ecosystems are also more susceptible to cybersecurity risks like data poisoning and adversarial attacks because their lack of controlled access, centralized oversight, standardization, and clear guidelines for acceptable use can hinder vulnerability identification and incident-response efforts.[53] Apart from their limitations in cybersecurity resilience, the collaborative nature of open-source AI ecosystems can sometimes lead to issues of quality control, where inconsistent contributions can complicate the integration of robust and reliable systems.[54] Additionally, open-source AI projects can suffer from limited accountability mechanisms.[55] Because there is often no singular entity responsible for overseeing the security or ethical use of such systems, responses to misuse or exploitation can be slower and more inconsistent. This diffusion of responsibility can ultimately create governance challenges, leaving gaps in enforceable rules for ethical open-source AI development and deployment.[56]

Several questions and knowledge gaps continue to complicate open-source AI's longer-term development.

In addition to these ongoing challenges, several questions and knowledge gaps continue to complicate open-source AI's longer-term development. The governance of open-source AI lacks clarity on the international stage about how to enforce accountability within its decentralized and globalized systems.[57] Questions about how to manage sensitive data, enforce licensing standards, and address intellectual property disputes remain largely unanswered, particularly as proprietary elements become increasingly intertwined with open-source projects.[58] Scalability is another notable issue because open-source systems often struggle to meet the growing computational and infrastructural demands required for state-of-the-art AI development.[59] Even with the expanded contributions made by leading technology companies over the past two years, it remains unclear whether these firms will maintain their commitment to open source or eventually revert to proprietary practices to protect their competitive advantages.[60] Furthermore, future research and development initiatives in securing open-source AI include standardizing protocols and practices to establish consistent security measures across projects; designing tailored tools for detecting and preventing threats like data poisoning and adversarial attacks; and developing reliable methods to trace and attribute malicious activities.[61]

50. Daryna Antoniuk, "Thousands of companies using Ray framework exposed to cyberattacks, researchers say," Recorded Future News, March 26, 2024. https://therecord.media/thousands-exposed-to-ray-framework-vulnerability.
51. Ibid.
52. Haiman Wong, "Lessons Learned from the 'ShadowRay' Campaign - The First Known Attack Targeting AI Workloads," R Street Institute, March 29, 2024. https://www.rstreet.org/commentary/lessons-learned-from-the-shadowray-campaign-the-first-known-attack-targeting-ai-workloads.
53. Maria Korolov, "10 things to watch out for with open source gen AI," CIO, May 15, 2024. https://www.cio.com/article/2104280/10-things-to-watch-out-for-with-open-source-gen-ai.html.
54. Ibid.
55. Ibid.
56. David Evan Harris, "How to Regulate Unsecured 'Open-Source' AI: No Exemptions," Tech Policy Press, Dec. 4, 2023. https://techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions.
57. Ibid.
58. Ibid.
59. Heaven. https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta.
60. Ibid.
61. Harris. https://techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions.

The "Closed" Approach to AI Development

In contrast, closed-source AI restricts access to proprietary AI data, models, and algorithms.[62] Through this approach, organizations can maintain rigorous oversight over their AI systems and development processes, reducing risks associated with misuse.[63] By emphasizing standardization, controlled access, and proprietary protections, closed-source AI allows companies to ensure the quality and reliability of their AI products, which is particularly critical in AI applications in healthcare, finance, and national security where data is highly sensitive.

The closed-source approach to AI development can also offer commercial advantages over open-source AI. A report released in November 2024 underscored this gap, finding that "the best open large language models have trailed the closed models" in benchmark performance and training compute "by anything from 5 to 22 months on average."[64] This disparity highlights how propriety models can also allow companies to stay ahead in AI advances. By monetizing proprietary advances, companies can recoup the substantial investments required for research and development while maintaining a competitive edge in the market.[65] This business model is also consistent with more conventional business strategies, where intellectual property protection serves as a key driver of profitability, influence, and innovation.[66]

Nevertheless, the closed-source approach also has drawbacks. The lack of transparency and explicability often raises ethical and trust-related concerns, particularly when proprietary AI systems are deployed at large scales in high-stakes environments.[67] Without access or insights into the underlying training data or models, researchers, regulators, and end users struggle to identify potential biases or unfair development practices embedded within these systems.[68] Moreover, the siloed nature of closed-source AI development can hinder collaboration and limit opportunities for cross-industry standardization and integration.[69]

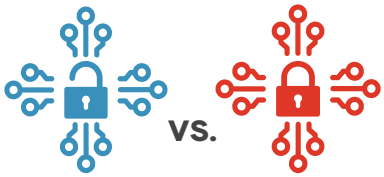
From a cybersecurity perspective, closed-source AI presents a paradox. While its restricted access reduces surface-level risks, such as tampering and unauthorized use, it can also create blind spots.[70] Hidden and unique vulnerabilities in proprietary systems may go undetected for longer periods of time because fewer people are working in these environments.[71] Furthermore, the inherently siloed nature of closed-source development may limit or even disincentivize opportunities for collaboration and information-sharing, impeding the creation of interoperable and resilient AI security frameworks.[72]

From a cybersecurity perspective, closed-source AI presents a paradox. While its restricted access reduces surface-level risks, such as tampering and unauthorized use, it can also create blind spots.

62. George Lawton, "Attributes of Open vs. Closed AI Explained," *TechTarget*, July 8, 2024. https://www.techtarget.com/searchenterpriseai/feature/Attributes-of-open-vs-closed-AI-explained.
63. Ibid.
64. John Werner, "Open AI Systems Lag Behind Proprietary and Closed Models," *Forbes*, Nov. 6, 2024. https://www.forbes.com/sites/johnwerner/2024/11/06/open-ai-systems-lag-behind-proprietary-and-closed-models/?utm_source=chatgpt.com.
65. Heaven. https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta.
66. Ibid.
67. Ben Chester Cheong, "Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making," *Frontiers in Human Dynamics* 6 (July 2, 2024). https://www.frontiersin.org/journals/human-dynamics/articles/10.3389/fhumd.2024.1421273/full.
68. Ibid.
69. Ibid.
70. Adebayo. https://www.forbes.com/sites/kolawolesamueladebayo/2025/01/28/the-biggest-winner-in-the-deepseek-disruption-story-is-open-source-ai.
71. Jon Bateman et al., "Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation AI Model Governance," Carnegie Endowment for International Peace, July 23, 2024. https://carnegieendowment.org/research/2024/07/beyond-open-vs-closed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance.
72. Ibid.

Importantly, these challenges also present several key opportunities for continued research and development. For example, research that supports the development of explicable AI (XAI) frameworks—which make AI decision-making processes interpretable for external stakeholders without revealing proprietary AI data, models, or development practices—could help address ongoing concerns over the lack of transparency in closed-source AI systems.[73] Another opportunity for future research and development is specialized APIs that allow companies to integrate open- and closed-source AI models while maintaining intellectual property protections.[74]

Beyond the Debate



As the open- vs. closed-source AI debate continues to unfold, the long-term success of AI development will depend on the ability of key stakeholders to objectively weigh and holistically evaluate the benefits, applications, and limitations of each approach.

Unsurprisingly, few stakeholders advocate for adopting a fully open-source approach or abandoning its principles altogether. Instead, there is a broad consensus that the ideal approach moving forward would be to find a way to marry the distinct benefits that each approach offers.[75] Most of the ongoing debate and outstanding challenges center around establishing what an integrated approach should look like, determining which solutions to prioritize, and crafting regulatory frameworks for effective management.

An Appraisal of Proposed Approaches

Some researchers, developers, and practitioners have already presented various "hybrid" AI approaches that aim to balance the openness and collaborative nature of open-source AI with the security and higher access controls of closed-source AI. This section explores three leading "hybrid" AI approaches—open-source AI with controlled access, tiered open-source AI access, and federated learning—examining their distinct features, benefits, and limitations, along with their implications for AI innovation, cybersecurity, and governance.

Open-Source AI with Controlled Access

Merging the transparent ethos of open-source principles with increased access controls such as licensing agreements, ethical guidelines, and user approval, open-source AI with controlled access aims to ensure that shared tools are used responsibly.[76]

Meta's Llama model, for example, requires users to apply for access and enforces a license that explicitly prohibits high-risk applications of the tool, including those intended to harm individuals, undermine public safety, or violate human rights.[77] By defining clear parameters for acceptable use, Llama demonstrates how transparency and collaboration can coexist with rigorous oversight.

73. Cheong. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4961260.
74. Asha Sharma, "DeepSeek R1 is now available on Azure AI Foundry and GitHub," Microsoft, Jan. 29, 2025. https://azure.microsoft.com/en-us/blog/deepseek-r1-is-now-available-on-azure-ai-foundry-and-github.
75. Irene Solaiman, "Generative AI Systems Aren't Just Open or Closed Source," *Wired*, May 24, 2023. https://www.wired.com/story/generative-ai-systems-arent-just-open-or-closed-source.
76. Will Henshall, "The Heated Debate Over Who Should Control Access to AI," *Time*, Nov. 4, 2024. https://time.com/6308604/meta-ai-access-open-source.
77. "Llama 2 Version Release," Meta, July 18, 2023. https://ai.meta.com/llama/license.

One of the most significant benefits of this approach to open-source AI is its cybersecurity potential. By regulating access, this "hybrid" approach reduces opportunities for malicious actors to exploit advanced tools for harmful purposes, such as launching cyberattacks, spreading disinformation, or compromising critical infrastructure.[78] Innovation can also thrive under this approach because contributions come from researchers and developers who are vetted through an application process, ensuring higher quality control and improved attribution.[79] Furthermore, from a governance perspective, licensing agreements can establish clear accountability by defining acceptable use and enabling enforcement mechanisms to hold users responsible for violations.[80]

A "hybrid" approach to open-source AI reduces opportunities for malicious actors to exploit advanced tools for harmful purposes, such as launching cyberattacks, spreading disinformation, or compromising critical infrastructure.

However, this approach also has limitations. Monitoring and enforcing compliance with licensing agreements in a controlled-access framework is resource-intensive, requiring significant commitment from the host to track usage, detect violations, and address any unauthorized use.[81] In addition, if the terms and restrictions outlined in the licensing agreements change or expand over time, they can become overly burdensome, potentially stifling collaboration and limiting creativity. User approval processes, while essential for maintaining oversight and quality control, can also introduce bureaucratic delays that hinder timely access to models, particularly in fast-paced research and development environments.[82] Moreover, the subjectivity inherent in granting and maintaining access introduces risks of inconsistency, bias, or errors, meaning that bad actors can still slip through access controls.[83] These limitations highlight the importance of continually refining enforcement mechanisms and establishing robust safeguards to ensure the efficacy of controlled access without compromising innovation or security.

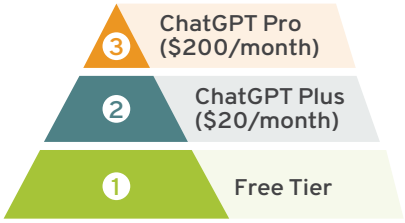
Tiered Open-Source AI Access

As the name suggests, tiered open-source AI access offers varying levels of access to AI models and capabilities.[84] While foundational resources are often openly available to the public, new and advanced features are gated behind tiers based on criteria such as payment or partnerships.[85] Unlike the binary nature of the controlled-access approach to open-source AI, which either grants or denies access outright, tiered access stratifies users into groups. In doing so, this approach uniquely balances openness and commercialization, enabling broad collaboration while reserving advanced features for users who meet specific conditions.

OpenAI's ChatGPT is one example of how this "hybrid" approach can be applied in practice. Casual users who do not want to pay a monthly fee have access to basic features powered by older GPT models, with more limitations on messaging volume and slower response times.[86] For users who want access to newer GPT models, faster

78. Bergmann. https://www.ibm.com/think/topics/llama-2.
79. Ibid.
80. Lee Tiedrich, "Open-Source and open access licensing in an AI Large Language Model (LLMs) world," OECD.AI, June 12, 2024. https://oecd.ai/en/wonk/open-source-and-open-access-licensing-in-an-ai-large-language-model-llms-world.
81. Stefano Maffulli, "Meta's LLaMa 2 license is not Open Source," Open Source Initiative, July 20, 2023. https://opensource.org/blog/metas-llama-2-license-is-not-open-source.
82. Ibid.
83. Tiernan Ray, "Cybercriminals are using Meta's Llama 2 AI, according to CrowdStrike," ZDNet, Feb. 21, 2024. https://www.zdnet.com/article/cybercriminals-are-using-metas-llama-2-ai-according-to-crowdstrike.
84. Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," arXiv, Feb. 5, 2023. https://arxiv.org/abs/2302.04844?utm_source=chatgpt.com.
85. Ibid.
86. Imad Khan, "ChatGPT Free vs. ChatGPT Plus: Worth the $20 Upgrade?," CNET, July 14, 2024. https://www.cnet.com/tech/services-and-software/chatgpt-free-vs-chatgpt-plus-worth-the-20-upgrade.

response times, lower limitations on messaging volume, and emerging features like Sora or DALL-E 3, the ChatGPT Plus subscription is available for $20 a month.[87] In late 2024, OpenAI released its highest tier—ChatGPT Pro—as a $200 month subscription aimed to meet the needs of users who want unlimited access to the newest AI models, early access to emerging and advanced features like Advanced Voice, and AI models optimized for professional and high-complexity applications that require cutting-edge reasoning capabilities.[88]



Similar to the controlled-access approach, tiered access can mitigate cybersecurity risks by limiting advanced AI features to only verified users.[89] Each tier also defines specific responsibilities and capabilities, facilitating deployments aligned with acceptable-use standards and emerging regulatory frameworks.[90] Yet the tiered-access approach's distinct advantage lies in its financial sustainability. By allowing free access to foundational AI models or tools, the tiered structure supports widespread experimentation and innovation while relying on paying subscribers to provide some financial capital for continuous improvement and rapid advances.[91]

This approach does have some drawbacks, however. The tiered-access approach relies on payments or partnerships to determine access risks, marginalizing researchers, developers, or organizations that cannot afford premium tiers.[92] Additionally, the tiered structure is not immune to exploitation by malicious actors who are capable of paying for advanced access.[93] Unauthorized use, such as jailbreaking or API key fraud to circumvent restrictions also presents cybersecurity, safety, ethical, and financial threats.[94] With continued refinement aimed at addressing these challenges, such as expanded subsidies for academic institutions, nonprofits, and individuals with demonstrated need; stronger authentication mechanisms; and adaptive policies that clearly outline criteria for tier progression and acceptable use, the tiered access approach offers another promising solution for effectively balancing innovation and accessibility with security and safety.

Federated Learning

In contrast to the tiered-access and controlled-access approaches, which regulate a user's access to AI models and tools, the federated learning approach is a decentralized training paradigm that enables AI models to be collaboratively developed across multiple organizations or devices while ensuring that the raw data remains private.[95] This means that only updates to the AI model—not the underlying data—are shared with a central aggregator, effectively promoting data privacy and security while fostering collaboration and scalability.

87. Ibid.
88. Reece Rogers, "Here's What OpenAI's $200 Monthly ChatGPT Pro Subscription Includes," *Wired*, Dec. 5, 2024. https://www.wired.com/story/openai-chatgpt-pro-subscription.
89. Solaiman. https://arxiv.org/abs/2302.04844?utm_source=chatgpt.com.
90. Ibid.
91. Shirin Ghaffary and Ed Ludlow, "OpenAI CFO Says 75% of Its Revenue Comes From Paying Consumers," Bloomberg News, Oct. 28, 2024. https://news.bloomberglaw.com/private-equity/openai-cfo-says-75-of-its-revenue-comes-from-paying-consumers.
92. Kyle Wiggers, "OpenAI might raise the price of ChatGPT to $44 by 2029," TechCrunch, Sept. 27, 2024. https://techcrunch.com/2024/09/27/openai-might-raise-the-price-of-chatgpt-to-22-by-2025-44-by-2029.
93. Vincenzo Ciancaglini and David Sancho, "Back to the Hype: An Update on How Cybercriminals Are Using GenAI," TrendMicro, May 8, 2024. https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/back-to-the-hype-an-update-on-how-cybercriminals-are-using-genai.
94. Derek B. Johnson, "'Severe' bug in ChatGPT's API could be used to DDoS websites," CyberScoop, Jan. 22, 2025. https://cyberscoop.com/ddos-openai-chatgpt-api-vulnerability-microsoft.
95. Kim Martineau, "What is federated learning?," IBM, Aug. 24, 2022. https://research.ibm.com/blog/what-is-federated-learning.

One prominent example of federated learning is Google's Android Gboard keyboard.[96] By training models locally on users' devices, Google is able to improve predictive text and personalized suggestions without centrally storing sensitive user data.[97] This decentralization reduces single points of failure, enhancing the system's overall resilience.[98] Keeping raw data on local devices also minimizes the risk and impact of large-scale data breaches.[99]

One prominent example of federated learning is Google's Android Gboard keyboard. By training models locally on users' devices, Google is able to improve predictive text and personalized suggestions without centrally storing sensitive user data.

Despite these cybersecurity and privacy benefits, federated learning can be vulnerable to model poisoning, which occurs when malicious users introduce corrupted updates that can degrade or compromise the central model's reliability and accuracy.[100] Additionally, the complexity of decentralized governance poses challenges, particularly when collaborations involve jurisdictions or organizations with conflicting privacy laws and guidelines.[101] These discrepancies can make it difficult to define and enforce responsibility and liability, complicating collaborations.

Federated learning may also hinder innovation and rapid iteration because of communication delays inherent in decentralized training processes.[102] Device heterogeneity and network limitations can further exacerbate these delays, slowing model convergence compared to conventional, centralized approaches.[103] Furthermore, implementing and maintaining federated systems requires robust infrastructure and coordination across devices, users, and organizations, which can be costly and technically demanding.[104] These barriers risk excluding smaller players or independent researchers, limiting the diversity of contributors to federated learning projects.

Overcoming these hurdles will require continued investments in secure aggregation techniques, anomaly detection for malicious updates, and standardized governance frameworks to ensure that federated learning can better balance its privacy protections with scalability and the ability to foster broader innovation.

From Industry Strategies to Policy Responses

Although each of the emerging hybrid approaches to open-source AI explored herein demonstrates significant progress in balancing priorities like cybersecurity, innovation, and governance, they also reveal persistent gaps that may require new technological and creative regulatory solutions. The controlled-access, tiered-access, and federated learning approaches provide valuable frameworks for navigating the open-source AI debate, offering practical strategies to mitigate cybersecurity and safety risks while fostering transparency and collaboration. However, challenges, such as inconsistent governance, unequal access, and vulnerabilities to misuse, underscore the need for further refinement and harmonization.

96. Ziteng Sun and Haicheng Sun, "Improving Gboard language models via private federated analytics," Google Research, April 19, 2024. https://research.google/blog/improving-gboard-language-models-via-private-federated-analytics.
97. Ibid.
98. Muhammad Raza, "Federated Learning in AI: How It Works, Benefits and Challenges," Splunk Blogs, Aug. 28, 2023. https://www.splunk.com/en_us/blog/learn/federated-ai.html.
99. Ibid.
100. Ibid.
101. Ibid.
102. Peter Kairouz et al., "Advances and Open Problems in Federated Learning," arXiv, March 9, 2021. https://arxiv.org/pdf/1912.04977.
103. Ibid.
104. Ibid.

Flexible policy frameworks that support industry-led innovation and uphold ethical standards could help address these challenges. The next section presents recent policy developments aimed at governing the expansion of open-source AI, outlines their intended objectives, and identifies opportunities for further improvement.

Recent Developments Aimed at Advancing Open-Source AI Governance

Over the past two years, AI development and governance efforts have reflected a push-and-pull dynamic, shifting from early attempts at more rigid regulatory control to a growing emphasis on strategic enablement and industry-driven governance.

Initially, open-source AI was framed more as a potential national security liability, prompting aggressive regulatory proposals aimed at limiting its risks. For example, President Joe Biden's Executive Order 14110, issued in October 2023, directed agencies such as the National Institute of Standards and Technology (NIST), National Telecommunications and Information Administration (NTIA), and National Science Foundation (NSF) to develop AI security and transparency guidelines.[105] The NTIA's July 2024 report, a key product of Biden's Executive Order, examined whether open-weight models should face additional restrictions but ultimately concluded that current evidence was insufficient to justify broad limitations.[106]

At the state level, California introduced SB-1047 in early 2024, which took an even more prescriptive approach by proposing liability measures that would have required AI developers to certify that their models posed no potential harm.[107] Although framed as a safeguard against AI misuse, the bill faced strong opposition from industry leaders and researchers, who argued that its broad liability provisions would disincentivize innovation and be legally ambiguous, practically unenforceable, and particularly burdensome for smaller developers.[108] By September 2024, Governor Gavin Newsom vetoed the measure, citing concerns that the bill's language was too imprecise and risked stifling innovation without meaningfully improving AI safety.[109] These earlier developments reflected a precautionary, risk-first approach that prioritized preemptive restrictions and regulatory oversight, but that ultimately struggled to balance innovation and security in a way that was both effective and enforceable.

In mid-2024, the policy conversation shifted toward strategic enablement, focusing on expanding investments in secure AI development, defining open-source AI standards, and establishing more flexible governance frameworks rather than imposing outright restrictions.

In mid-2024, the policy conversation shifted toward strategic enablement, focusing on expanding investments in secure AI development, defining open-source AI standards, and establishing more flexible governance frameworks rather than imposing outright restrictions. This transition was driven by a growing recognition that open-source AI is not just a risk but also a potential asset to maintaining U.S. technological leadership and competitiveness. Later that year, in December 2024, a report on AI from the

105. "FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence," The White House, Oct. 30, 2023. https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence.

106. "Dual-Use Foundation Models with Widely Available Model Weights Report," National Telecommunications and Information Administration, July 30, 2024. https://www.ntia.gov/programs-and-initiatives/artificial-intelligence/open-model-weights-report; "Democratizing the future of AI R&D: NSF to launch National AI Research Resource pilot," U.S. National Science Foundation, Jan. 24, 2024. https://new.nsf.gov/news/democratizing-future-ai-rd-nsf-launch-national-ai; "Department of Commerce Announces New Guidance, Tools 270 Days Following President Biden's Executive Order on AI," National Institute of Standards and Technology, July 26, 2024. https://www.nist.gov/news-events/news/2024/07/department-commerce-announces-new-guidance-tools-270-days-following.

107. Danny Tobey et al., "California's SB-1047: Understanding the Safe and Secure Innovation for Frontier Artificial Intelligence Act," DLA Piper, Feb. 20, 2024. https://www.dlapiper.com/en/insights/publications/2024/02/californias-sb-1047.

108. Scott Kohler, "All Eyes on Sacramento: SB 1047 and the AI Safety Debate," Carnegie Endowment for International Peace, Sept. 11, 2024. https://carnegieendowment.org/posts/2024/09/california-sb1047-ai-safety-regulation?lang=en.

109. Gavin Newsom, "SB-1047 Veto Message," Office of the Governor, Sept. 29, 2024. https://www.gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf.

House Bipartisan Task Force reflected this shift by calling for federal investments in open-source AI research at the NSF, NIST, and the Department of Energy to strengthen AI model security, governance, and privacy protections.[110] Rather than jumping to push for broad regulatory constraints, the report emphasized the importance of taking a risk-based approach that would monitor potential harms over time while sustaining open development.[111] This complemented the Senate AI Working Group's earlier May 2024 report, which had urged Senate committees to examine the policy implications of different product-release strategies for AI systems and to understand the distinctions between closed- and fully open-source models.[112] Meanwhile, the Open Source Initiative, a consortium of 70 researchers, lawyers, policymakers, activists, and representatives from leading technology companies, introduced its first formal definition of "open-source AI" in an effort to establish a universal standard to distinguish between truly open systems and those that incorporated restrictive licensing terms.[113] While praised for providing much-needed clarity, the definition also sparked debate over whether rigid criteria might discourage hybrid models that balance openness with security safeguards.[114]

As the Biden administration transitioned out, it left behind a complex legacy of ambitious initiatives that elevated the role of open-source AI in the context of AI governance and development. Although the former president's 2023 Executive Order has since been rescinded, agency-led initiatives remain under review. At the same time, Congress has yet to pass a comprehensive federal AI law, despite several legislative proposals being introduced and debated in recent years.[115] Meanwhile, states have continued to propose and pass new AI policies, raising the possibility that a fragmented regulatory landscape could emerge in the absence of federal leadership.[116] Finally, although industry-led governance efforts such as the Open Source Initiative's definition and voluntary security frameworks have gained traction, they remain nonbinding and leave questions about long-term enforcement and standardization unanswered.

Looking ahead, the debate over open-source AI governance will likely center on whether policymakers can reconcile national security concerns with the economic and strategic benefits of open-source AI development. With the second Trump administration and the 119th Congress now in place, the challenge will be determining whether AI policy can be consolidated into a cohesive and focused national strategy or will have to remain shaped by a mix of state-led regulations and industry-led frameworks.[117] While open-source AI remains at the heart of the broader AI policy debate, its future role depends on whether governance frameworks can leverage its advantages while mitigating evolving security risks.

With the second Trump administration and the 119th Congress now in place, the challenge will be determining whether AI policy can be consolidated into a cohesive and focused national strategy or will have to remain shaped by a mix of state-led regulations and industry-led frameworks.

110. House Committee on Science, Space, and Technology, "House Bipartisan Task Force on Artificial Intelligence Delivers Report," United States House of Representatives, Dec. 17, 2024. https://science.house.gov/2024/12/house-bipartisan-task-force-on-artificial-intelligence-delivers-report.
111. Ibid.
112. The Bipartisan Senate AI Working Group, "Driving U.S. Innovation in Artificial Intelligence: A Roadmap for Artificial Intelligence Policy in the United States Senate," United States Senate, May 2024. https://www.schumer.senate.gov/imo/media/doc/Roadmap_Electronic1.32pm.pdf.
113. Rhiannon Williams and James O'Donnell, "We finally have a definition for open-source AI," MIT Technology Review, Aug. 22, 2024. https://www.technologyreview.com/2024/08/22/1097224/we-finally-have-a-definition-for-open-source-ai.
114. Steven Vaughan-Nichols, "We're a big step closer to defining open source AI - but not everyone is happy," ZDNet, Aug. 23, 2024. https://www.zdnet.com/article/were-a-big-step-closer-to-defining-open-source-ai-but-not-everyone-is-happy.
115. Caitlin Andrews, "The outlook for AI safety regulation in the US," IAPP, Feb. 12, 2025. https://iapp.org/news/a/the-outlook-for-ai-safety-in-the-u-s-.
116. Adam Thierer, "California and Other States Threaten to Derail the AI Revolution," R Street Institute, May 2, 2024. https://www.rstreet.org/commentary/california-and-other-states-threaten-to-derail-the-ai-revolution.
117. Adam Thierer, "AI Policy in the Trump Administration and Congress after the 2024 Elections," R Street Institute, Nov. 7, 2024. https://www.rstreet.org/commentary/ai-policy-in-the-trump-administration-and-congress-after-the-2024-elections.

Identifying Policy Priorities, Emerging Technological Solutions, and Best Practices

There are numerous reasons to be optimistic about the future of AI development and governance. First, open-source and closed-source AI are not mutually exclusive options; both offer substantial benefits, and harnessing their strengths together is essential. Second, ongoing efforts to develop hybrid solutions that balance openness with safety and security demonstrate a willingness to adapt existing governance and business frameworks. Third, policymakers and industry leaders already recognize the critical role open-source AI plays in driving innovation and sustaining America's leadership in technology. Finally, the wealth of knowledge and best practices that have been accumulated over decades of open-source development provide a strong foundation for addressing current challenges. By drawing from this expertise and fostering creative, flexible approaches to governance, innovation, and technological solutions, the United States is well-positioned to navigate the open-source AI landscape.

The following recommendations highlight immediate priorities that the second Trump administration, the 119th Congress, industry leaders, researchers, and the open-source community should pursue.

Policy Priorities

1. **Develop and Issue Clear Guidelines for Secure Open-Source AI Deployment.** The White House should direct NIST, the Cybersecurity and Infrastructure Security Agency, and the NTIA to collaboratively establish transparent, risk-based guidelines that clarify best practices for the use and deployment of open-source AI models, systems, tools, and resources.[118] These guidelines should be voluntary, adaptable, and tailored to the model's potential impact and application. For higher-risk contexts, such as open-weight models used in critical infrastructure (e.g., energy grid management), additional security and testing considerations may be necessary to mitigate risks. Guidelines could also include a checklist of rigorous pre-release testing protocols to help developers identify and address vulnerabilities before deployment. Rather than imposing a formal approval process, this elective approach would provide industry-aligned best practices to enhance security and accountability while preserving flexibility in AI development.

2. **Invest in Public–Private Partnerships for Open-Source AI Model Validation.** Congress should fund public–private partnerships to develop tools and protocols for assessing the safety, transparency, and reliability of open-source models.[119] These partnerships or initiatives could be modeled after the Defense Advanced Research Projects Agency's existing cybersecurity initiatives, incentivizing collaboration between the public, private firms, government agencies, and academic institutions to create scalable validation frameworks.[120] For example, a new program could facilitate the development of automated systems that verify

118. Haiman Wong and Brandon Pugh, "Key Cybersecurity and AI Policy Priorities for Trump's Second Administration and the 119th Congress," R Street Institute, Jan. 6, 2025. https://www.rstreet.org/research/key-cybersecurity-and-ai-policy-priorities-for-trumps-second-administration-and-the-119th-congress.
119. Ibid.
120. Jonathan Greig, "DARPA awards $14 million to semifinal winners of AI code review competition," The Record, Aug. 14, 2024. https://therecord.media/darpa-awards-14-million-ai-code-review.

the security and safety of open-source contributions. Furthermore, Congress should consider offering tax incentives for startups specializing in scaling model-validation capabilities, particularly those focusing on open-source, AI-specific risks, such as adversarial vulnerabilities.[121] By investing in these initiatives, Congress can ensure that the growing volume of open-source AI projects is matched by robust and scalable validation mechanisms.

3. **Implement Risk-Tiered Liability Shields for Open-Source AI Development.** Congress should consider establishing liability protections that correspond with the risk levels associated with different types of open-source AI projects and applications.[122] Under this framework, developers of lower-risk models, such as tools for educational purposes, could benefit from broad liability shields that encourage innovation while limiting their legal exposure in cases of third-party misuse. This approach would protect developers by offering clearer legal boundaries and reducing uncertainty.[123]

Emerging Technological Solutions

1. **Develop and Implement Embedded Provenance Tracking.** Organizations should develop embedded provenance tracking systems to enhance transparency and accountability in open-source AI development.[124] These systems could use cryptographic tagging or distributed ledger technologies to provide an immutable record of updates, contributors, and deployment contexts.[125] For example, embedding provenance tracking into Hugging Face repositories could allow developers to verify and audit contributions in real time, ensuring a clear history of changes and reducing the risk of tampering. These systems could deter malicious actors while bolstering trust and accountability across open-source communities.

2. **Deploy AI-Driven Anomaly-Detection and Behavioral Analysis Systems.** Open-source AI systems incorporate AI-driven anomaly-detection tools to proactively identify deviations from expected behavior.[126] An anomaly-detection system integrated into an open repository on Hugging Face, for instance, could flag unusual activities, such as spikes in downloads or malicious code commits. These systems could also leverage machine learning classifiers trained on historical misuse patterns to distinguish benign anomalies from suspicious and malicious activities.[127] This continuous monitoring would enable timely intervention, enhancing the security and reliability of open-source AI projects.

121. Jeff Campbell, "U.S. Tax Reform Can Fuel AI and Cybersecurity Innovation," Cisco Blogs, Sept. 15, 2024. https://blogs.cisco.com/news/u-s-tax-reform-can-fuel-ai-and-cybersecurity-innovation.
122. Kathryn Bosman Cote, "Outsmarting Smart Devices: Preparing for AI Liability Risks and Regulations," *San Diego International Law Journal* 101 (July 3, 2024). https://digital.sandiego.edu/cgi/viewcontent.cgi?article=1350&context=ilj.
123. Dylan Walsh, "The legal issues presented by generative AI," MIT Management Sloan School, Aug. 28, 2023. https://mitsloan.mit.edu/ideas-made-to-matter/legal-issues-presented-generative-ai.
124. Harris. https://techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions; Amruta Kale et al., "Provenance documentation to enable explainable and trustworthy AI: A literature review," *Data Intelligence* 5:1 (Winter 2023), pp. 139-162. https://direct.mit.edu/dint/article/5/1/139/109494/Provenance-documentation-to-enable-explainable-and.
125. "AI Output Disclosures: Use, Provenance, Adverse Incidents," National Telecommunications and Information Administration, March 27, 2024.https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/information-flow/ai-output-disclosures.
126. Joel Barnard and Cole Stryker, "What is anomaly detection?," IBM, Dec. 12, 2023. https://www.ibm.com/think/topics/anomaly-detection.
127. Ibid.

3. **Design Adaptive Model Guardrails.** Developers should create and implement adaptive guardrails in open-source AI models that are capable of dynamically responding to emerging risks and misuse patterns.[128] These guardrails could serve as real-time filters that prevent harmful outputs or actions by refining thresholds based on user interactions. For example, a large language model could block or flag harmful content by learning from previously flagged outputs and adjusting its filters accordingly. Reinforcement learning techniques could further enhance these guardrails, ensuring they evolve with emerging threats.

Best Practices for the Open-Source AI Community

1. **Expand Existing Best Practices for Open-Source Libraries, Packages, and Software Supply Chains.** The open-source AI community should build upon established best practices by incorporating rigorous cybersecurity measures, such as sandboxing, dependency management, and regular vulnerability assessments. Moreover, the open-source community could adapt existing software supply chain security controls to AI-specific challenges, such as model integrity checks, to mitigate risks associated with using open-source models.[129]

2. **Encourage Voluntary Adoption of Copyleft Agreements in AI Development.** To ensure that derivative works and contributions remain open and accessible, the open-source AI community should promote the use of licenses with copyleft features, such as the GNU GPL or the Lesser General Public License.[130] Specifically, the owners of open-source AI datasets, models, tools, and resources could model this approach by applying these licenses to the projects they make available. Open-source AI repositories could also include templates for applying copyleft licenses to simplify the process for contributors and promote greater adoption. This approach would prevent proprietary capture and ensure that the broader community could benefit from collaborative advances.

3. **Establish Community-Driven Accountability Mechanisms.** The open-source AI community should implement peer-review systems and transparent contribution-tracking mechanisms to ensure responsible AI development.[131] For example, community-driven reporting and moderation boards could review flagged issues or concerns and maintain a transparent record of resolutions. This decentralized approach would ensure shared oversight and responsibility, aligning with the collaborative ethos of open-source development.

Through sustained commitments and collaborative efforts among policymakers, industry, researchers, and the open-source community, these recommendations hold the potential to ensure that open-source AI remains a catalyst for secure innovation and technological progress.

128. Jinwei Hu et al., "Adaptive Guardrails For Large Language Models via Trust Modeling and In-Context Learning," arXiv, Aug. 16, 2024. https://arxiv.org/html/2408.08959v1.
129. Jack Cable, "With Open Source Artificial Intelligence, Don't Forget the Lessons of Open Source Software," Cybersecurity & Infrastructure Security Agency, July 29, 2024. https://www.cisa.gov/news-events/news/open-source-artificial-intelligence-dont-forget-lessons-open-source-software.
130. Xinwei Guo, "Copyleft for Alleviating AIGC Copyright Dilemma: What-if Analysis, Public Perception and Implications," arXiv, Feb. 19, 2024. https://arxiv.org/html/2402.12216v1; Steffen Herbold et al., "Legal Aspects for Software Developers Interested in Generative AI Applications," arXiv, April 25, 2024. https://arxiv.org/html/2404.16630v1.
131. Megan Shahi et al., "Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone," Center for American Progress, Feb. 1, 2024. https://www.americanprogress.org/article/generative-ai-should-be-developed-and-deployed-responsibly-at-every-level-for-everyone.

Conclusion

Open-source AI is indispensable for maintaining America's technological leadership. By lowering barriers to entry, open-source AI empowers startups, academic institutions, and independent developers to drive advances that might otherwise be confined to a handful of technology companies. This decentralization not only fosters competition but also mitigates the risk of over-reliance on a few technology giants, reduces vulnerabilities in critical systems, and ensures a more resilient ecosystem. Initiatives like Google's Project Oscar and CrewAI exemplify how open-source AI continues to evolve, streamlining workflows and expanding access to tools like AI agents, which are poised to shape the next frontier of AI development.[132] These efforts also illustrate how the distinctions between open- and closed-source AI are increasingly blurred, as companies seek to leverage the strengths of both approaches to drive innovation.

On the world stage, competition and open-source AI are not only advantages—they are imperatives. As adversarial nations like China and Russia pursue centralized AI development strategies, America must leverage its open-source ecosystems to sustain its technological edge. China's centralized approach, characterized by government-directed priorities and heavy investment in state-aligned AI initiatives, is complemented by strategic open-source contributions, such as the release of DeepSeek's R1 model.[133] These calculated efforts enhance China's global influence, foster innovation, and attract international collaboration, all while maintaining rigid state oversight. This duality underscores the critical need for America to strike its own strategic balance between openness and safeguarding its national security and technological leadership.

The stakes of open-source AI governance and development are profound. While the policy priorities, emerging technological solutions, and best practices identified in this study provide a foundation, addressing the complexities of an increasingly globalized AI landscape requires further action. International communication and coordination that shapes shared norms and expectations is vital for navigating the evolving risks posed by borderless, open-source ecosystems.[134] Policymakers must anticipate the evolving challenges of advanced AI systems and foster innovation and interdisciplinary cooperation across public and private sectors to navigate these complexities effectively. If open-source AI is supported and guided intelligently, America has a unique opportunity to channel its culture of innovation and entrepreneurship into a force that strengthens national security, drives economic growth, and solidifies its position as a global leader.

If open-source AI is supported and guided intelligently, America has a unique opportunity to channel its culture of innovation and entrepreneurship into a force that strengthens national security, drives economic growth, and solidifies its position as a global leader.

About the Author

Haiman Wong is a resident fellow in the Cybersecurity and Emerging Threats team at the R Street Institute. Her research examines the intersection of cybersecurity and emerging technologies, including artificial intelligence, edge computing, and connected vehicles.

132. "Oscar, an open-source contributor agent architecture," Google Open Source, last accessed Jan. 16, 2025. https://go.googlesource.com/oscar/+/refs/heads/master/README.md; Vanna Winland et al., "What is crewAI?," IBM, Aug. 2, 2024. https://www.ibm.com/think/topics/crew-ai.
133. Zeyi Yang, "Why Chinese companies are betting on open-source AI," MIT Technology Review, July 24, 2024. https://www.technologyreview.com/2024/07/24/1095239/chinese-companies-open-source-ai.
134. Adam Thierer, "Existential Risks and Global Governance Issues Around AI and Robotics," R Street Institute, June 15, 2023. https://www.rstreet.org/research/existential-risks-and-global-governance-issues-around-ai-and-robotics.