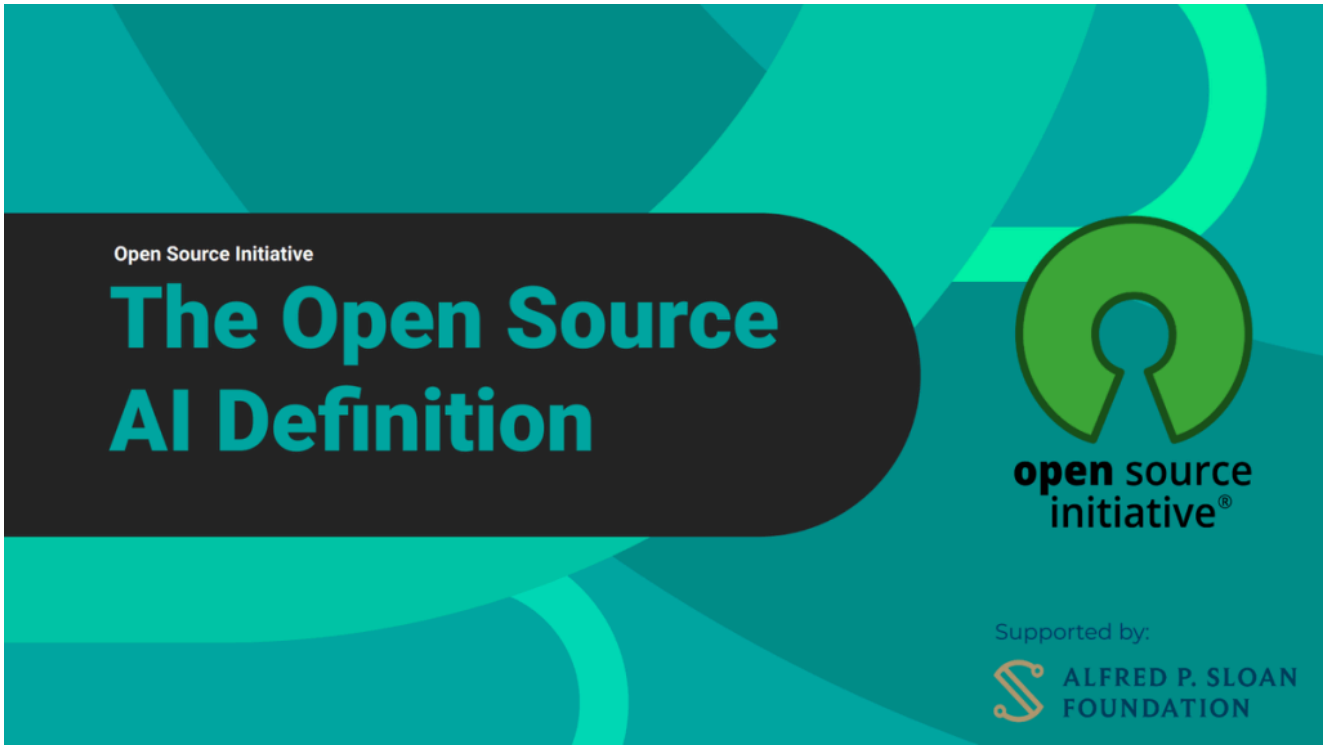
AI / OPEN SOURCE

The Case Against OSI's Open Source AI Definition

The chorus of criticism around OSI's open source AI definition continues, with a few people even suggesting the stakes aren't just for AI systems but the future of open source itself.

Dec 5th, 2024 6:00am by David Cassel



Last month the Open Source Initiative released its official definition for open source AI. The website notes that the definition has already been endorsed by at least 20 organizations, including Suse, Mozilla and the Eclipse Foundation.

But there's also some growing discontent. A blog post by the Software Freedom Conserv[illegible]n said "substantial acrimony" filled the

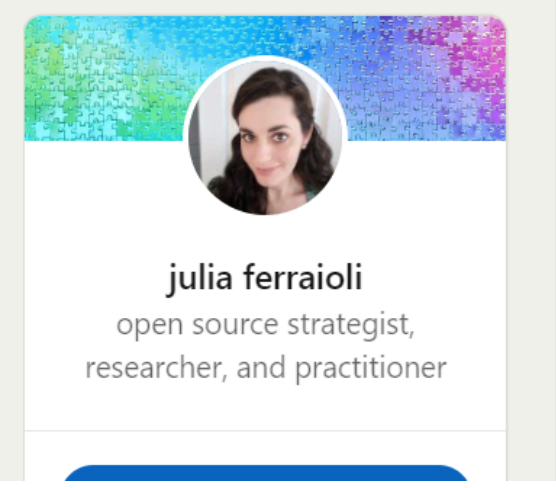
information about their training data so that a 'skilled person can recreate a substantially equivalent system using the same or similar data,' which goes further than what many proprietary or ostensibly Open Source models do today."

Carlo Piana, a board member and former board chairman of the Open Source Initiative (OSI), told TNS that the definition was leaving room for future flexibility, since "our collective understanding of what AI does, what's required to modify language models, is limited now. The more we use it, the more we'll understand."

But the chorus of criticism is continuing, with a few people even suggesting the stakes aren't just the licensing classifications for AI systems — but the future of open source itself.

The Case Against

The FOSS news site LWN.net quickly rounded up some early statements of concern about training data, from Amazon Web Services' open source technical strategist Tom Callaway to AWS open source AI/ML strategist Julia Ferraioli. "It damages every established understanding of what 'open source' is," Callaway later posted on LinkedIn, "all in the name of hoping to attach that brand to a 'bigger tent' of things ... I am deeply disappointed."



This isn't about AI. It may have started with AI, but that's not what is at stake.

It's about the very meaning of open source, and the OSI's insistence on undermining it on a technical AND cultural level. I worry about the opportunities that future generations will not have as technology trends closed. I worry about future contributors who will decide to steer clear because of renewed toxicity. I worry about voices silenced, progress reversed, potential lost.

Zoom

5. Why LLMs Within Software Development May Be a Dead End

RedMonk industry analyst Stephen O'Grady is also skeptical. "I do not believe the term open source can or should be extended into the AI world," O'Grady wrote last month.

Amanda Brock, CEO of the nonprofit OpenUK, also recently said in a LinkedIn post that "I don't agree with what's being done. I am not alone. Most (not all) senior open source software engineers I have spoken to agree with me.

"Sometimes we disagree. That's ok. Hearing all voices and concerns is how we demonstrate a healthy discussion of the right issues is taking place. So please, be kind."

And posting in response was Bruce Perens himself, the creator of the original open source definition (and an original co-founder of the OSI). Perens stated unequivocally that the open source AI definition "is flawed," and that OSI "hasn't done a great job and so weren't necessarily the best team to do this. In my opinion, the result is less than Open Source."

Reached for comment by The New Stack, Perens went further, saying "I don't think the Open Source AI Definition was a good idea at the start, and it hasn't turned out well. The plain old Open Source Definition that we've had for 26 years can be applied to AI, and I think it would have been a better idea for the OSI to release a procedure on how to do that."

Perens thinks that original definition could be applied to both AI software and its training data — and like others, Perens believes the training data *is* the source code. ("This is complicated by the fact that the AI may be real-time and the source code is thus ever increasing. But that's how some machine learning systems work and the 26-year-old Open Source Definition still works with them.")

s a 4:25 he Open Source brand," warning of the ssibility that "that might be the demise of OSI and Open Source. Now that we

troubling rights that the old Open Source Definition gives you." Does that lead us

What About Medical Data?

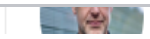
The OSI's official FAQ for the definition makes another argument against open training data. They "want Open Source AI to exist also in fields where data cannot be legally shared, for example medical AI."

Even the venerable Free Software Foundation (FSF) acknowledges the possibility of "valid moral reasons for not releasing training data, such as personal medical data." But the FSF concludes this just leads to the unusual scenario of non-free applications where "using it could be ethically excusable if it helps you do a specialized job that is vital for society, such as diagnosing disease or injury."

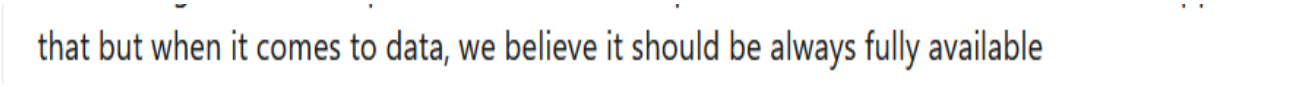
The FSF has been working on their own definition since May, but in October were ready to declare that machine-learning applications "are only partially software," and that the FSF's eventual criteria "will require the software, as well as the raw training data and associated scripts, to grant users the four freedoms." (Freedom No. 1? "[T]he freedom to study how the program works...")

FSF executive director Zoë Kooyman confirmed to The New Stack last week that the group's opinion has not changed on training data. "We believe that we cannot say a ML application is free unless all its training data and the related scripts for processing it respect all users, following the four freedoms." Also weighing in was the U.N.-endorsed Digital Public Goods Alliance (whose board members include the government of Sierra Leone and UNICEF). Their position? An upcoming AI-related update will propose that their own standards "continue requiring open training data for AI systems to be considered Digital Public Goods."

There were also some thoughts from Nextcloud — which released its own Ethical AI Rating systems in 2023.



Founder and CEO at Nextcloud GmbH



Zoom

Johannes Poortvliet

Co-founder, Director Communications @Nextcloud | Regain control over your data 💪 | We're gr...

2w

The Open Source Initiative (OSI) today released 1.0 of their #OpenSource #AI definition. Now it's 'just' a 1.0 and there has already been a fair bit of criticism, but as their director Stefano Maffulli made clear to us in a call, this is the start of a journey and it is a definition we can build on. In case of Nextcloud, that means we keep our Ethical AI Definition, which does contain a requirement to open up all data, in parallel with the OSI definition.

Zoom

The Open-Data LLMs

The official FAQ for the definition argues that requiring fully open training data "would relegate Open Source AI to a niche of AI trainable only on open data... That niche would be tiny, even relative to the niche occupied by Open Source in the traditional software ecosystem."

But is that changing? On Nov. 13, 2024, French AI company Pleias released "the largest fully open multilingual dataset for training LLMs, containing over 2 trillion tokens of permissibly licensed content with provenance information."

This year, Ai2, the Seattle-based nonprofit AI research institute founded in 2014 by the late Paul Allen, has also been releasing LLMs with open training data. And on Oct. 31, AMD announced they'd release a series of "fully open 1 billion parameter language models," open sourcing all training details. AMD's announcement says this "allows organizations to tailor the model's architecture and training process to meet their unique requirements" and "empowers a diverse community of users, developers, and researchers to explore, utilize, and train state-of-the-art large language models."

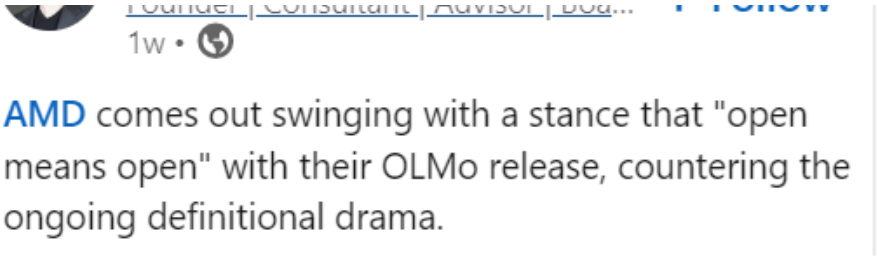
Long-time open source advocate Karsten Wade thinks chipmakers like AMD (and

roper Open Source AI will outpace its

sed sourced competitors.

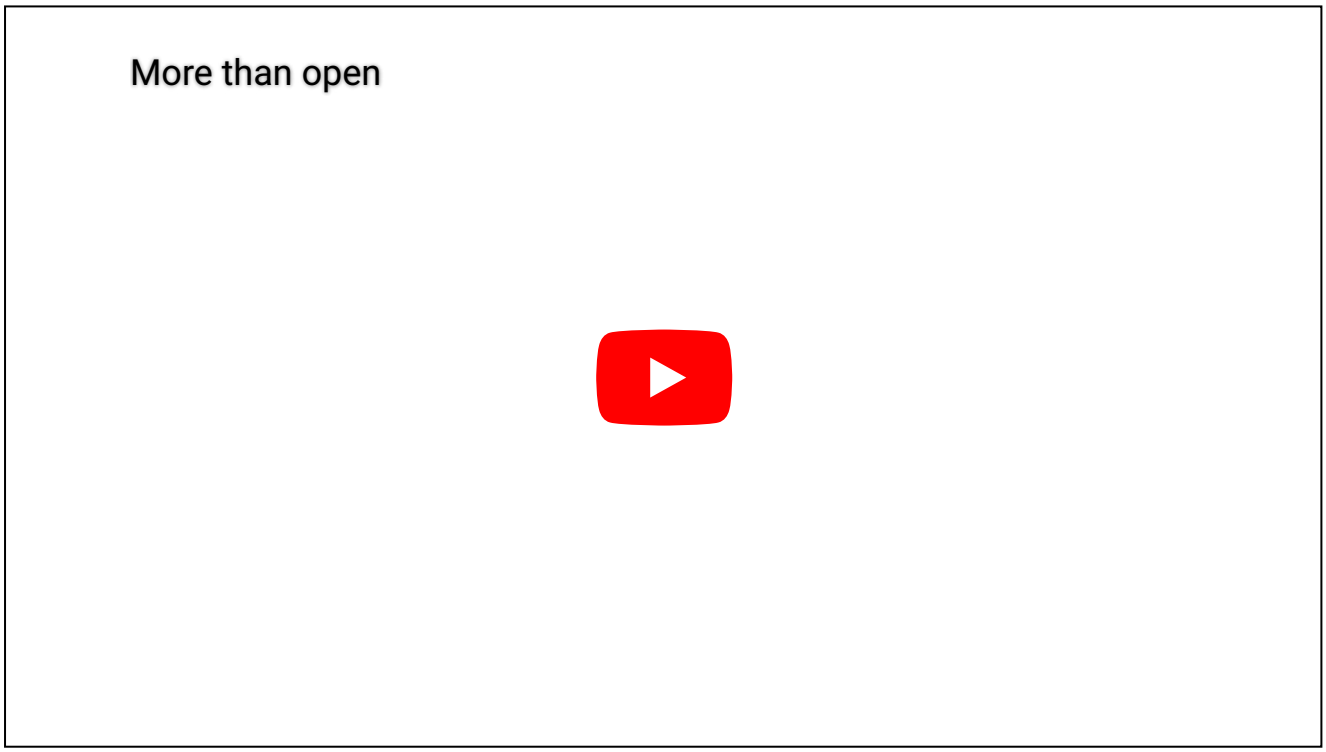
THENEWSTACK

were," and, with October's announcement of AMD's OLMo, "now, most definitely."



Zoom

And on Oct. 31st, Ai2 released a video wherein team members shared their own thoughts — such as, "In order to be truly open, every part of the pipeline has to be open, in my opinion."



YOUTUBE.COM/THENEWSTACK
Tech moves fast, don't miss an episode. Subscribe to our YouTube channel to stream all our podcasts, interviews, demos, and more.

SUBSCRIBE

4:25

concerns down to simple talking points.

their profiles...") Johnston criticizes the methodology used to create the definition ("co-design" rather than "rough consensus"), arguing that the process ultimately ignored working groups which had actually favored a requirement for open training data. Panning the OSI's final definition, he writes that "Their FAQ deceptively divides data into four categories, only to accept any data, or none."

But more importantly, Johnston notes calls by some Debian developers for a resolution that the definition isn't compliant with Debian's Free Software Guidelines. Johnston believes this could lead to scenarios where "OSAID-compliant software would be rejected by Debian and its dependent distros." Even without a formal resolution, Johnston predicted to The New Stack that a statement of disapproval from Debian could prompt a thousand more people to stand in opposition.

So what happens next? In an email interview, Johnston said he ultimately wants a definition that will "allow future generations of practitioners and models to stand on the shoulders of the last, rather than rendering us subservient to foreign vendors of opaque, open-washed models." Transparent training data lets users confirm their models weren't trained on proprietary data, according to Johnston, but that's not the only advantage he sees. "The transparency of training data allows minorities like the co-designers to assess and address ethical issues including fairness and bias."

Looking to the future, Johnston said he still remains optimistic about the future of the "open source" definition — as it is, or with a (different) update that also includes open AI training data. And he feels like public opinion is on his side. "There's a small risk that the OSAID fork takes hold before being repealed, but it's already been rejected by the existing Open Source industry, including the Free Software Foundation and Debian, on which it was based."

And so the pushback continues. In his blog post, the Software Freedom Conservancy's Bradley Kuhn argued that the new definition "erodes" the meaning of open source, writing that it's "the moment that software freedom advocates fe[illegible] 'open source' — with which OSI was

4:25

platform.

Correction: *This article has been corrected to update Carlo Piana's status as an OSI board member.*

TNS

David Cassel is a proud resident of the San Francisco Bay Area, where he's been covering technology news for more than two decades. Over the years his articles have appeared everywhere from CNN, MSNBC, and the Wall Street Journal Interactive...

Read more from David Cassel →