Show Your Training Data: New AI definition challenges what it means to be open source

October 28, 2024
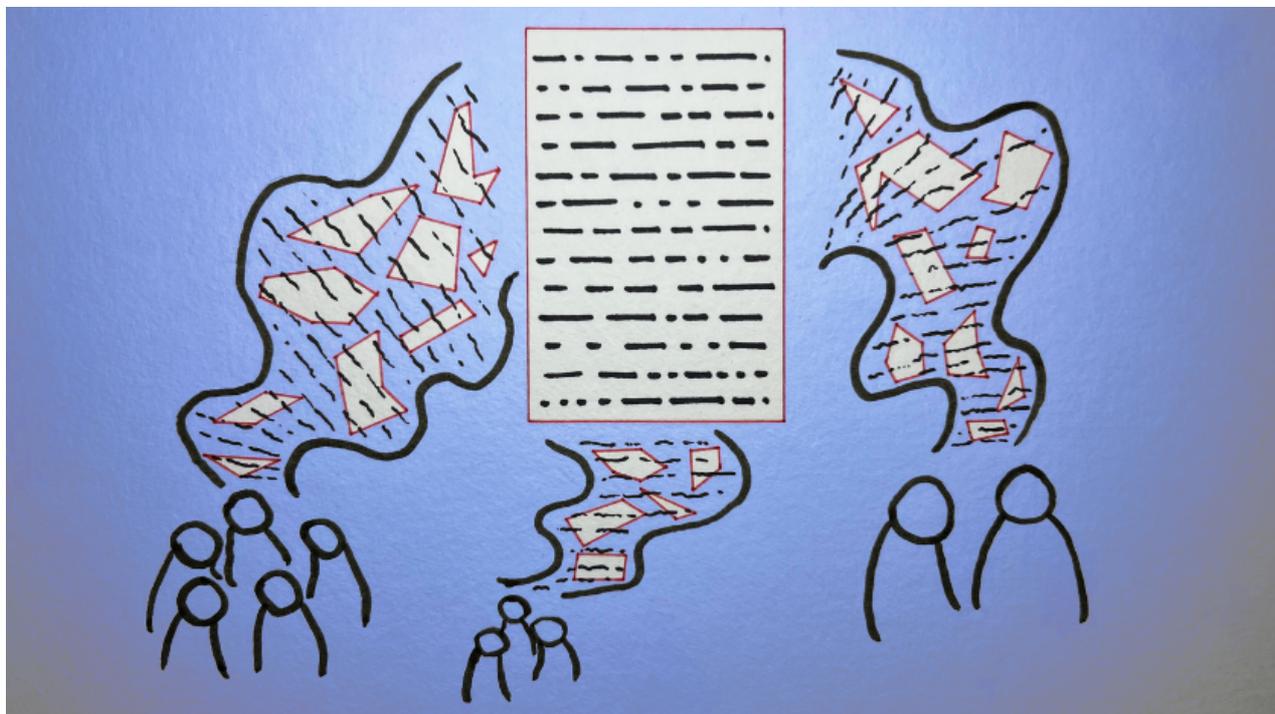
Yasmine Boudiaf & LOTI / Better Images of AI / Data Processing / CC-BY 4.0

Open source has been a driving factor behind the ever-continuing rise in AI. From the early days of OpenAI (before it closed its doors) and Meta's ever-growing herd of llamas, open source AI has powered tools and technologies used by industries all over the world.

Yet, until now, the term 'open source AI' has remained nebulous, creating grey areas in licensing, innovation, and trust.

The Open Source Initiative (OSI), the nonprofit behind the definition of open source software has spent the best part of two years gathering insights and ideas to create a community-led definition as to what constitutes open source AI.

The Open Source Definition (OSAID) v.1.0, published at the All Things Open 2024 conference, offers a standard to validate whether or not an AI system can be deemed truly open source.

"The co-design process that led to version 1.0 of the open source AI definition was well-developed, thorough, inclusive and fair," said Carlo Piana, OSI board chair. "We're energised about how this definition positions OSI to facilitate meaningful and practical open source guidance for the entire industry."

What is open source AI: OSI's definition

An open source AI system, as defined by the OSI, is an AI system made available under terms and in a way that grants the freedom to:

- Use the system for any purpose and without having to ask for permission.

- Study how the system works and inspect its components.

- Modify the system for any purpose, including to change its output.

- Share the system for others to use with or without modifications, for any purpose.

The freedoms apply to both a fully functional system and its discrete elements. To exercise these freedoms, access to the preferred form for making modifications is considered essential.

What does a definition mean for open source AI?

The need to define what constitutes open source AI is ultimately about transparency. The team are OSI have sought to set clear standards so AI systems can not only be properly evaluated but also to prevent misuse or dilution of the term.

"This is a starting point for a continued effort to engage with the communities to improve the definition over time as we develop with the broader open source community the knowledge to read and apply OSAID v1.0," said Stefano Maffulli, executive director of the OSI.

The definition has already garnered supporters, with the likes of Mozilla, AI research lab Eleuther AI, and the engineering team at Bloomberg have all endorsed the definition.

"Transparency is at the core of EleutherAI's non-profit mission. The Open Source AI Definition is a necessary step towards promoting the benefits of open source principles in the field of AI," said Stella Biderman, executive director at the EleutherAI Institute. "We believe that this definition supports the needs of independent machine learning researchers and promotes greater transparency among the largest AI developers."

Headache for Meta?

The definition is, however, likely to upset some people in the AI community, particularly the companies and researchers that open source their models but opt not to disclose the training data used to train the model.

The rules require open source models to provide enough information about their training data so that a 'skilled person can recreate a substantially equivalent system using the same or similar data.'

One outlet that rules around training data likely to be irked is Meta. The Facebook and Instagram parent has followed its own path, opting to publish its AI models, like the recently released Llama 3.2, as open source, meaning enterprises don't have to pay to access the model.

In July, CEO Mark Zuckerberg posted a lengthy affirmation of the company's favourability towards open source, saying: "Open source AI represents the world's best shot at harnessing this technology to create the greatest economic opportunity and security for everyone."

Meta also counts an open source champion in one Yann LeCun, the company's chief AI scientist, a vocal advocate for open source AI development.

However, Meta has routinely opted not to disclose information on the selection of data sources its engineers used to train its model — instead opting to release its Llama models under a modified license which does not provide any information about the underlying data — which would go against the OSI definition for an AI model to be considered open source.

"[The training requirement] is the starting point to addressing the complexities of how AI training data should be treated, acknowledging the challenges of sharing full datasets while working to make open datasets a more commonplace part of the AI ecosystem," said Ayah Bdeir, AI strategy lead at Mozilla.

"This view of AI training data in open source AI may not be a perfect place to be, but insisting on an ideologically pristine kind of gold standard that will not actually be met by any model builder could end up backfiring."

Seth Dobrin, founder and CEO of Qantm AI and former chief AI officer of IBM, likened Meta's use of the term open source to describe its models as a "marketing ploy," adding, "The OSI definition makes that crystal clear."

"While I support the need for companies to maintain proprietary IP as a VC and a former tech exec, if you are going to put something in the open source community, it should be fully open-source," Dobrin said. "I am unsure where the ambiguity has come from, as all of these AI systems are built on top of fully open-source software, and everyone building the AI systems knows what open source is and isn't. But I am glad that OSI has stepped in and taken a stance to remove the ambiguity."

Capacity has contacted Meta for comment.

From definitions to aspirational statements

The definition also comes as the Software Freedom Conservancy (SFC) released its own aspirational statement on AI systems — specifically focusing on programming assistants like GitHub Copilot.

While the SFC's approach is more stringent, requiring all components including training data to be free and open source software (FOSS) licenced, it positions itself as an idealistic vision rather than a practical standard.

"While our proposal may seem unrealistic, nearly every proposal in the history of FOSS has seemed unrealistic — until it happened," the organisation noted.