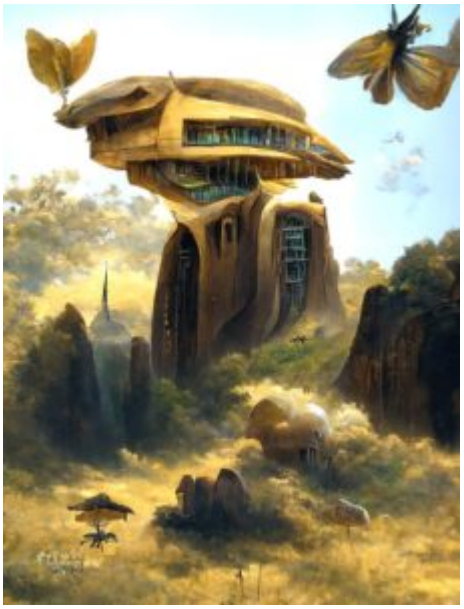
Kluwer Copyright Blog

Open Source AI – definition and selected legal challenges

Yaniv Benhamou (University of Geneva) · Monday, April 15th, 2024



"Artificial intelligence architecture" by Wendelin Jacober is marked with CC0 1.0.

In the generative AI era, there is a proliferation of open source claims (*i.e.* operators that claim to release AI models sufficiently open to be part of the open source or open innovation movement, as opposed to closed-source model), such as open source and open access foundation models (e.g. Google BERT, Meta LLaMA Large Language Model (LLM), OpenAI API). While an open source approach to AI is valued as important for fostering innovation and competition, the notion raises many questions: *(1)* What is open source AI? Which elements shall be available as open source? Can it be everything (*i.e.* all elements composing the AI model) or only specific components (e.g. training data, weighting factors)? *(2)* What is the intersection and the difference between 'open data' and 'open source'? *(3)* What is the effect of open source licenses on the AI model that uses only some open source components? *(4)* What is the liability of open source contributors? *(5)* What is the impact of new regulation on open source AI?

This post follows a panel organized by the Global Partnership on AI (GPAI), with McCoy Smith, Shun-Ling Chen, Yaniv Benhamou(panellists) and Yann Dietrich (moderator). It covers some of the basics on open source AI focusing on its definition and legal challenges.

1. **What is Open Source AI**

Open source AI refers to the use of open source components within an AI model, *i.e.* elements composing the AI model (e.g. documentation, software codes, copyrighted training data) that are under open source licenses (OSL), i.e. licenses that comply with the open source definition (in brief that allow software or data to be freely used, studied, modified, and shared, also known as the "four freedoms").

There are many AI models that claim to be open source – just as there are multiple forms of open

source licenses, from permissive licenses (e.g. MIT License or Apache License) to less permissive licenses (e.g. GNU GPL or the BSD license). So open source AI exists across a spectrum of openness, from fully open to fully closed. The level of openness depends on how much the inner working of the AI model is shared with the public, i.e. whether all or certain components of the AI model are made publicly available (e.g. documentation, methods, weighting factors, information on the model architecture or usage). A recent report ranked AI models based on their level of openness and 13 components composing AI models, with Meta's Llama2 being the second lowest ranked (due to a permissive license but with additional commercial terms for users with more than 700 million monthly active users), and ChatGPT being the lowest ranked (which explains why Elon Musk is suing OpenAI for breach of contract as a *de facto* closed-source model).

So when we speak generally about open source AI, it would be more accurate to instead specify the level of openness based on the open components (e.g. "open code AI", "open training Data AI", "open weighting factors AI" etc.).

This being said, the need to define what constitutes open source AI remains, not only to avoid stakeholders to use the terms for marketing purposes only (kind of "open source washing") but also to know which legal consequences are attached to such qualification, such as the legal effects of OSL on the AI models' components that are under proprietary licenses, limitations of liability and exception regimes (e.g. AI Act providing exceptions to transparency and documentation for open source AI).

The exact definition of what constitutes open source AI is still subject to discussion. If we rely on the European regulation, in particular on the definition of the new AI Act, "free and open source AI" is defined as "*AI components [that] are made accessible under a free and open-source license*" (recital 89) in particular their "*parameters, including the weights, the information on the model architecture, and the information on model usage*" (recital 102) and open source AI components cover the software, the data and the AI models (including tools, services or processes of an AI system). However, the scope of these exceptions is limited as it does not exempt AI systems that are monetized (i.e. provided against a price or otherwise monetised, including through the use of personal data), or considered high-risk (recital 103-104). Unfortunately, the AI Act does not specify the number of components (threshold) that shall be made available to qualify as open source AI.

According to experts of the open source community, merely releasing a model under an open source license (e.g. through open repositories) without providing access to other components should not qualify as open source (but eventually as "open access AI"). So, AI models should qualify as open source only if they release different components beyond the simple releasing of the model (e.g. documentation, methods, weighting factors, information on the model and on the architecture). Finally, the Open Source Initiative (OSI) is currently working on a definition for open source AI. Its "Open Source AI Definition – draft v. 0.0.6" requires at least the following 3 components to be available to the public under terms that grant the "four essential freedoms" (use, study, modify, share): data (including training data, methodologies and techniques), code (including the model architecture) and model parameters (including the weighing factors).

2. **Intersection between open data and open source software**

Given the importance of data when it comes to AI, one may wonder whether open source comes with open data?

AI models rely on a massive amount of data (training data), some of which are under open license terms. Indeed, open source AI components do not relate only to software but also to data. So this raises the question of the intersection between open source software and open data (in the sense of elements, software or data, under permissive licenses such as Google BERT under Apache or ChatGPT trained on Wikipedia data under CC). Three comments can be made about this arrangement.

First, not all training data are under permissive licenses, as some are just publicly available (like copyright images or texts that are publicly available, viewable but not reusable). Think of social media, whose data are scraped and used to train Large Language Models (LLM) (e.g. Reddit or X (Twitter) data for ChatGPT) and which try to ban AI data scraping via technical tools and contractual terms (see class action against OpenAI for privacy and copyright infringement).

Second, open data and software are not the only open source components of AI, as the different elements of an AI model can include also documentation, weighting factors or information on the model architecture). So it is better to speak about level of openness depending on these components – from fully open to fully closed open source.

Third, data also include non-copyrighted elements, such as personal data (e.g. social media data), databases and trade secrets (e.g. a dataset combining technical, machine-generated and mixed data). So, data may be subject to multiple, sometimes conflicting, legal regimes, such as copyright, trade secrets or data protection. This leads to fragmentation and has become a major challenge in the AI era. Solutions to address this issue include contractual mechanisms (e.g. open licenses that extend to non-copyrighted elements), as well as regulatory interventions (e.g. EU Digital Market Act and competition laws that force access to certain data, see below).

So when we hear the term “Open Source AI”, we usually think of the software (code or documentation), not necessarily the training data, the model itself or the weighting factors. But, given the spectrum of openness, open source software or open source AI does not necessarily come with open data: it can have only open code, documentation, weighting factors, architecture, open training data.

3. **What is the effect of open source licenses on AI models, such as their output?**

While all eyes are on “open source AI” and their level of openness, a less debated issue is the impact of open software or open data on the AI model. In particular, does the use of open software or open data make the whole AI model open, including the output of these models?

This relates to the propagating effect of certain open source licenses (OSL) that require any code deriving from software under OSL to remain under the same permissive type of license. This led for instance the FSF to sue Cisco Systems in 2008 for violating the GPL. It has major

repercussions in the AI context, as such propagation could render entire or some components of open source AI models fully open (e.g. when AI output qualify as derivatives of the input data).

However, we consider that there are good arguments to be very careful in the manner in which one approaches the definition of derivative in the AI context ("AI Derivatives") that differs from the software context. For instance, AI models involve several actors and are based on several components (see above, the OSI definition or the recent report relying on multiple components, each of which may or may not be under different license terms, such as OSL, and/or qualify as AI Derivatives).

4. **What is the liability of open source contributors?**

With open licenses, there are multiple contributors. This creates a contractual chain between the primary upstream developer and the downstream users (who can, depending on the applicable open source license, make copies or create derivatives).

This raises questions of liability. On the one hand, developers can be held liable (in tort) if the codes or the data are dysfunctional and cause harm or infringe rights. Illustrations of this include DAO being held liable to its users (USD 50 million) due to a vulnerable open source code and Canada Airline for its chatbot giving incorrect information to a traveller). On the other hand, downstream users can be held liable (contractual liability) if they do not respect the license terms. This happens, for instance, if they omit to mention the upstream developers when required as in the lawsuit developers vs Microsoft-Github/OpenAI-Copilot (some consider even that as a kind of "open source laundering"). Liability exclusions, like that in the MIT License stating that the software is provided "as is" with no warranty of any kind, are not valid in civil law jurisdictions for gross negligence, when the primary or derivative contributor knowingly or involuntarily causes damage.

With open source AI, there may be liability issues too, in particular for anyone participating in the contractual chain. The major difference, if any, between the AI and software context is the increased number of contributors, who may have participated in the AI lifecycle and who may be held liable for different acts and the impact on the whole AI lifecycle (e.g. most open source licenses provide a termination in case of breach of contract, which would impact the functioning of the AI model).

5. **What is the impact of new regulation on open source AI?**

In the EU, a number of regulations may impact Open Source AI, such as the EU AI Act, Data Act and Digital Market Act.

The **EU AI Act** may impact open source AI, as it makes requirements lighter for stakeholders that release their models under open source licenses. On a basic level, among open source AI, a

distinction is made between: *(i)* AI systems (deployed AI systems and applications, think ChatGPT) for which the AI Act does not apply, unless they represent a "*high risk*" and *(ii)* the underlying General Purpose AI (GPAI) models (pre-trained models, like GPT4) for which lighter transparency and documentation obligations apply ("open source exceptions"), unless they represent a "*systemic risk*" or monetize their services, i.e. provide technical support or services through a software platform, or use personal data for reasons other than improving security, compatibility or interoperability of the software. One **criticism** at least is that open source can get away with being less transparent and less documented than proprietary GPAI models, an incentive to use open licenses for actors seeking to avoid transparency and documentation obligations, while violating the spirit of open source.

The **EU Data Act** may impact open source AI, as it provides rules on how data sharing contracts shall be drafted, for instance to protect EU businesses from unfair contractual terms. It provides rules for B2B mainly, so it remains to be seen how it may impact general contracts addressed to an undefined number of third party users (such as open licenses, general terms of use of AI models towards end users or business terms of AI models towards business clients in relation to their APIs or other business products).

The **EU Digital Market Act** and competition law may also impact open source AI, as it could force access to data (e.g. certain training data and datasets under the *essential facilities* doctrine), which seems fine with copyright data, more difficult for personal data that shall be protected by privacy laws.

6. **Conclusion**

Open source AI models require different notions and terminology than open-source software, in particular as they are more complex in their composition. AI models are based on several components (e.g. from code to weighting factors and training data) and often involve several actors. Therefore, there is a need to understand what exactly open source AI means, and what are the legal effects of the associated license on the entire AI model. While many actors are calling their systems "open source AI" despite the fact that their license contain restrictions (e.g. Meta Llama2) and there is still debate, some regulations (e.g. AI Act) start referring to "free and open source AI" and the open source community is about to adopt a definition based on required components (data, code, model) to be released under OSL.

---

legal technology will increase for next year. With Kluwer IP Law you can navigate the increasingly global practice of IP law with specialized, local and cross-border information and tools from every preferred location. Are you, as an IP professional, ready for the future?

Learn how **Kluwer IP Law** can support you.

This entry was posted on Monday, April 15th, 2024 at 10:30 am and is filed under Artificial Intelligence (AI), Software
You can follow any responses to this entry through the Comments (RSS) feed. You can leave a response, or trackback from your own site.