Moving a step closer to defining Open Source AI

Stefano Maffulli - Executive director
Open Source Initiative

FOSDEM 2024

The objective for 2024

Open Source AI Definition version 1.0

Leave comments for this text

stating the intentions of this document; the Definition of Open Source AI itself; and a checklist to evaluate licenses.
We follow the definition of AI adopted by UNESCO:

An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

Preamble

Why we need Open Source Artificial Intelligence (AI)

Open Source has demonstrated that massive benefits accrue to everyone when you remove the barriers to learning, using, sharing and improving software systems. These benefits are the result of using licenses that adhere to the Open Source Definition. The benefits can be distilled to autonomy, transparency, and collaborative improvement.

Everyone needs these benefits in AI. We need essential freedoms to enable users to build and deploy AI systems that are reliable and transparent.

How we can get the benefits of Open Source AI

A precondition for a system to be Open Source software is that developers must have unrestricted access to the "preferred form to make modifications to the work".
For AI systems, the preferred form to make modifications to the work depends on the specific kind of AI.

[Provide an example, based on machine learning?]

Out of scope issues

The Open Source AI Definition doesn't say how to develop and deploy an AI system that is ethical or responsible, although it doesn't prevent it. What makes an AI system ethical or responsible is a separate discussion.

What is Open Source AI

To be Open Source, an AI system needs to make its components available under licenses that individually grant the freedoms to:

- **Study** how the system works and inspect its components.

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

- **Modify** the system to change its recommendations, predictions or decisions to adapt to your needs.

- **Share** the system with or without modifications, for any purpose.

[Provide an example, based on machine learning?]

Checklist to evaluate licenses

TODO

Leave comments for this text

How Open Source came to be

1: Legal framework

Copyright applied to software, first.

This new artifact became privatized work.

Researchers complained.

2: Principles

The GNU Manifesto lays the ground to oppose privatization.

A community forms around these principles.

3: Licenses

Copyleft is a hack on copyright.

Incorporating the principles, serving as the Constitution of a forming community.

Golden Rule applied to AI

If I like an **AI system** I must be free to share it with other people.

What we've learned so far

- We need to define Open Source AI, in general, not just machine learning
- OECD's definition of AI is well accepted (with caveat!)

"An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to **generate outputs** such as predictions, content, recommendations, or *decisions* that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment." (2023)

Matching expectations

AI deserves to enjoy the benefits of Open Source

- autonomy
- transparency
- collaborative improvement
- ensuring the agency of the user

Policy makers, academia and industry are focusing on

- transparency
- trustworthiness
- reliability
- transparency
- explainability
- fairness
- safety etc

What basic freedoms do we need?

What is Open Source AI

To be Open Source, an AI system needs to be available under legal terms that grant the freedoms 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 to change its recommendations, predictions or decisions to adapt to your needs.
- **Share** the system with or without modifications, for any purpose.

What is the **preferred form to make modifications** to an AI system?

Getting the specifications

AI systems	List of components	Legal frameworks	Legal documents	Checklist
As defined by the OECD.	What elements are necessary to: - use - study - modify - share an AI system?	For each artifact, evaluate which laws apply. Some will be under "Intellectual Property" regimes, some will be under other regimes.	We'll match the components and the identified legal frameworks with the terms of the legal documents already in use, where available.	After repeating this exercise enough times, we'll be able to generalize the outcomes and write the specs to evaluate the freedoms granted.

Small working groups to analyze systems

For each in:

- Pythia
- Llama2
- BLOOM
- OpenCV
- Mistral
- Phi2
- Olmo
- …

- What do you need to give an input and get an output? (use)
- What do you need to give an input and get a different output? (modify)
- What do you need to understand why given an input, you get that output? (study)
- What do you need to let others give an input and get an output? (share)

What's the preferred form to make modifications to an AI system?

New in draft 0.0.5

Component	Necessary to Use	Necessary to Study	Necessary to Modify	Necessary to Share
Code				
…				
Data				
…				
Model				
…				
Other				
…				

Then the rest

- get the legal framework for each component
- get the legal documents
- analyze the documents
- write up a summary

Repeat for at least 4-5 AI systems, ideally not just LLMs and "Generative AI"

2024 timeline

	February	March	April	May	June …	… October
System testing work stream	Call For Volunteers + Activity Feedback and Revision	Virtual System Review Meetings Begin	Virtual System Review Meetings Continue	Virtual System Review Meetings END	Feedback Informs Content of OSI In-Person Stakeholder Meeting	Monthly Virtual Meetings
Stakeholder consultation work stream	Bi-Weekly Virtual Public Townhalls	Bi-Weekly Virtual Public Townhalls	Bi-Weekly Virtual Public Townhalls	Bi-Weekly Virtual Public Townhalls	Townhall + **OSI In-Person Stakeholder Meeting** (date + place TBD)	**Release version 1.0**
Release schedule	Draft 0.0.5	Draft 0.0.6	Draft 0.0.7	Draft 0.0.8	RC1	v. 1.0

Criteria for RC1 and v. 1.0

RC1

- Expected outcome of in-person meeting end May/early June!
- The draft is completed in all its parts
- The draft is supported by at least 2 representatives for each of the 6 stakeholder groups

version 1

- Expected outcome of in-person and online meetings through the summer/early autumn
- The draft is endorsed by at least 5 reps for each of the stakeholder groups
- Announced in late October

The stakeholders we're engaging with

System Creator	License Creator	Regulator	Licensee	End User	Subject
Makes AI system and/or component that will be studied, used, modified, or shared through an open source license (e.g., ML researcher in academia or industry)	Writes or edits the open source license to be applied to the AI system or component; includes compliance (e.g., IP lawyer)	Writes or edits rules governing licenses and systems (e.g. government policy-maker)	Seeks to study, use modify, or share an open source AI system (e.g. AI engineer, health researcher, education researcher)	Consumes a system output, but does not seek to study, use, modify, or share the system (e.g., student using a chatbot to write a report, artist creating an image)	Affected upstream or downstream by a system output without interacting with it intentionally; includes advocates for this group (e.g. people with loan denied, or content creators)
✅	✅	⚠️	✅	⚠️	⚠️
Enough to start	Enough to start	Leads to US, EU, Singapore, no commitment yet	Enough to start	Which org is squarely in this space?	ACLU, Algorithmic Justice League

It doesn't end with v. 1.0

We'll need to define rules for maintenance and review of the Definition

OSI's immediate next steps

- ~~more publicity to the process~~
  - ~~public discussion forum~~ https://discuss.opensource.org
  - bi-weekly townhalls **ONGOING**
  - more opportunities to volunteer
- ~~update project landing page~~
- reach out to more stakeholders
- raise funds for 2024 meetings
- setup the board for review and approval of v. 1.0 - **STARTED**

Draft v. 0.0.5 of the Open Source AI Definition

Open to public comments

https://opensource.org/deepdive/drafts

Become a member of OSI

https://members.opensource.org/join
Support more workshops in 2024

@ed@opensource.org

stefano@opensource.org

The Open Source Definition

Introduction

Open source doesn't just mean access to the source code. The distribution terms of open-source software must comply with the following criteria:

1. Free Redistribution

The license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale.

2. Source Code

The program must include source code, and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well-publicized means of obtaining the source code for no more than a reasonable reproduction cost, preferably downloading via the Internet without charge. The source code must be the preferred form in which a programmer would modify the program. Deliberately obfuscated source code is not allowed. Intermediate forms such as the output of a preprocessor or translator are not allowed.

3. Derived Works

The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.

4. Integrity of The Author's Source Code

The license may restrict source-code from being distributed in modified form *only* if the license allows the distribution of "patch files" with the source code for the purpose of modifying the program at build time. The license must explicitly permit distribution of software built from modified source code. The license may require derived works to carry a different name or version number from the original software.

5. No Discrimination Against Persons or Groups

The license must not discriminate against any person or group of persons.

6. No Discrimination Against Fields of Endeavor

The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a business, or from being used for genetic research.

7. Distribution of License

The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.

8. License Must Not Be Specific to a Product

The rights attached to the program must not depend on the program's being part of a particular software distribution. If the program is extracted from that distribution and used or distributed within the terms of the program's license, all parties to whom the program is redistributed should have the same rights as those that are granted in conjunction with the original software distribution.

9. License Must Not Restrict Other Software

The license must not place restrictions on other software that is distributed along with the licensed software. For example, the license must not insist that all other programs distributed on the same medium must be open-source software.

10. License Must Be Technology-Neutral

No provision of the license may be predicated on any individual technology or style of interface.