OPEN SOURCE AI DEFINITION

Online public townhall

July 26, 2024

last updated: July 26, 2024 (NV)
Community agreements

- **One Mic, One Speaker** -- Please allow one person to speak at a time.
- **Take Space, Make Space** -- If you tend to talk more, we invite you to make space for others to share. If you tend not to share, we invite you to speak up.
- **Kindness** -- This work is hard, but we don’t have to be. Gentleness and curiosity help. Those who use insults or hate speech will need to leave the meeting.
- **Forward Motion** -- We advance by focusing on what is possible in the moment and doing it. Obstacles are marked for later discussion, not used to stop the process. If we hit a boulder, we note it on the map and keep walking. We’ll come back and unearth it later on.
- **Solution-Seeking** -- This work is so complex that focusing on what won’t work will stop it. Suggesting new ideas, options, and proposals is vulnerable, but crucial. All of us are needed to make this work.
- **Anything else?**
OSI’s objective for 2024
Open Source AI Definition
Open Source AI Definition

Current Version

OSAID v.0.0.8
Open Source AI Definition
What We’re Working On
OSAID v.0.0.9
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 summarized as autonomy, transparency, frictionless reuse, and collaborative improvement.

Clarifying that the recipients of the freedoms are developers, deployers and end-users.
Open Source AI Definition

Four Freedoms v.0.0.9 plans

What is Open Source AI

An Open Source AI is an AI system made available under 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 for any purpose, including to change its output.
- **Share** the system for others to use with or without modifications, for any purpose.

Clarifying that the four freedoms of open source AI are derived from the Free Software Definition.
Underlining that components and systems must be free from encumbrances that prevent any developer, deployer, or users from exercising those freedoms.
Open Source AI Definition
Preferred Form
v.0.0.9 plans

Preamble

Why we need Open Source Artificial Intelligence (AI)

Open Source has demonstrated that inclusive benefits accrue to everyone when others freely share, build upon, and improve upon another’s work. These benefits are realized only because that work is shared as the Open Source Definition. The benefits can be summed as increased transparency, technical merit, and information and competition.

Everyone needs these benefits. As AI, we need a specific framework to enable users to build, develop, and deploy systems that are transparent and interoperable.

What is Open Source AI

As Open Source is an open-source suitcase, under terms that grant the beneficiaries

• The right to use any software without making any use for permission.
• The right to inspect software and request the components.
• The right to inspect, copy, and modify the components.
• The right to share software with others, including in modified forms.
• The right to use the system for one or without modifications, for any purpose.

Preferential form to make modifications to machine-learning systems

The add definitions of...

...the “OSD compliant” requirement for data information...

...and the “OSD conformant” requirement for model parameters

..so legal requirements are clear for each component

Checklist to evaluate machine learning systems

<table>
<thead>
<tr>
<th>Table of default required components</th>
<th>Legal Frameworks</th>
</tr>
</thead>
<tbody>
<tr>
<td>-- component required</td>
<td>-- legal framework</td>
</tr>
<tr>
<td>Training data and annotations</td>
<td>Available under OSD- compliant licenses</td>
</tr>
<tr>
<td>Testing data and annotation</td>
<td>Available under OSD- compliant licenses</td>
</tr>
<tr>
<td>Training data (including full data, annotated)</td>
<td>Available under OSD- compliant licenses</td>
</tr>
<tr>
<td>Testing data (with labels, test set)</td>
<td>Available under OSD- compliant licenses</td>
</tr>
<tr>
<td>Training data cleaning strategy</td>
<td>Available under OSD- compliant licenses</td>
</tr>
<tr>
<td>Code</td>
<td>-- legal framework</td>
</tr>
<tr>
<td>-- before processing</td>
<td>Available under OSD- approved license</td>
</tr>
<tr>
<td>-- after modeling</td>
<td>Available under OSD- approved license</td>
</tr>
<tr>
<td>License</td>
<td>Available under OSD- approved license</td>
</tr>
<tr>
<td>Model architecture</td>
<td>Available under OSD- approved license</td>
</tr>
<tr>
<td>Model parameters</td>
<td>Available under OSD- compliant licenses</td>
</tr>
</tbody>
</table>

<...>
Checklist will be a separate document and process and its components will be updated to follow the Model Openness Framework (MOF) precisely.
Open Source AI Definition

**System Validation**

OSAID v.0.0.8 (and soon v. 0.0.9)
We were interested in reviewing about 10 AI systems self-described as open to validate the definition.

1. Arctic
   1. Jesús M. Gonzalez-Barahona
      Universidad Rey Juan Carlos

2. BLOOM
   2. Danish Contractor
      BLOOM Model Gov. Work Group
   3. Jaan Li
      University of Tartu, One Fact Foundation

3. Falcon
   1. Casey Valk
      Nutanix
   2. Jean-Pierre Lorre
      LINAGORA, OpenLLM-France

4. Grok
   1. Victor Lu
      independent database consultant
   2. Karsten Wade
      Open Community Architects

5. Llama 2
   1. Davide Testuggine
      Meta
   2. Jonathan Torres
      Meta
   3. Stefano Zacchioli
      Polytechnic Institute of Paris
   4. Victor Lu
      independent database consultant

6. LLM360
   5. [Team member TBD]
      LLM360

7. OLMo
   4. Amanda Casari
      Google
   5. Abdoulaye Diack
      Google

8. OpenCV*
   1. Rasim Sen
      Oasis Software Technology Ltd.

9. Phi-2
   6. Seo-Young Isabelle Hwang
      Samsung

10. Pythia*
    1. Seo-Young Isabelle Hwang
        Samsung
    2. Stella Biderman
        EleutherAI
    3. Hailey Schoelkopf
        EleutherAI
    4. Aviya Skowron
        EleutherAI

11. T5
    5. Jaan Li
       University of Tartu, One Fact Foundation

Viking

6. Merlijn Sebrechts
   Ghent University
# Validation Updates

Thanks to Arctic and LLM360 for helping identify documentation!

<table>
<thead>
<tr>
<th>AI System</th>
<th>Meets OSAID requirements?</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of system with link to its review sheet</td>
<td>Based on OSAID v. 0.0.8 and/or v.0.0.6</td>
<td>Summary explanation of status (as of 6/11/24)</td>
</tr>
<tr>
<td>Arctic</td>
<td>Expect Yes</td>
<td>Verbal confirmation from Snowflake, which is adding legal documents to review sheet (6/3/24)</td>
</tr>
<tr>
<td>BLOOM</td>
<td>Confirmed No (license fails)</td>
<td>Usage restrictions in RAIL license</td>
</tr>
<tr>
<td>Falcon</td>
<td>Expect No</td>
<td>Documents on training methodologies and techniques and training, validation and testing are missing</td>
</tr>
<tr>
<td>Grok</td>
<td>Expect No</td>
<td>Very little public information on system</td>
</tr>
<tr>
<td>Llama 2</td>
<td>Confirmed No</td>
<td>Data pre-processing + training, validation and testing code are not available</td>
</tr>
<tr>
<td>LLM360</td>
<td>Expect Yes</td>
<td>Self-certified as compliant on the forum, awaiting addition of reviewable documents to their sheet</td>
</tr>
<tr>
<td>Mistral</td>
<td>Confirmed No</td>
<td>Some data information and code components missing, no training code available</td>
</tr>
<tr>
<td>OLMo</td>
<td>Expect Yes</td>
<td>Supporting libraries and tools unclear, but all other legal documentation is present</td>
</tr>
<tr>
<td>OpenCV</td>
<td>Unclear</td>
<td>Model requirement unclear because OpenCV does not store, but instead supports external deep learning frameworks</td>
</tr>
<tr>
<td>Phi-2</td>
<td>Unclear</td>
<td>Data information, code, and model information missing</td>
</tr>
<tr>
<td>Poro</td>
<td>Unclear</td>
<td>Most review documentation not yet located; Located documentation meets OSAID requirements</td>
</tr>
<tr>
<td>Pythia</td>
<td>Confirmed Yes</td>
<td>Only non-alignment was absence of labeling documentation, which was not created. v 0.0.8 adds &quot;if used&quot; to requirement, resolving this</td>
</tr>
<tr>
<td>T5</td>
<td>Expect Yes</td>
<td>Only possible restriction is in supporting libraries and tools because gcloud command requires special hardware. Hardware requirements are out of scope for the OSAID, so this is likely not a recognized restriction.</td>
</tr>
</tbody>
</table>
Open Source AI Definition

What’s Next?

June - October 2024

- Complete validation phase
- Resolve comments, release v. 0.0.9 after validation
- Cut the release candidate with sufficient endorsement
The OSI Board requires a definition that is:

- Supported by diverse stakeholders
- Provides real-life examples
- Ready by October 2024

The definition needs to have approval by end users, developers, deployers and subjects of AI, globally.

The definition must include relevant examples of AI systems that comply with it at the time of approval, so cannot have an empty set.

A usable version of the definition needs to be ready for approval by the board at the October board meeting.

Approved June 21, 2024
2024 Timeline

- **February**: Call For Volunteers + Activity Feedback and Revision
- **June**: Virtual System Review
- **July**: Virtual System Review
- **August**: Virtual System Review
- **September**: Virtual System Review Ends
- **October**: Townhalls +
  - All Things Open (Raleigh)
  - Data Workshop (Europe TBD)

- **June**: Townhalls +
  - OSPOs for Good (NYC)
  - Sustain Africa (virtual)

- **July**: Townhalls +
  - AI-dev (Hong Kong)

- **August**: Townhalls +
  - Nerdearla (Buenos Aires)

- **February**: Bi-Weekly Virtual Public Townhalls
- **June**: Bi-Weekly Virtual Public Townhalls

- **June**: Draft 0.0.5
- **July**: Draft 0.0.8
- **August**: Draft 0.0.9
- **September**: RC1
- **October**: RC1

**System testing work stream**

- Stakeholder consultation work stream

**Release schedule**

- Draft 0.0.5
- Draft 0.0.8
- Draft 0.0.9
- RC1
- RC1
- Stable Version
<table>
<thead>
<tr>
<th>Region</th>
<th>Country</th>
<th>City</th>
<th>Conference</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>North America</td>
<td>United States</td>
<td>Pittsburgh</td>
<td>✓ PyCon US</td>
<td>May 17</td>
</tr>
<tr>
<td>Europe</td>
<td>France</td>
<td>Paris</td>
<td>✓ OW2</td>
<td>June 11 - 12</td>
</tr>
<tr>
<td>North America</td>
<td>United States</td>
<td>New York</td>
<td>✓ OSPOs for Good</td>
<td>July 9 - 11</td>
</tr>
<tr>
<td>Africa</td>
<td>Virtual</td>
<td>Virtual</td>
<td>✓ Sustain Africa</td>
<td>July 15</td>
</tr>
<tr>
<td>Asia Pacific</td>
<td>China</td>
<td>Hong Kong</td>
<td>AI_dev</td>
<td>August 23</td>
</tr>
<tr>
<td>Latin America</td>
<td>Argentina</td>
<td>Buenos Aires</td>
<td>Nerdearlᵃ</td>
<td>September 24 - 28</td>
</tr>
<tr>
<td>Europe</td>
<td>France</td>
<td>Paris</td>
<td>Data governance</td>
<td>October</td>
</tr>
<tr>
<td>North America</td>
<td>United States</td>
<td>Raleigh</td>
<td>All Things Open</td>
<td>Oct 27 - 29</td>
</tr>
</tbody>
</table>
Co-Design Workshop in Monterey, USA

Linux Foundation Member Summit | October 2023

Share

Share the system, with or without modifications, for any purpose, [without limitations].

Study

Study how the AI system works, and inspect its components. Access to the AI system components is a precondition of this.
Co-Design Workshop in Addis Ababa, Ethiopia
How to Participate :)

● Public forum: discuss.opensource.org

● Become an OSI member
  ○ Free or or full
  ○ SSO with other OSI websites

● Biweekly virtual townhalls… like this one!

● Volunteer to help with validation (email or DM Mer Joyce)
Thank you

We realize this is difficult work and we appreciate your help and openness in improving the definition.