OPEN SOURCE AI DEFINITION

Online public townhall

April 5, 2024

last updated: March 21, 2024 (SM)

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?**

The objective for 2024

Open Source AI Definition version 1.0

hackmd.io/@opensourceinitiative/osaid-0-0-5

Definition of AI system

Preamble

Out of scope issues

4 freedoms

Legal checklist

version 0.0.3

Leave comments for this text

About Programs Licenses Open Source

```
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

Definition of AI system

Preamble

Out of scope issues

4 freedoms

Legal terms checklist

version 0.0.3

Leave comments for this text

About Programs Licenses Open Source

```
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

Done … ish?

Revising draft

Open Source AI Definition v. 0.0.6

An Open Source AI is an AI system made available to the public 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.

Precondition to exercise these freedoms is to have access to the preferred form to make modifications to the system. For machine learning systems that means having public access to:

- **Data**: Sufficiently detailed information on how the system was trained, including the training methodologies and techniques, the training data sets used, information about the provenance of those data sets, their scope and characteristics; how the data was obtained and selected, the labeling procedures and data cleaning methodologies.
- **Code**: The code used for pre-processing data, the code used for training, validation and testing, the supporting libraries like tokenizers and hyperparameters search code (if used), the inference code, and the model architecture.
- **Model**: The model parameters, including weights. Where applicable, these should include checkpoints from key intermediate stages of training as well as the final optimizer state.

transparency requirements only

System Review Workgroups

- Creating content for definition v. 0.0.7
- Release by next Friday, April 12th

Workgroups

Selected for diversity of approaches to AI openness:

1. **Pythia**: open science project, with a permissive license
2. **BLOOM**: open science project, with lots of details released but shared with a restrictive license
3. **Llama 2**: commercial project, accompanied by limited amount of science and with a restrictive license
4. **OpenCV**: open source project, with ML components outside of the generative AI space

Members

Llama 2

1. **Bastien Guerry** DINUM, French public administration
2. **Ezequiel Lanza** Intel
3. **Roman Shaposhnik** Apache Software Foundation
4. **Davide Testuggine** Meta
5. **Jonathan Torres** Meta
6. **Stefano Zacchiroli** Polytechnic Institute of Paris
7. **Mo Zhou** Debian, Johns Hopkins University
8. **Victor Lu** independent database consultant

BLOOM

1. **George C. G. Barbosa** Fundação Oswaldo Cruz
2. **Daniel Brumund** GIZ FAIR Forward - AI for all
3. **Danish Contractor** BLOOM Model Gov. WG
4. **Abdoulaye Diack** Google
5. **Jaan Li** University of Tartu, Phare Health
6. **Jean-Pierre Lorre** LINAGORA, OpenLLM-France
7. **Ofentse Phuti** WiMLDS Gaborone
8. **Caleb Fianku Quao** Kwame Nkrumah University of Science and Technology, Kumasi

Pythia

1. **Seo-Young Isabelle Hwang** Samsung
2. **Cailean Osborne** University of Oxford, Linux Foundation
3. **Stella Biderman** EleutherAI
4. **Justin Colannino** Microsoft
5. **Hailey Schoelkopf** EleutherAI
6. **Aviya Skowron** EleutherAI

To achieve better global representation, we conducted outreach to Black, Indigenous, and other People of Color, particularly women and individuals from the Global South.

OpenCV

1. **Rahmat Akintola** Cubeseed Africa
2. **Ignatius Ezeani** Lancaster University
3. **Kevin Harerimana** CMU Africa
4. **Satya Mallick** OpenCV
5. **David Manset** ITU
6. **Phil Nelson** OpenCV
7. **Tlamelo Makati** WiMLDS Gaborone, Technological University Dublin
8. **Minyechil Alehegn Tefera** Mizan Tepi University
9. **Akosua Twumasi** Ghana Health Service

Phase 1: Deciding Required Components

Component Voting

Code All code used to parse and process data, including:	Required to Use?	Required to Study?	Required to Modify?	Required to Share?
Data preprocessing code		SZ	SZ EL	
Training code		SZ	SZ	
Test code				
Code used to perform inference for benchmark tests				
Validation code				
Inference code	SM EL DT SM JT SZ			
Evaluation code				
Other libraries or code artifacts that are part of the system, such as tokenizers and hyperparameter search code, if used.	BG,EL, SM, SZ			

example: Llama 2

January 15

Vote Compilation

Recommendation Report

March 10

Definition v. 0.0.6

OSI: Open Source AI Definition

File Edit View Insert Format Data Tools Extensions Help

Likely required for all four freedoms

Components of an AI system	Recommendation Should it be required?	Rationale Why should it be required?	Total All Votes	Votes (MOF update)
Code	last update: 2/21/24 (MJ)	last update: 2/21/24 (MJ)		
■* Data preprocessing code	Lean yes	Likely required to study and modify	13	
■ Training, validation and testing code	Yes	Likely required to study and modify	21	
■ Inference code	Yes	Likely required to use, possibly to study and modify	23	
■ Evaluation code	Lean no	Likely not required to study	3	
Data				
■ Datasets	Maybe	Requirement to study offset by lack of necessity for use	8	
▸ Training datasets	Lean no	Possibly required for study	4	
▸ Testing datasets	Lean no	Possibly required for study	2	
▸ Validation datasets	No	Likely not required for study	0	
▸ Benchmarking datasets	Lean no	Possibly required for study	2	
■ Data card	No	Likely not required for study	-1	
■ Evaluation Data	Lean no	Likely not required for study	3	
■ Evaluation Results	Lean no	Likely not required for study	4	
All other data documentation	Lean no	Possibly required for study	4	
Model				

Report on working group recommendations

Open Source AI process

Recommendations

The recommendations below respond to the question:

- *Should X component be required for an AI system to be licensed as open?*

Based on the number of votes for each component across all
follows:

Required

- Training, validation, and testing code
- Inference code
- Model architecture
- Model parameters
- Supporting libraries & tools*

Likely Required

- Data preprocessing code

Maybe Required

- Training datasets
- Testing datasets
- Usage documentation
- Research paper

The Open Source AI Definition

Checklist to evaluate legal documents

This table is work in progress. See slide 7 of Jan 26 town hall for more details.

Required components	Legal frameworks
Code	
- Data pre-processing	Available under OSI-compliant license
- Training, validation and testing	Available under OSI-compliant license
- Inference code	Available under OSI-compliant license
- Supporting libraries and tools	Available under OSI-compliant license
Model	
- Model architecture	Available under OSI-compliant license
- Model parameters (including weights)	To be defined in the next phase

The following components are not required, but their inclusion in public releases is appreciated.

Optional components

- Code used to perform inference for benchmark tests

Process: From mid-January through February, system workgroups voted on which components should be required for a system to be defined as open. These votes were then publicly tabulated and a recommendations report was publicly shared on the forum. The recommendations became version 0.0.6 of the definition.

Phase 2: Finetuning the Component Checklist

Definition v. 0.0.6

The Open Source AI Definition

Checklist to evaluate legal documents

This table is work in progress. See slide 7 of Jan 26 town hall for more details.

Required components
Code
- Data pre-processing
- Training, validation and testing
- Inference code
- Supporting libraries and tools
Model
- Model architecture
- Model parameters (including weights)

The following components are not required appreciated.

Optional components
- Code used to perform inference for ben

Checklist for Doc Review

Required Components
Code
Data pre-processing
Training, validation and testing
Inference
Supporting libraries and tools (including tokenizers and hyperparameters search code, if used)
Model
Model architecture
Model parameters (including weights)
Documentation
Training methodologies and techniques
Training data scope and characteristics
Training data provenance (including how data was obtained and selected)
Training data labeling procedures
Training data cleaning methodology

Definition to Checklist: A week after the v.0.0.6 release in mid-March, we went back the the workgroup members to ask them to help us finetune the requirements checklist implied by version 0.0.6 (left).

This v 0.0.6 checklist includes the components categorized as required or likely required by voting in Phase 1, plus a list of data transparency requirements already in force under the EU's AI Act.

Requirements: Open Source AI Definition 0.0.6

Document Review Spreadsheet

Required Components	Legal Framework			BLOOM Analysis		terms: OSAID 0.0.6
source: Open Source AI Definition v. 0.0.6	for each required component	Links to Legal Document	**Use** for any purpose and without having to ask for permission	**Study** how the system works and inspect its components	**Modification** for any purpose, including to change its output	**Sharing** for others to use, with or without modifications, for any purpose
Code						
Data pre-processing	Available under OSI-compliant license	[add link]				
Training, validation and testing	Available under OSI-compliant license	[add link]				
Inference	Available under OSI-compliant license	[add link]				
Supporting libraries and tools (including tokenizers and hyperparameters search code, if used)	Available under OSI-compliant license	[add link]				
Model						
Model architecture	Available under OSI-compliant license	[add link]				
Model parameters (including weights)	TBD	[add link]				
Documentation						
Training methodologies and techniques	?	[add link]				
Training data scope and characteristics	?	[add link]				
Training data provenance (including how data was obtained and selected)	?	[add link]				
Training data labeling procedures	?	[add link]				
Training data cleaning methodology	?	[add link]				

example; BLOOM review spreadsheet

To be added in version 0.0.7

Document Reviewers

Llama 2

Affiliated

1. **Davide Testuggine** Meta
2. **Jonathan Torres** Meta

Unaffiliated

3. **Stefano Zacchiroli** Polytechnic Institute of Paris
4. **Victor Lu** independent database consultant

BLOOM

Affiliated

1. **Danish Contractor** BLOOM Model Governance Workgroup

Unaffiliated

2. **Jaan Li** University of Tartu, Phare Health

Pythia

Affiliated

1. **Stella Biderman** EleutherAI
2. **Aviya Skowron** EleutherAI
3. **Hailey Schoelkopf** EleutherAI

Unaffiliated

4. **Seo-Young Isabelle Hwang** Samsung

OpenCV

Affiliated

1. ***none***

Unaffiliated

2. ***none***

Volunteers needed! Email or DM Mer

Representation: Relation to Open Source AI

Stakeholder	Description	Example
1. **System Creator**	Makes AI system and/or component that will be studied, used, modified, or shared through an open source license	ML researcher in academia or industry
2. **License Creator**	Writes or edits the open source license to be applied to the AI system or component, includes compliance	IP lawyer
3. **Regulator**	Writes or edits rules governing licenses and systems	government policy-maker
4. **Licensee**	Seeks to study, use modify, or share an open source AI system	AI engineer in industry, health researcher in academia
5. **End User**	Consumes a system output, but does not seek to study, use, modify, or share the system	student using a chatbot to write a report, artist creating an image
6. **Subject**	Affected upstream or downstream by a system output without interacting with it intentionally + advocates for this group.	photographer who finds their image in training dataset (upstream), mortgage applicant evaluated by a bank's AI system (downstream)

Representation: Global Inclusion and Equity

open source initiative®

Seeking document reviewers for Pythia and OpenCV

Open Source AI process

Mer 5d

TASK: As part of the systems review track, we're looking for volunteers to review licenses for the Pythia and OpenCV systems and fill out this spreadsheet 4 to check the compatibility of version 0.0.6 3 of our definition with current AI systems.

TIMELINE: Our goal was to complete this review by next Tuesday, April 2nd, though we'll likely extend the deadline in consultation with the volunteers who respond.

VOLUNTEERS: For transparency, reviewers will have their names and affiliations made public. Black, Indigenous, Latine, and other people of color, women, queer, transgender, and non-binary people, people with disabilities, and people from poor and working class backgrounds are encouraged to respond.

LEARN MORE Reviewers are already assigned in the Llama 2 and BLOOM groups. We have two reviewers for Pythia and are seeking more. We have no reviewers yet for OpenCV. Further information on the workgroups and their past activities can be found here 2.

Next Steps

2024 timeline

System testing work stream
Stakeholder consultation work stream
Release schedule

OSAID v. 0.0.7 by next Friday, April 12th

February	March	April	May	June ...	... October
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**
Bi-Weekly Virtual Public Townhalls	Bi-Weekly Virtual Public Townhalls	Bi-Weekly Virtual Public Townhalls	Townhalls + **PyCon Workshop** (≈ May 17th, Pittsburgh)	Townhall + **OSI In-Person Stakeholder Meeting** (date + place TBD)	**Release version 1.0**
Draft 0.0.5	Draft 0.0.6	Draft 0.0.7	Draft 0.0.8	RC1	v. 1.0

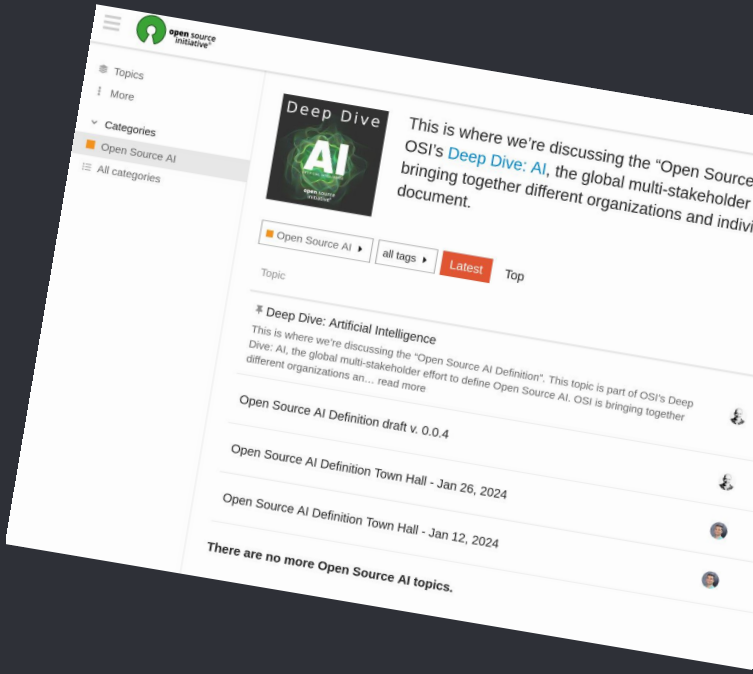
Deep Dive AI in-person meetings

Region	Country	City	Conference	Date
North America	United States	Pittsburgh	PyCon US	May 17*
Europe	?			May ?
Africa	Nigeria	Abuja	OSCA	June 6 - 8
Latin America	Mexico	Mexico D.F.	Latam OSS	July 19 - 20
Asia Pacific	Hong Kong	Hong Kong	AI_dev	August 23
North America	United States	Raleigh	All Things Open	Oct 27 - 29

*confirmed

Join the conversation

- discuss.opensource.org
- Public forum
- Join as OSI member
  - Free or full
  - SSO with other OSI websites



Q & A

Thank you

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

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

Help us find stakeholders

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

Finetuning with Document Review

AI systems	List of components	Legal frameworks	Legal documents	Finetuned **Checklist**
Active workgroups: - Llama2 - Pythia - BLOOM Recruiting reviewers - OpenCV	What elements are necessary to: - use - study - modify - share an AI system? These are listed in definition v. 0.0.6.	For each component, evaluate which laws apply. Some will be under “Intellectual Property” regimes, some will be under other regimes.	Next, match the components and legal frame-works with the terms of the legal documents, if they exist.	After repeating this exercise through multiple systems, we’ll be able to generalize the outcomes and write the specs to evaluate the freedoms granted. These will appear in definition 0.0.7.