LFAI Trusted AI Committee

Principles Working Group

Proposed Approach to Workstream

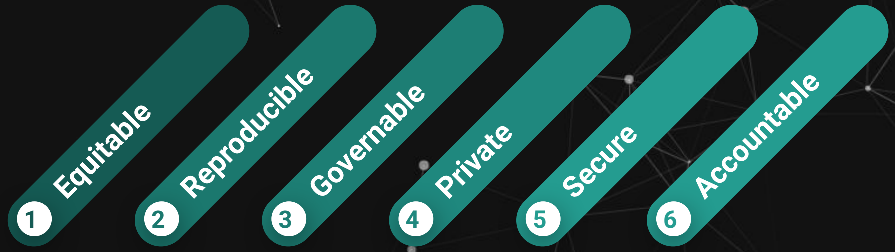
Alejandro Saucedo | a@ethical.institute

@AxSaucedo

Executive Summary

LF AI Trusted AI Principles:

- Identified 6 principles, which were abstracted from existing initiatives, and have been structured to be relevant for any open source projects
- Agreed that principles will focus “beyond the algorithms” into the OSS governance process



1 Equitable
2 Reproducible
3 Governable
4 Private
5 Secure
6 Accountable

Document with full details on each principle:
https://docs.google.com/document/d/1t03G7JOc1KpzzX7KyZq0YHsg2yCZaTqV4OZodVqHsjM/edit

Next steps:

- Align with broader LF AI Trusted AI Committee with the proposed LF AI Trusted AI Principles
- Identify potential areas for application of LF AI Trusted AI Principles such as:
    - Case studies of alignment of existing projects with the Principles
    - Incubating / graduation project assessment with the Principles
- Create checklist that would allow parties to evaluate OSS projects with the LF AI Trusted AI Principles

Structure of Principles

Objective to identify 4-8 high level principles

Each of the principles identified would have the following:

1. High level overview / description
2. Evaluation criteria “Maturity Model” (defined in next slide)
3. Examples of how current OSS projects have addressed this

Implementation Level of Principles

These principles would be implemented in the following situations:

- Assessing the feasibility of incubation and/or graduation of an LFAI Project
- Re-assessing the current state of processes in regards to maintenance and governance of an existing incubating and/or graduated LFAI Project

Focus beyond the algorithms

Concerns exist in regards to the principles becoming a barrier to entry if used during the project evaluation

- Principles could be focused on the open source development and maintenance process instead of the features of the underlying project
- Principles could also be positioned in a more generic way, instead of being used during the evaluation process of the open source projects

Evaluation Criteria: Maturity Model

The evaluation criteria provides an objective scorecard:

1. Given it’s hard to agree on “best practice”, it focuses on defining what is objectively “bad practice” for an LFAI project
2. Allows for answers to not just be “pass / fail” but instead descriptions of how it’s addressed, or how it will be addressed
3. This will ensure that projects have a minimum set of processes required in the context of trusted AI

Evaluation Criteria: Example

An example set of evaluation criteria for many non-particularly defined principles can be the following:

- Contributors have not assessed the impact of incorrect performance of core project functionality as well as ways in which it can be addressed in the context of
- Contributors for projects that handle data have not outlined the required considerations when using the open source project to make sure that data is handled correctly
- Contributors do not have a process and/or infrastructure to revert models in production without unreasonable level of disruption
- Contributors don't have process to assess human review process requirements based on the impact of incorrect predictions
- No process and/or infrastructure to ensure machine learning data encrypted on transport/rest

The answers for some of these questions can still be "no" as long as a reasonable answer or a roadmap focus/action is added

Abstracting from Existing Projects

Given there are several very renowned projects in the LFAI, this provides an opportunity to identify existing best practices

This can also provide with practical case studies on how existing projects align with the principles chosen

Why this is important?

Although it’s obvious for our group on why we’re pushing for the principles, it would be key to verbalise why the project

The world runs on open source, the world will run on AI. It’s not only important to ensure the right principles in the use of technology, but also in the core development processes

Proposed Principles

1 Equitable

2 Reproducible

3 Governable

4 Private

5 Secure

6 Accountable

Document with full details on each principle:
https://docs.google.com/document/d/1t03G7JOc1KpzzX7KyZq0YHsg2yCZaTqV4OZodVqHsjM/edit

Next steps

- Get alignment & input from the broader LFAI working group
- Identify the practical way in which this can be applied into the process of open source maintenance / development
- Explore re-using the practical assessment points of https://ec.europa.eu/futurium/en/ai-alliance-consultation
- Explore further phases: 1) define principles, 2) define checklists, 3) identify tools that can automate checks