Ethics in Artificial Intelligence (Module 3)

Last update: 08 March 2025

Academic Year 2024 – 2025
Alma Mater Studiorum · University of Bologna

Contents

1 Human agency and oversight

AI act, article 14 Article related to human oversight. It states that: AI act, article 14

- Human centric AI is one of the key safeguarding principles to prevent risks.
- AI systems must be designed and developed with appropriate interfaces to allow humans to oversee them.

Human agency AI systems should empower human beings such that they can: Human agency

- Make informed decisions.
- Foster their fundamental rights.

This can be achieved with methods like:

- Human-centric approaches,
- AI for social good,
- Human computation,
- Interactive machine learning.

Human oversight Oversight mechanisms to prevent manipulation, deception, conditioning from AI systems. Human oversight

Possible methods are:

- Human-in-the-loop,
- Human-on-the-loop,
- Human-in-command.

Human-centered AI framework Approach centered on high autonomy while keeping human control. Human-centered AI framework

Remark. Human agency and oversight happens at different levels:

Development team Responsible for the technical part.

Organization Decides who is in charge of accountability, validation, ...

External reviewers (e.g., certification entities).

1.1 Governance and methodology

Human-out-of-the-loop The environment is static and cannot integrate human knowledge. The AI system is a black-box that cannot be used in safety-critical settings. Human-out-of-the-loop

Human-in-the-loop (HITL) The environment is dynamic and can use expert knowledge. The AI system is explainable or transparent and suitable for safety-critical settings. Human-in-the-loop (HITL)

In practice, the AI system stops and waits for human commands before making a decision.

Society-in-the-loop The society, with its conflicting interests and values, is taken into account. Society-in-the-loop

Human-on-the-loop (HOTL) The AI system operates autonomously and the human can intervene if needed. Human-on-the-loop (HOTL)

Remark. Limitations of human-centric AI are:

- It does not scale well as human intervention is involved.
- It is hard to evaluate its effectiveness.
- Performance of the AI system might degrade.

1.2 HITL state-of-the-art approaches

1.2.1 Active learning

Active learning The system is in control of the learning process and the human acts as an oracle for labeling data. Active learning

The learner can query, following some strategy, the human for the ground-truth of unlabeled data. A general algorithm works as follows:

1. Split the data into an initial (small) pool of labeled data and a pool with the remaining unlabeled ones.
2. The model selects an example(s) to be labeled by the oracle.
3. The model is trained on the available labeled data.
4. Repeat until a stop condition is met.

The selection strategy can be:

Random

Uncertainty-based Select examples classified with the least confidence according to some metric.

Diversity-based Select examples that are rare or representative according to some metric.

Remark. This approach is effective in settings with lots of unlabeled data and annotating all of it is expensive.

Remark. This approach is sensitive to the choice of the oracle.

1.2.2 Interactive machine learning

Interactive machine learning Users interactively supply information that influences the learning process. Interactive machine learning

Remark. Compared to active learning, with interactive machine learning it is the human that selects the learning data.

1.2.3 Machine teaching

Machine teaching Human experts are completely in control of the learning process. There can be different types of teachers: Machine teaching

Omniscient teacher Complete access to the components of the learner (i.e., feature space, parameters, loss, optimization algorithm, ...).

Surrogate teacher Access to the loss.

Imitation teacher The teacher uses a copy of the learner that it can query to create a surrogate model.

Active teacher The teacher queries the learner and evaluates it based on the output.

Adaptive teacher The teacher selects examples based on the current hypothesis of the learner.