Ethics in Artificial Intelligence

Matteo Donati

March 4, 2023

Contents

1 Science Oriented AI

1.1 Ethics Guidelines for Trustworthy AI

The Ethics Guidelines for Trustworthy AI is a document prepared by the High-Level Expert Group on Artificial Intelligence which has been made public on the 8th of April, 2019. According to such document, AI should be:

- lawful, complying with all applicable laws and regulations;
- ethical, ensuring adherence to ethical principles and values;
- robust, both from a technical and social perspective.

These requirements should be met throughout the system's entire life cycle. In particular, the first chapter of the document states that the development, deployment and use of AI systems should:

- adhere to ethical principles in such a way to respect the human autonomy, prevent harm, achieve fairness and explicability.
- pay particular attention to situations involving more vulnerable groups of people and situations that are characterized by asymmetry of power or information;
- bring substantial benefits to individuals and society.

The second chapter of the document ensures that the development, deployment and use of AI systems meets the seven key requirements for trustworthy AI:

1. human agency and oversight;
2. technical robustness and safety;
3. privacy and data governance;
4. transparency;
5. diversity, non-discrimination and fairness;
6. environmental and societal well-being;
7. accountability.

and considers technical and non-technical methods to ensure the implementation of such requirements. Moreover, this chapter also focuses on:

- research and innovation;
- communicating information to stakeholders about the AI system's capabilities and limitations;
- facilitating the traceability of AI systems;

- involving stakeholders throughout the AI system's life cycle;
- being mindful of the tensions between different principles and requirements.

The third chapter of the document focuses on:

- adopting a trustworthy AI assessment list;
- keeping in mind that such an assessment list will never be exhaustive.

Human-centric AI. Commitment to the use of AI in the service of humanity and the common good, with the goal of improving human welfare and freedom. Need to maximize the benefits of AI systems while preventing and minimizing their risks.

Ethics vs. law. Ethics include norms indicating what should be done, with regard to all interests at stake. Lw include norms that are adopted through institutional processes, and that are coercively enforced. In general, AI should be lawful, meaning that it should comply with EU primary law, secondary law, UN Human Rights treaties and the Council of Europe conventions, laws of EU Member State (i.e. Italian law).

1.1.1 Ethical Principles (Based on Human Rights)

There are four ethical principles that have to be respected:

- Respect for human autonomy. Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process.
- Prevention of harm. AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings.
- Fairness. This principle is based on two dimensions:
  - Substantive dimension, which is about: ensuring equal distribution of both benefits and costs; ensuring that individuals and groups are free from unfair bias and discrimination; promoting equal opportunity in terms of access to education, goods, services and technology; never leading to people being deceived or impaired in their freedom of choice; AI practitioners respecting the principle of proportionality between means and ends.
  - Procedural dimension, which is about the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them.
- Explicability. This principle is about ensuring contestability. Namely: processes need to be transparent; capabilities and purpose of AI systems have to be communicated; decisions have to be explainable to whomever is affected. Moreover, it should be noticed that an explanation as to why a model has generated a particular output or decision is not always possible.

Possible tensions between principles have to be considered.

1.1.2 Requirements of Trustworthy AI

There are seven requirements for a trustworthy AI:

- Human agency and oversight. AI systems should support human autonomy and decision-making. Thus, they should support: fundamental rights; human agency; human oversight; technical robustness and safety; resilience to attack and security; fallback plan and general safety; accuracy; reliability and reproducibility.
- Technical robustness and safety. AI systems should be developed with a preventive approach to risks and in such a manner that they reliably behave as intended while minimizing unintentional and unexpected harm, and preventing unacceptable harm.
- Privacy and data governance. Prevention of harm necessitates privacy and data governance: privacy and data protection; quality and integrity of data; access to data.
- Transparency. This requirement is bout traceability, explainability, and communication.
- Diversity, non-discrimination and fairness. Inclusion and diversity should be enables throughout the entire AI system's life cycle.
- Societal and environmental well-being. The broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle.
- Accountability. This requirement is about ensuring responsibility and accountability for AI systems and their outcomes.