Ethics in Artificial Intelligence

Matteo Donati

March 7, 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 to help assess AI systems and to achieve requirements. This can be done by disseminating results to the open public, and by systematically training a new generation of experts in ethics in AI;

- communicating information to stakeholders about the AI system's capabilities and limitations. This enables realistic expectation settings and transparency;
- facilitating the traceability of AI systems;
- involving stakeholders throughout the AI system's life cycle. This implies foster training and educating so that all stakeholders are aware of trustworthy AI;
- being mindful of the tensions between different principles and requirements. Continuously identify, evaluate, document and communicate these trade-offs and their solutions.

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. Ensuring trustworthy AI is not about ticking boxes, but about continuously identifying and implementing requirements, evaluating solutions, ensuring improved outcomes throughout the AI system's lifecycle, while involving stakeholders during the process.

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)

According to the aforementioned document, 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. In particular:
  - AI systems should not subordinate, coerce, deceive, manipulate, condition or herd humans (e.g. face recognition to identify people in public spaces);
  - AI systems should be designed to augment, complement and empower human cognitive, social and cultural skills;
  - the allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity to human choice. This also implies human oversight.

- **Prevention of harm**. AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. Moreover, AI systems should be safe and secure.

- **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. In particular, specific methods of accountable deliberation, used to deal with such tensions, should be established.

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. Namely, they should support:

  - **fundamental rights** of humans;
  - **human agency**. Users should be able to make informed and autonomous decisions regarding AI systems;
  - **human oversight**. This helps ensuring that an AI system does not undermine human autonomy nor causes other adverse effects;
  - **technical robustness and safety**;
  - **resilience to attack and security**. AI systems should be protected against vulnerabilities which can be exploited by some other entities;
  - **fallback plan and general safety**. AI systems should have safeguards that enable a fallback plan in case of problems;
  - **accuracy**. AI systems should have the ability of making correct judgments;
  - **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**. This requirement is about:
  - **prevention of harm necessitates privacy and data governance**. AI systems must guarantee privacy and data protection throughout a system's entire lifecycle;
  - **quality and integrity of data**. The data used to train a system should not contain socially constructed bias, inaccuracies, errors and mistakes;
  - **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. In particular, this requirement focuses on:
  - **the avoidance of unfair bias**. This is about prevent unintended prejudice and discrimination against certain groups of people;
  - **accessibility and universal design**. AI systems should be user-centric and designed in such a way to allow people to use AI products or services regardless of their age, abilities or characteristics;
  - **diversity and inclusive design teams**. The teams that design, develop, test and maintain, deploy and procure AI systems should reflect the diversity of users and of society.
- **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. In particular, this requirement focuses on:
  - **sustainable and environmentally friendly AI**. Different measures that help securing the environmental friendliness of AI systems should be encouraged;
  - **social impact**. The effects of these systems on individuals, groups and society must be carefully monitored and considered;
  - **society and democracy**.
- **Accountability**. This requirement is about ensuring responsibility and accountability for AI systems and their outcomes. In particular, this last requirement focuses on:
  - **auditability**. It should be possible to enable the assessment of algorithms, data and design processes;
  - **minimization and reporting of negative impacts**. This is related to the ability to report on actions or decisions that contribute to produce a system's output, and to respond to the consequences of such an outcome;
  - **trade-offs** between requirements;
  - **redress**. Accessible mechanisms that ensure adequate redress should be foreseen.

2 Introduction to Ethics

2.1 Consequentialism

According to the concept of consequentialism:

- An action if morally required:
  - if and only if it delivers the best outcome, relative to its alternative;
  - if and only if its good outcome outweight its negative outcomes to the largest extent;
  - if and only if it produces the highest quality.
- Morality is considered an optimization problem.
- There are various kinds of consequentialism. This is due to how the good and bad things to be optimized are chosen.

The utilitarianism of John Stuart Mill and Jeremy Bentham is a well known example of consequentialism. According to the principle of utility: actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure, and the absence of pain; by unhappiness, pain, and the privation of please. In particular, utilitarianism:

- is conceptually simple;
- is egalitarian (i.e. everybody's utility counts in the same way);
- fits with some basic intuitions;
- in many cases is workable, in other cases is problematic.

There exist two versions of utilitarianism:

- act utilitarianism (i.e. do the action that maximizes utility, do the optifimic action);
- rule utilitarianism (i.e. follow the rule whose consistent application maximized utility, follow the optifimic rule).

2.2 Ethics / Morality

Positive (conventional) morality is defined as the set of moral rules and principles that are accepted in a society. Critical morality, instead, refers to the morality that one believes is correct, rational and just. One can criticize positive morality based on one's critical morality. In particular, morality can be absorbed from the society, thus it can be considered a social phenomenon. In order to make AI learn from society, there is the need to understand how we learn morality from the society. Moreover, morality is considered a place for widespread disagreement (e.g. abortion, migration, capital punishment, etc.), but there always should be something on which everyone may agree.

Normative ethics is concerned with determining what is morally requires, how one ought to behave.

Metaethics is the study of the nature, scope, and meaning of moral judgment.

Absolutism vs. relativism. There is a single true ethics; when two people express incompatible ethical judgment, then one of them must be wrong. Ethical judgments are always relative to particular frameworks or attitudes.

Pro-tanto and all-things-considered moral judgment. Many moral prescription are defeasible. These prescriptions state general propositions that are susceptible of exceptions (e.g. what if a lie would save a person's life?). In general, an act is a prima facie duty when there a moral reason in favor of doing the act, but one that can be outweighted by other moral reasons.