ACAI 2021
The advanced course on AI on Human Centered AI

Tutorial Outline

Mythical Ethical Principles for AI and How to Operationalise Them
Marija Slavkovik, University of Bergen

To have ethical AI two questions need to be answered: what is the ethical impact that an AI system can have and what does it mean for an AI system to behave ethically. The answers to both of these questions hinder the identification of what are the values or principles that we want upheld by AI and for AI. Identifying these principles is not enough, we also want to define them so that they can be operationalised, or at least understand what operationalisation means. Dennett (1988), argued that ethical behaviour cannot be efficiently reached by following moral theories or values, but require guidelines that help a resource bound agent to attain them.

There is not a finite list of which principles or values should characterise an ethical AI system. Operationalization is even further behind. The goal of this tutorial is to make ethical AI more attainable by training AI professionals into asking the right questions: what is an ethical principle for AI and what does it mean for such a principle to be well defined. The tutorial is aimed at AI researchers that are interested in pursuing AI ethics research. There is a gap between moral philosophy and ethically behaving AI. The tutorial outcome is helping AI researchers close this gap by enabling them to interpret an abstract principle from moral philosophy into a property that can be formally specified and measured or computationally implemented.

The tutorial uses recent articles in AI ethics that attempt to define and identify pertinent ethical principles. Based on the specialist publishing venues today, we can distinguish as being of particular interest in fairness (Chouldechova and Roth 2020), (algorithmic) accountability (Wieringa 2020), transparency (Diakopoulos 2020), explainability (Miller 2019), privacy, and responsibility (Dignum 2019). These terms are not mapped to traditionally recognised values in moral philosophy. We discuss the definitions in the literature and argue they strengths and weaknesses from an operational point of view. This analysis will be done as in a workshop, group discussion style together with the participants.

Piano (2020) gives an overview of the literature that attempts to account for the ethical principles considered pertinent for AI. The work of Floridi and Cowls (2019) can be distinguished because they attempt to map a set of 47 principles onto five dimensions: beneficence, non-maleficence, autonomy, justice and, explicability. We discuss how these five dimensions map to the above mentioned goal oriented efforts towards accountability, fairness, transparency, explainability, privacy and responsibility.

References

Chouldechova and Roth (2020) https://cacm.acm.org/magazines/2020/5/244336-a-snapshot-of-the-frontiers-of-fairness-in-machine-learning/fulltext

Dennett (1988) https://tannerlectures.utah.edu/_documents/a-to-z/d/dennett88.pdf

Diakopoulos (2020) https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190067397.001.0001/oxfordhb-9780190067397-e-11

Dignum (2019) https://www.springer.com/gp/book/9783030303709

Floridi and Cowls (2019) https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/7

Miller (2019) https://dl.acm.org/doi/10.1145/3313107

Piano (2020) https://www.nature.com/articles/s41599-020-0501-9

Wieringa (2020) https://dl.acm.org/doi/pdf/10.1145/3351095.3372833