Ethical Implications of AI WS 20/21

Series of Lectures

Frankfurt Big Data Lab, Goethe University Frankfurt

Final Report Requirements

The goal of the final report is to continue to work on the selected AI system (i.e. an AI-product and or an AI-based service) used in healthcare, and finish the evaluation process.

Scope

The report must be delivered using the same google doc. The **new part** must be min 3, max. 5 pages long (excluding references). The **combined references** (mid-term + final) must be **no more than 2 pages**. You should use the same style as for the mid-term report (normal text: Arial, 11pt, 1.15 line spacing, extra space after paragraph, for references you can reduce font size to 9pt).

The FINAL report should cover the following 5 points:

1. Identify any CLAIMS made by the producer of the AI system

According to the definition provided by Brundage et al. (2020, p.65) "**Claims** are assertions put forward for general acceptance. They're typically statements about a property of the system or some subsystem. Claims asserted as true without justification are assumptions, and claims supporting an argument are subclaims." Furthermore "AI developers regularly make claims regarding the properties of AI systems they develop as well as their associated societal consequences. Claims related to AI development might include, e.g.:

- We will adhere to the data usage protocols we have specified;
- The cloud services on which our AI systems run are secure;
- We will evaluate risks and benefits of publishing AI systems in partnership with appropriately qualified third parties;
- We will not create or sell AI systems that are intended to cause harm;
- We will assess and report any harmful societal impacts of AI systems that we build; and
- Broadly, we will act in a way that aligns with society's interests."

(Brundage et al., 2020, p.64)

2. Develop of an evidence base

Following Brundage et al. (2020), this step consists of reviewing and creating an evidence base to verify/support any claims made by producers of the AI system and other relevant stakeholders:

"**Evidence** serves as the basis for justification of a claim. Sources of evidence can include the design, the development process, prior experience, testing, or formal analysis" (Brundage et al., 2020, p.65) and "**Arguments** link evidence to a claim, which can be deterministic, probabilistic, or qualitative. They consist of "statements indicating the general ways of arguing being applied in a particular case and implicitly relied on and whose trustworthiness is well established" [144], together with validation of any scientific laws used. In an engineering context, arguments should be explicit" (Brundage et al. 2020, p.65)

NOTE: You can for example use the "helping hand" from the Claims, Arguments, and Evidence (CAE) framework (Adelard LLP, 2020)

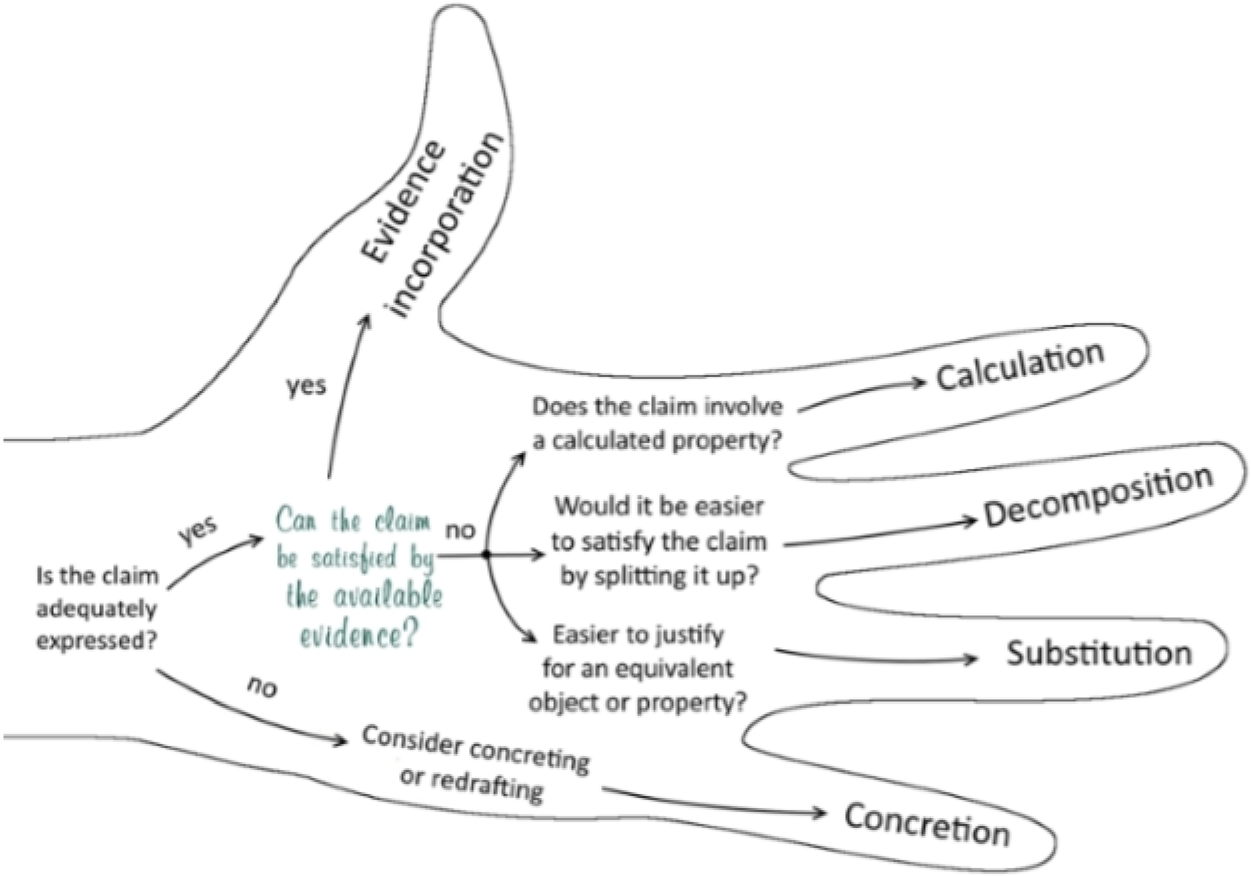
Evidence incorporation
yes
Is the claim adequately expressed?
yes
Can the claim be satisfied by the available evidence?
no
Does the claim involve a calculated property?
Calculation
Would it be easier to satisfy the claim by splitting it up?
Decomposition
Easier to justify for an equivalent object or property?
Substitution
no
Consider concreting or redrafting
Concretion

Figure 1: "Helping Hand" of the CAE framework (Arelard LLP, 2020)

3. Map Ethical issues to Trustworthy AI Areas of Investigation

The basic idea of the process in this step is to identify from the list of ethical issues which areas require inspection. Therefore map Ethical issues to some or all of the seven requirements for trustworthy AI:

- Human agency and oversight,
- Technical robustness and safety,
- Privacy and data governance,
- Transparency,
- Diversity, non-discrimination and fairness,
- Societal and environmental wellbeing
- Accountability

(High-Level Expert Group on Artificial Intelligence, 2019, p.14)

4. Use the **ALTAI web tool** to answer the questions for the corresponding areas of trustworthy AI that you have mapped

5. Critically evaluate the result of the ALTAI assessment if it is relevant for the use case you have chosen

Motivate your analysis.

Grading

Each team receives 0-5 points for the final report.

To pass the course you need to receive at least one point for both the mid-term and the final report and have a total of at least 3 points.

The points will then be translated into a grade (more points = better grade).

References

Adelard LLP (2020). *Helping Hand – CAE FRAMEWORK*. Retrieved December 16, 2020, from https://claimsargumentsevidence.org/notations/helping-hand/

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., … Anderljung, M. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *ArXiv:2004.07213 [Cs]*. http://arxiv.org/abs/2004.07213

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai