Abstract for the SAS 23 Conference: Reliable or Trustworthy AI?

Reliability in Context: Challenges in Operationalizing and Weighing a Central Value

Reliability is widely recognized as a central value in the ethics of AI (cf. eg. Hallensleben et al., 2020). However, adequately operationalizing and weighing this value in concrete AI implementations poses several often-overlooked challenges. As I aim to argue, a number of these challenges have to do with the fact that sensibly operationalizing reliability requires a comprehensive understanding of the concrete and various normative requirements for an AI application. By drawing on concrete examples, I aim to highlight and systematize some of these challenges: 1) On a rather conceptual level, it needs to be to decided which *criteria* are best suited to measure the level of reliability of an AI application. There are a number of candidate criteria (statistical reliability, robustness, resilience; cf. Hallensleben 2022). Yet, as I aim to argue, which one is most relevant in a concrete case depends on an understanding of its relevant normative requirements. 2) As reliability is a goal-relative value (cf. Cartwright 2020), determining the right kind of *indicators* to measure reliability requires an adequate understanding of the goals an AI implementation is meant to pursue. However, for some AI applications (like ChatGPT) identifiying their relevant goals can be a non-trivial task (cf. Alfano et al 2020, Carter 2016). 3) Lastly, I will discuss the challenges of *weighing* the importance of reliability in concrete cases. This assessment involves considering factors such as the impact on non-epistemic values affected by the application and the potential influence of other values like transparency and accountability. The fulfillment of these values may shape the relative urgency assigned to reliability.