In the wake of increasing social embeddedness of Artificial Intelligence (AI) in recent years, and its real-life effects and implications, calls for trustworthiness have gained substantial traction. Whereas a call for trustworthy AI seems more actionable than vague and general demands for some sort of ethical AI, the question remains if trustworthiness is actually applicable to AI and its use-cases.

Understanding AI as a complex socio-technical system might suggest a relational account that goes beyond the engineer's notion of mere reliability, and opens up the possibility of identifying AI itself as a trustworthy entity. In turn, such an understanding could place AI more in line with an inter-agential trust relation, with all the philosophical assumptions that come with it. On the other hand, this notion runs counterintuitive to the understanding of technological artifacts solely as tools. Therefore, we want to address the possibility of trustworthy AI seriously, without reducing it to reliability.

Our contribution investigates the validity of the concept of trustworthy AI: Is it possible to call AI trustworthy without committing a category mistake, equivocation fallacy or undue anthropomorphization? And if so, is trust in AI a good pathway to ethical AI development and use?

We start by analyzing the value of trust through the lens of its constitutive aim: The mitigation of the characteristic kind of risk that arises in the context of interaction with other agents in virtue of their autonomy (agential risk). Then, we describe the notion of trust both formally and substantively with the aim of establishing how well-founded trust can serve this functional role and what being a trustworthy agent amounts to.

We identify a dilemma in the AI context: Trustworthy AI is necessary because AI agents are sources of agential risk which needs to be managed. But at the same time, AI agents cannot fulfill the preconditions necessary to be called trustworthy because they lack the capacity to possess normatively well-grounded motivations for their actions.

We resolve this dilemma by explaining how AI agents' aims can be normatively grounded through their creators and the design process, without simply reducing trust in an AI agent to trust in its creators. We conclude that trustworthy AI is a meaningful concept that can help structure our approach to ethically sound human-AI-interaction.