Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)

What to Trust When We Trust Artificial Intelligence (Extended Abstract)

Duncan Purves, Schuyler Sturm, John Madock

University of Florida
dpurves@ufl.edu, tsturm@ufl.edu, johnmadock@ufl.edu

Abstract

So-called "trustworthy AI" has emerged as a guiding aim of industry leaders, computer and data science researchers, and policy makers in the US and Europe. Often, trustworthy AI is characterized in terms of a list of criteria. These lists include fairness, accountability, and transparency. Fairness, accountability, and transparency are valuable objectives, and they have begun to receive attention from philosophers and legal scholars. However, those who put forth criteria for trustworthy AI have failed to explain why satisfying the criteria makes an AI system—or the organizations that make use of the AI system—worthy of trust. To explain why fairness, accountability, and transparency are suitable criteria for trustworthy AI one needs an analysis of trustworthy AI. An analysis of trustworthy AI provides the fundamental features of an AI system in virtue of which it is (or is not) worthy of trust. This paper puts forward an analysis of trustworthy AI that can be used to critically evaluate criteria for trustworthy AI such as fairness, accountability, and transparency. In this paper we first clarify the target concept to be analyzed: trustworthy AI. We argue that AI, in its current form, should be understood as a distributed, complex system embedded in a larger institutional context. This characterization of AI is consistent with recent definitions proposed by national and international regulatory bodies (e.g., NIST), and it eliminates some unhappy ambiguity in the common usage of the term. We further limit the scope of our discussion to AI systems which are used to inform decision-making about qualification problems, problems wherein a decision-maker must decide whether an individual is qualified for some beneficial or harmful treatment. We argue that, given reasonable assumptions about the nature of trust and trustworthiness, only AI systems that are used to inform decision-making about qualification problems are appropriate candidates for attributions of (un)trustworthiness. We then distinguish between two models of trust and trustworthiness that we find in the existing literature. We motivate our account by highlighting this as a dilemma in in the accounts of trustworthy AI that have previously been offered. These accounts claim that trustworthiness is either exclusive to full agents (and it is thus nonsense when we talk of trustworthy AI), or they offer an account of trustworthiness that collapses into mere reliability. We offer that one of the core challenges of putting forth an account of trustworthy AI is to avoid reducing to one of these two camps. It is thus a desideratum of our account that it avoids being exclusive to full moral agents, while it simultaneously avoids capturing artefacts such as mere tools. We go on to propose our positive account which we submit avoids these twin pitfalls. We subsequently argue that if AI can be trustworthy, then it will be trustworthy on an institutional model. Starting from an account of institutional trust offered by Purves and Davis, we argue that trustworthy AI systems have three features: they are competent with regard to the task they are assigned, they are responsive to the morally salient facts governing the decision-making context in which they are deployed, and they publicly provide evidence of these features. The resulting account allows us to accommodate the core challenge of finding a balance between agential accounts and reliability accounts. We go on to refine our account, answer objections, and revisit the list criteria from above as explained in terms of competence, responsiveness, and evidence.