BudPT21 - Budapest Workshop on Philosophy of Technology 2021
9-10 December 2021
Budapest University of Technology and Economics

Trust and Trustworthiness in AI Ethics

Due to the progress of research in Artificial Intelligence (AI) as well as its deployment and application, the public debate on AI systems has also gained momentum in recent years. With the publication of the *Ethics Guidelines for Trustworthy AI* (2019), notions of trust and trustworthiness gained particular attention within AI ethics: Despite the consensus that AI should be trustworthy, it is less clear what trust and trustworthiness entail in the field of AI, and what ethical standards, technical requirements and practices are needed for the realization of Trustworthy Artificial Intelligence (TAI).

In this paper I will give a detailed overview on the notion of trust employed in AI Ethics Guidelines thus far. Based on this overview I will assess the overlaps and omissions of these guidelines from the perspective of practical philosophy. I will argue that, currently, AI Ethics overloads the notion of trustworthiness. The notion of "trustworthiness" thus runs the risk of becoming a buzzword that cannot be operationalized into a working concept for AI research. On top of that, we can observe that the notion of "trust" deployed in AI research so far is mainly an instrumental notion: Trust is perceived as a condition of a widespread deployment of AI applications. The ambiguities of trust are, however, rarely discussed.

The argumentative aim of this paper is, thus, twofold: I will show that, on the one hand, what is needed with regard to TAI is an approach that is informed with findings on trust from other fields, for instance, social sciences and humanities, especially practical philosophy. On the other hand, practical philosophy will also

gain insights about the phenomenology of trust when taking the debates on TAI into account.

In pursuing my argumentative objectives, I will give an overview on the current state in AI Ethics on trust and trustworthiness via an analysis of current guidelines with regard to these notions. This overview is based on a combined corpus of documents from the respective analyses of Jobin et al. (2019), Zeng et al. (2019), Fjeld et al. (2020), Hagendorff (2020) and Thiebes et al. (2020), amended by guidelines published after the publication of these articles. I will focus on the following questions: To what extent and how is the term currently used in these guidelines? What notions of "trust" and "trustworthiness" are prevalent in these guidelines? What political, social and moral role is attributed to trust? Which concepts are associated with the term trustworthiness? After these observations on the current state of conceptualizing trustworthiness in AI ethics guidelines, I will then discuss overlaps between the guidelines and their omissions: In which areas do these guidelines converge? What has not sufficiently been addressed? I will formulate some points to consider for future research on TAI from the perspective of practical philosophy to close these gaps and avoid a too narrow understanding of trust. Finally, I will also draw some more general conclusions on how we should conceptualize trust with regard to AI – and which mistakes we should avoid.