Through the Sands of Time: A Reliabilistic Account of Justified Credence in the Trustworthiness of AI Systems

Andrea Ferrario[1,2]

[1]ETH, Zurich, Switzerland

[2]Mobiliar Lab for Analytics at ETH, Zurich, Switzerland

Abstract

We address an open problem in the epistemology of artificial intelligence (AI), namely, the justification of the epistemic attitudes we have towards the trustworthiness of AI systems. We start from a key consideration: the trustworthiness of an AI is a time-relative property of the system, with two distinct facets. One is the actual trustworthiness of the AI, and the other is the perceived trustworthiness of the system as assessed by its users while interacting with it. We show that credences, namely, beliefs we hold with a degree of confidence, are the appropriate attitude for capturing the facets of trustworthiness of an AI over time. Then, we introduce a reliabilistic account providing justification to the credence in the trustworthiness of AI, which we derive from Tang's probabilistic theory of justified credence. Our account stipulates that a credence in the trustworthiness of an AI system is justified if and only if it is caused by an assessment process that tends to result in a high proportion of credences for which the actual and perceived trustworthiness of the AI are calibrated. Our approach informs research on human-AI interactions and trustworthy AI by providing actionable recommendations on how to measure the reliability of the process through which users perceive the trustworthiness of the system and its calibration to the actual levels of trustworthiness of the AI. It also allows investigating the relation between reliability and the appropriate reliance on the system.

Keywords

artificial intelligence, trustworthiness, trustworthy AI, epistemology, justification, reliabilism

Extended Abstract

The trustworthiness of AI systems is a key topic in the philosophy of AI [1, 2, 3, 4]. On the one hand, researchers investigate which properties of AI systems make them worthy of users' trust [1, 5, 6, 7]. On the other hand, they discuss how to make users trust the AI systems in an appropriate way, that is, how to entrench trust—whatever that may mean—in the trustworthiness of the systems [8, 3, 9, 10]. Further, trustworthiness is important for trusting from both an epistemological and ethical perspective. Epistemological, as justifying our stance, e.g., a belief, in the trustworthiness of an AI is seen as a necessary step to appropriately relying on the system. In fact, as Durán and Jongsma clearly state in the case of medical AI: "A physician is not morally justified in giving a certain treatment to a patient unless the physician has reliable knowledge

EWAF'23: European Workshop on Algorithmic Fairness, June 7–9, 2023, Winterthur, Switzerland

aferrario@ethz.ch (A. Ferrario)

0000-0001-9968-9474 (A. Ferrario)

that the treatment [suggested by the AI] is likely to benefit the patient" [4, p. 331]. Ethical, as the deployment of trustworthy AI systems mitigate the risk of unethical decisions affecting the lives of many people [5].

Recent studies on the epistemological perspective of trusting AI systems have focused on the grounds on which we can state that our beliefs regarding the trustworthiness of AI predictions are justified [11, 4]. To this end, authors discuss the account of "computational reliabilism" (CR) for AI systems, which states that we are justified in believing that the predictions of an AI system are trustworthy because they are generated by a reliable process, i.e., the AI system itself [12, 11, 4]. The reliability of the AI is described as the tendency to generate predictions that are worthy of our trust [4]. Then, to avoid circularity, a clear definition of trustworthiness is needed. However, the current implementation of CR to AI systems falls short of providing it. As a result, at the time of writing, it is still not clear what we hold when stating that we believe that an AI system—or any of its components—is worthy of trust. Relatedly, the reasons why a belief could be the appropriate propositional attitude to describe our stance towards this "trustworthiness" need further investigation. These research gaps affect the epistemology of AI and the endeavors in human-AI interactions that investigate how to entrench, e.g., "calibrate," trust in the trustworthiness of these systems [4, 3, 9].

Therefore, our goal is to introduce an account of justification of the epistemic attitudes describing our stance towards the trustworthiness of AI systems. Our plan proposes three steps: (1) reviewing related work on the provision of justifications of the beliefs on the trustworthiness of AI, (2) answering the question "what is the trustworthiness of AI?," and (3) arguing that credences[1] are more appropriate than beliefs to describe our stance towards the trustworthiness of these systems [15]. This allows us introducing our account of justified credence in the trustworthiness of AI following Tang's probabilistic theory of reliability [14].

In step (1), we propose to review related work on CR with a focus on its recent implementation for AI systems [4], showing that this account has a problem with trustworthiness. In step (2), we propose to take a *time-relative* perspective on AI systems, which we see as artefacts with a life cycle that alternates between design and deployment phases. Doing so, we can show that, at each step of the AI life cycle and moment of time, the trustworthiness of the AI comprises two, measurable facets. The first—called "actual trustworthiness"—is the objective trustworthiness of the AI that is measured during the design process of the system and is generally unknown after its deployment. The second—called "perceived trustworthiness"—provides a point-in-time perspective on how a deployed system is subjectively perceived as being worthy trust by its different users [16]. It is measured by the users of the system instead. Finally, in step (3), we propose to talk about credences rather than beliefs in the trustworthiness of an AI [15]. In fact, we can show that a probabilistic account of vindicated credences [14] allows encoding the different facets of the trustworthiness of an AI. In particular, their vindication calibrates the degree of confidence of the credence, i.e., the level of perceived trustworthiness, to the actual trustworthiness of the AI. Finally, we can introduce our account of justified credence in the (actual) trustworthiness of AI systems. It is derived from Tang's "(Probability)" reliabilistic theory of justified credence [14] and it states that a credence in the trustworthiness of an AI

[1] A (binary) belief is the attitude we have "whenever we take something to be the case or regard it as true" [13]. A credence is the epistemic attitude towards p that expresses a degree of confidence in p [14, 15].

system is justified if and only if it is the result of a reliable assessment process, i.e., a process to assess the trustworthiness of AI that tends to result in a high proportion of vindicated credences.

The advantages of our approach are manifold. First, it overcomes the limitations affecting the existing account of CR for AI systems by appropriately encoding both facets of trustworthiness in a reliabilist account of justified credence. Then, it allows introducing a measurable definition of calibration between the perceived and actual trustworthiness of an AI. Finally, it allows measuring the reliability of the assessment process of an AI explicitly, generalizing existing reliability scoring of credence-forming processes [17, 18, 19, 20]. This, in turn, suggests the design of empirical studies to investigate the relation between the reliability of the assessment process, the different manipulations of the trustworthiness of the AI over time, the provision of methodologies to support users' assessments—similar to the "reliability indicators" in CR [11, 4]—and the appropriate reliance on the AI's predictions [21].

References

[1] L. Floridi, Establishing the rules for building trustworthy AI, Nature Machine Intelligence 1 (2019) 261–262.

[2] T. Grote, P. Berens, On the ethics of algorithmic decision-making in healthcare, Journal of Medical Ethics 46 (2020) 205–211.

[3] A. Jacovi, A. Marasović, T. Miller, Y. Goldberg, Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021) 624–635.

[4] J. M. Durán, K. R. Jongsma, Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical ai, Journal of Medical Ethics 47 (2021) 329–335.

[5] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399.

[6] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, B. Zhou, Trustworthy AI: From principles to practices, arXiv preprint arXiv:2110.01167 (2021).

[7] E. Petersen, Y. Potdevin, E. Mohammadi, S. Zidowitz, S. Breyer, D. Nowotka, S. Henn, L. Pechmann, M. Leucker, P. Rostalski, et al., Responsible and regulatory conform machine learning for medicine: A survey of challenges and solutions, IEEE Access (2022).

[8] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors 46 (2004) 50–80.

[9] A. Ferrario, M. Loi, How explainability contributes to trust in AI, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1457–1466.

[10] M. Loi, A. Ferrario, E. Viganò, How much do you trust me? A logico-mathematical analysis of the concept of the intensity of trust, Synthese 201 (2023).

[11] J. M. Durán, N. Formanek, Grounds for trust: Essential epistemic opacity and computational reliabilism, Minds and Machines 28 (2018) 645–666.

[12] J. M. Durán, Explaining simulated phenomena: A defense of the epistemic power of computer simulations (2014).

[13] E. Schwitzgebel, Belief, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Winter 2021 ed., Metaphysics Research Lab, Stanford University, 2021.

[14] W. H. Tang, Reliability theories of justified credence, Mind 125 (2016) 63–94.

[15] E. G. Jackson, The relationship between belief and credence, Philosophy Compass 15 (2020) e12668.

[16] Q. V. Liao, S. S. Sundar, Designing for responsible trust in AI systems: A communication perspective, arXiv preprint arXiv:2204.13828 (2022).

[17] B. Lam, Calibrated probabilities and the epistemology of disagreement, Synthese 190 (2013) 1079–1098.

[18] R. Pettigrew, Epistemic utility and norms for credences, Philosophy Compass 8 (2013) 897–908.

[19] J. Dunn, Reliability for degrees of belief, Philosophical Studies 172 (2015) 1929–1952.

[20] R. Pettigrew, What is justified credence?, Episteme 18 (2021) 16–30.

[21] M. Schemmer, P. Hemmer, N. Kühl, C. Benz, G. Satzger, Should I follow AI-based advice? Measuring appropriate reliance in human-ai decision-making, arXiv preprint arXiv:2204.06916 (2022).