## *Caveat Usor:* Trust and Epistemic Vigilance Towards Artificial Intelligence

As artificial intelligence becomes part of our everyday lives, we are increasingly faced with the question of how to use it responsibly. In public discourse, as well as in science and philosophy, this issue is often framed in terms of trust, for example by asking whether, to what extent and under what conditions it is appropriate to trust AI systems. The answers to such questions are quite diverse. On the one hand, there is the ubiquitous rhetoric that "trustworthy AI" should be developed, and that once this is achieved, we can and should trust it.[1] We might call those who espouse this rhetoric "trust enthusiasts". On the other hand, many philosophers have rejected this rhetoric, arguing that trust is a genuinely interpersonal relationship, and that speaking of "trust relationships" with beings other than humans amounts to a kind of category mistake. Instead, we should speak only of "reliance" on AI systems and other technical devices.[2] Let us call those who hold this view "trust sceptics".

In this paper I argue that both positions are partly right and partly wrong. While I agree with the trust sceptics that some forms of trust require the trustee to have a strong normative status or to possess affective states and motivations that current AI systems lack, I argue that trust is a multi-faceted cluster concept and that some forms of trust can also exist toward AI systems. In particular, trust can be conceptualized as an *unquestioning attitude*.[3] When one trusts in this sense, one stops questioning whether the trusted person or object is successfully performing its function.[4]

However, rejecting the position of the trust sceptics does not imply an unqualified endorsement of the trust enthusiasts. Although I agree with the trust enthusiasts that trust serves important socio-epistemic functions and that AI systems should be designed to be as trustworthy as possible, I argue that we should never *blindly* trust an AI system, no matter how well designed it may be. The reason is that trusting is always risky, and the assessment that the trustee is trustworthy can always turn out to be wrong. As a result, epistemic trust should always be accompanied by a sufficient degree of *epis-*

[1] An example is the "Ethics Guidelines for Trustworthy AI" published by the European Commission's High-Level Expert Group on AI in 2019; for a critical analysis of this rhetoric see also Reinhardt, Karoline (2022): Trust and trustworthiness in AI ethics. *AI Ethics*, https://doi.org/10.1007/s43681-022-00200-5.

[2] See, e.g., Goldberg, Sanford C. (2020): Epistemically engineered environments. *Synthese* 197, 2783-2802; Hatherley, Joshua James (2020): Limits of trust in medical AI. *J Med Ethics* 46, 478-481; McMyler, Benjamin (2020): On not making up one's own mind. *Synthese* 197, 2765-2781; Ryan, Mark (2020): In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics* 26, 2749-2767; Kerasidou, Charalampia/Angeliki Kerasidou/Monika Buscher/Stephen Wilkinson (2022): Before and beyond trust: reliance in medical AI. *J Med Ethics* 48, 852-856; Freiman, Ori (2022): Making sense of the conceptual nonsense 'trustworthy AI'. *AI and Ethics*, https://doi.org/10.1007/s43681-022-00241-w.

[3] See Nguyen, C. Thi (2023): Trust as an unquestioning attitude. In: Tamar Gendler et al. (eds.): *Oxford Studies in Epistemology, Vol. 7.* Oxford, 214-244. A similar approach has also been developed by Ferrario, A./Loi, M./Viganò, E. (2020): In AI We Trust Incrementally: A Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. *Philos. Technol.* 33, 523-539.

[4] A version of this definition can also be applied to AI systems that are used as epistemic instruments or informational sources in the following way: to trust an AI system as an epistemic instrument or informational source in domain D is to have an attitude of not questioning the system's deliverances concerning D.

*temic vigilance*.[5] Epistemic vigilance towards an AI system can be characterized as the disposition to (temporarily) suspend one's unquestioning attitude towards the system when certain signs of malfunction or unreliability of the system appear. As a *disposition*, it is not to be confused with an *active search* for such signs (which would be incompatible with trust).[6] Ultimately, my position can be captured by the slogan *caveat usor*: let the user beware (echoing the Roman law principle *caveat emptor*: let the buyer beware).[7]

[5] For a discussion of the concept of epistemic vigilance, see Sperber, Dan/Fabrice Clément/Christophe Heintz/Olivier Mascaro/Hugo Mercier/Gloria Origgi/Deirdre Wilson (2010): Epistemic Vigilance. *Mind and Language* 25, 359-393.

[6] See also Levy, Neil (2022): In Trust We Trust: Epistemic Vigilance and Responsibility. *Social Epistemology* 36, 283-298.

[7] McBrayer recently applied an adaptation of this principle to the context of testimonial exchanges between human speakers and hearers (see McBrayer, Justin P. (2022): *Caveat Auditor*: Epistemic Trust and Conflicts of Interest. *Social Epistemology,* https://doi.org/10.1080/02691728.2022.2078250).