Hybrid Trust: Making Sense of Trustworthy AI

(*abstract: 513 words*)

Interactions between humans and AI based systems are rising considerably in both public and private contexts. Such intelligent and autonomous systems (IATs, henceforth) are increasingly capable of executing complex functions without human supervision and intervention. IATs systems are able to perform human-like actions and we, as human agents (HAs, henceforth), delegate important task and decisions to them. Because of these IATs' capabilities, HAs tend to anthropomorphize them and their perception of IATs is shifting from mere tools to real agents. Indeed, such systems are now widely recognized to be artificial agents (Brezeal et al. 2004, Fossa 2020, Formosa 2021).

Just as delegation of tasks and decisions between HAs requires some form of trust, so does delegation between HAs (Fossa, 2020, Taddeo, 2017). Indeed, many stakeholders are working to make sure that HAs can trust and IATs, and institutions work on legislations and regulations that are being built around the concept of *trustworthy AI*. So, although the notions of trust and trustworthiness concerning IATs are widely used, the philosophical literature on trust and trustworthiness is mainly centered on HAs interactions, and the mainstream attitude is to reject the application of trust and trustworthiness to IATs.

According to the philosophical orthodoxy, the concepts of trust and trustworthiness are not applicable to HAs-IATs interactions: these two notions are morally loaded, and since IATs are not moral agents (contrary to HAs), HAs cannot trust an IAT and IATs cannot be trustworthy. IATs then do not have moral commitments towards HAs, therefore the only attitude a HA can have towards an IAT can be one of reliance and IATs can only bear the property of reliability.

In this paper, I argue that the restriction imposed by the philosophical literature on the application of trust towards non-human agents should be rejected; and that the property of trustworthiness should be somehow extended to IATs. There is a third kind of attitude that is middle way between trust and reliance that can accommodate our intuitions of a different kind of relationship with such IATs, that is what I will call hybrid trust. One way to accommodate the intuition that the kind of attitude that we have towards some IATs is closer to trust than the kind of reliability that we have towards mere inanimate objects is to adopt a novel account of trust: i.e. Tallant's Ought account (2022). On his view,

> x trusts y to f if x believes that y ought to f, and, believes that, where y called upon to f, then y would f

The *ought* can be differently read (morally, normatively, pragmatically, predictively) and at different kinds of *oughts* correspond different kinds of trust. So, there can be cases where non-moral trust is possible: that is where the *ought* is not grounded in moral obligations of the trustee. But there's more: since in the AI community, people are working toward the implementation of moral values into IATs (Townsend et al., 2022) there is a sense in which IATs have also moral obligations towards HAs. Thus, there can be cases of hybrid (moral-plus-non-moral) trust between HAs and IATs.

References

Breazeal, C., Gray, J., Hoffman, G., & Berlin, M. 2004. “Social robots: Beyond tools to partners.” *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, pp. 551–556.

Formosa, P., 2021. “Robot Autonomy vs. Human Autonomy: Social Robots, Artificial Intelligence (AI), and the Nature of Autonomy”, *Minds & Machines* 31, 595–616. https://doi.org/10.1007/s11023-021-09579-2

Fossa, F., 2019. "I don't trust you, you faker!" On Trust, Reliance, and Artificial Agency. Teoria 39 (1):63-80.

Taddeo, M., 2017. “Trusting Digital Technologies Correctly.” *Minds & Machines* 27, 565–568. https://doi.org/10.1007/s11023-017-9450-5

Tallant, J., 2022. “Trusting What Ought to Happen.” *Erkenntnis*. https://doi.org/10.1007/s10670-022-00608-9

Tallant, J., *manuscript*

Townsend, B., Paterson, C., Arvind, T.T. *et al.* 2022. “From Pluralistic Normative Principles to Autonomous-Agent Rules.” *Minds & Machines* 32, 683–715. https://doi.org/10.1007/s11023-022-09614-w