Trust does not need to be human: it is possible to trust medical AI

Andrea Ferrario ,[1] Michele Loi ,[2] Eleonora Viganò [2]

ABSTRACT

In his recent article 'Limits of trust in medical AI,' Hatherley argues that, if we believe that the motivations that are usually recognised as relevant for interpersonal trust have to be applied to interactions between humans and medical artificial intelligence, then these systems do not appear to be the appropriate objects of trust. In this response, we argue that it is possible to discuss trust in medical artificial intelligence (AI), if one refrains from simply assuming that trust describes human–human interactions. To do so, we consider an account of trust that distinguishes trust from reliance in a way that is compatible with trusting non-human agents. In this account, to trust a medical AI is to rely on it with little monitoring and control of the elements that make it trustworthy. This attitude does not imply specific properties in the AI system that in fact only humans can have. This account of trust is applicable, in particular, to all cases where a physician relies on the medical AI predictions to support his or her decision making.

In the paper 'Limits of trust in medical AI,'[1] Hatherley provides a concise and clear overview of the current progress in the design and implementation of AI systems in medicine. At first, he describes the problem, called the 'epistemic authority and anthropocentric epistemology'[1] stemming from the performance of medical AIs, and the subsequent necessity to identify ways to design human–AI interactions in the clinical practice that could take into account the different specificities and epistemic stances of the agents involved.

At the core of Hatherley's paper lies the discussion of the limits of interpersonal trust in medical AI. This is based on the widespread distinction in the philosophical debate between reliance and trust. Classically, 'I rely on you when I predict that you will behave in a certain way, though I trust you when I judge that you ought to behave in a certain way.'[2] Accordingly, Hatherley argues that in the case of medical AIs, it is not possible to talk about trust, but only about reliance, for two reasons. First, if we believe that to trust is to consider motivations and interests, then AIs 'lack the right kind of motivation for trust—either in the form of encapsulated interest or a sense of good will—since they lack motivation entirely.'[1] Second, 'relations with AI systems cannot be said to be trusting relations, as one might have with a human clinician, since trust generates normative obligations that cannot be borne by an AI.'[1]

In summary, Hatherley states that, if we endorse the classical difference between reliance and trust typical of interpersonal trust—where the difference is based on normative and descriptive expectations—we cannot trust AI, but only rely on it.

We agree with Hatherley when he states that 'AI threatens to produce a deficit in trusting clinical relationships between doctors and patients.'[1] However, we believe that, although convenient, the choice of applying human trust to describe human–AI interactions is not fully justified. This begs the question against AI. Rather, we shall strive, as much as possible, to identify a meaningful concept of trust that is applicable to human–human and human–AI relations. If such a concept exists and can be meaningfully distinguished from mere reliance, then we can talk about trust (and not only about simple reliance) in medical AI.

We propose to redefine the dichotomy 'reliance vs trust' using the activity of monitoring: in other words, trust involves economising on monitoring. This account is called 'simple trust.'[3] Trust, that is, denotes a reliance property that describes the willingness of the physician to rely on the medical AI without intentionally generating and/or processing further information about the medical AI's capabilities to achieve the goal at hand (eg, by monitoring the medical AI).[3] Therefore, according to this account, simple trust is a property of a class of reliance relations, which is not characterised by specific trustworthiness properties,[i] but rather by a diminished willingness to actively update the belief on the trustworthiness of the medical AI one relies on. According to this account, to (simply) trust is not to identify selected properties in the medical AI we deem worthy of trust, but to rely on the medical AI without updating beliefs on its trustworthiness. The focus here is on the process of updating the trustor's belief about the trustee's trustworthiness rather than on the trustee's features reflecting qualities that only humans can have, such as normative obligations. In fact, when trust is achieved, the physician may not engage herself in costly mental processes aiming at updating her beliefs on the trustworthiness of the medical AI (ie, monitoring).[3] Notice that simple trust is not necessarily irrational. Simple trust can be accompanied by reflective trust,[3] which is the belief that simple trust in the entity is justified.[ii] In reflective trust, the physician comes to believe that an attitude of simple trust towards the medical AI is appropriate.

Let us clarify our position with an example. Medical AIs support decision making by the provision of predictions, often in the form of machine learning model outcomes, to identify and plan better prognoses, diagnoses and treatments.[iii] These outcomes are the result of complex computational processes on high-dimensional data that are difficult to understand by physicians. Therefore, it may be convenient to look at the medical AI as a 'black box', or an input–output system whose internal mechanisms are not directly accessible or understandable. Through a sufficient number of interactions with the medical AI, its developers and AI-savvy colleagues, and by analysing different types of outputs (eg, those of young patients or multimorbid ones), the physician may develop a mental model, that is, a set of beliefs, on the performance and error patterns of the AI. We describe this phase in the relation between the physician and the AI as the 'mere reliance' phase,

[1]Department of Management, Technology and Economics, ETH Zurich, Zurich, Switzerland
[2]Digital Society Initiative (DSI) and Institute of Biomedical Ethics and History of Medicine (IBME), University of Zurich, Zurich, Switzerland

Correspondence to Dr Andrea Ferrario, Management, Technology and Economics, ETH Zürich, Zurich 8092, Switzerland; aferrario@ethz.ch

[i]Such as autonomy and accountability with respect to normative obligations.

[ii]Ferrario *et al* call this a trustworthiness belief.[3] 'Trustworthiness' here is used in a special sense, referring to the belief that the entity has those features—whatever they may be—that justify simple trust in it.

[iii]However, assuming a two-step process that sees the AI contributing to the decision-making of the physician with its prediction and, later, to the formation of the final decision and its dissemination to the patient, we can assume that the primary relation is with the physician and we focus on it.

which does not need to involve trust (or at best involves very little trust).

What we are saying here is that after a sufficient number of trials, the physician would eventually entertain beliefs on the performance and error patterns of the medical AI. Therefore, at the next interaction with the medical AI, the physician could trust the AI by relying on it without updating these beliefs.[iv] This is expressed by a disposition of the physician to exert little efforts and time in further activities instrumental to belief updating, such as generating further evidence of the medical AI accuracy. This is the phase of trust.

Notice two aspects of this relation. First, it involves a combination of simple and reflective trust: the physician is not only disposed to rely on the AI with no (or little, at most) monitoring, but he also reflectively believes that this stance is appropriate, given the known AI performance, the severity of the possible harm at stake. Second, as the time and effort no longer used for monitoring can be expended in other activities, it is easy to explain why trust has been described as the lubricant of social interactions.[4]

In summary, in our account of trust, it is not the content of the physician's belief (eg, the acknowledgement of autonomy and accountability with respect to normative obligations) that defines trusting a medical AI, but rather the way reliance is supported by a process of belief generation or update (ie, no or little monitoring). It is this peculiar concession of the physician to economise on monitoring that characterises the trust relation and distinguishes it from sheer reliance.

We argue that the discussion on 'trustworthy AI' in medicine can benefit from considering this perspective, which does not start by the identification of the characteristically human desiderata of interpersonal trust, and applies them to the physician–AI relation. Rather, we can provide a meaningful account of trust in AI, distinct from mere reliance, and redefine trustworthiness on its basis. Trustworthy AI, in our view, means AI that deserves to be trusted, that is, relied on with little or no monitoring.

[iv] We are not saying that simple trust descends from rational beliefs only; in fact, we neither pose constraints on the nature of the beliefs, nor on their existence. We refer to[3] for all details.

Contributors AF is the corresponding author for this work; he drafted the manuscript. ML and EV provided important intellectual inputs to substantially improve the first draft. All authors equally contributed to paper revisions and its finalisation.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

OPEN ACCESS

To cite Ferrario A, Loi M, Viganò E. *J Med Ethics* Epub ahead of print: [please include Day Month Year]. doi:10.1136/medethics-2020-106922

Received 21 September 2020
Accepted 25 October 2020

J Med Ethics 2020;**0**:1–2.
doi:10.1136/medethics-2020-106922

ORCID iDs
Andrea Ferrario http://orcid.org/0000-0001-9968-9474
Michele Loi http://orcid.org/0000-0002-7053-4724
Eleonora Viganò http://orcid.org/0000-0002-1640-2763

REFERENCES

1 Hatherley JJ. Limits of trust in medical AI. *J Med Ethics* 2020;46(7):478–81.
2 Nickel PJ, Vaesen K. Risk and Trust. In: Roeser S, Hillerbrand R, Sandin P, *et al*, eds. *Handb risk theory Epistemol Decis theory ethics soc Implic risk*. Dordrecht, Netherlands: Springer, 2012.
3 Ferrario A, Loi M, Viganò E. In AI we trust Incrementally: a Multi-layer model of trust to analyze Human-Artificial intelligence interactions. *Philos Technol* 2020;33(3):523–39.
4 Arrow KJ. *The limits of organization*. New York: Norton, 1974.