Medical AI: is trust really the issue?

Jakob Thrane Mainz

ABSTRACT
I discuss an influential argument put forward by Hatherley in the *Journal of Medical Ethics*. Drawing on influential philosophical accounts of interpersonal trust, Hatherley claims that medical artificial intelligence is capable of being reliable, but not trustworthy. Furthermore, Hatherley argues that trust generates moral obligations on behalf of the trustee. For instance, when a patient trusts a clinician, it generates certain moral obligations on behalf of the clinician for her to do what she is entrusted to do. I make three objections to Hatherley's claims: (1) At least one philosophical account of interagent trust implies that medical AI is capable of being trustworthy. (2) Even if this account should ultimately be rejected, it does not matter much because what we care mostly about is that medical AI is reliable. (3) It is false that trust in itself generates moral obligations on behalf of the trustee.

Artificial intelligence (AI) is making its way into the field of medicine, where it promises to make significant improvements. The literature frequently reports that (A) medical AI often outperforms medical clinicians when it comes to making correct diagnoses and selecting correct treatments for specific diseases. However, the literature also reports that (B) medical AI often work as 'black boxes' with little or no transparency in the decision-making process. The tension between (A) and (B) has generated a lot of discussion on trust in medical AI. Hatherley has recently made an important contribution to the discussion. I focus on two of his main claims. The first claim is that medical AI is not capable of being trustworthy. According to the influential philosophical accounts of trust, medical AI is capable of being reliable, but not trustworthy[1] (Bjerring JC et al: p. 478).[1] On Russell Hardin's account of trust, for instance, '…reliance is insufficient for trust because trusting someone also requires a belief that one's interests are encapsulated in the interests of the trusted person.' (Bjerring JC et al: p. 480).[1] Hatherley's point is that all the influential accounts of interpersonal trust require that the trustee is capable of having certain motives, intentions or attitudes. When it comes to medical AI, these requirements seem to be demonstrably unsatisfied. Hatherley's second claim is that trust generates moral obligations. For instance, if a patient trusts a clinician, then it generates a moral obligation on behalf of the clinician to do what she is entrusted to do (Bjerring JC et al: p. 480).[1]

I make three objections to Hatherley's claims:

1. First, at least one philosophical account of interagent trust implies that medical AI is capable of being trustworthy.
2. Second, even if this account should ultimately be rejected, it does not matter much because what we care mostly about is that medical AI is reliable.
3. Third, it is false that trust in itself generates moral obligations on behalf of the trustee.

(1) and (2) are objections to Hatherley's first claim, and (3) is an objection to his second claim. Let us begin with. (1) Coeckelbergh's account of trust straightforwardly implies that medical AI is capable of being trustworthy. His account holds that—despite the prima facie inconsistency with the interpersonal accounts of trust—artificial agents are capable of being trustworthy.[2] He defends—inter alia—a functionalist view of trust, and claims that to say that an artificial agent is trustworthy is to say that it is believed that it will do the things that it is supposed and expected to do (Coeckelbergh M: p. 58).[2] Hatherley may be correct that all accounts of interpersonal trust imply that medical AI cannot be trustworthy, but it is not true of all accounts of interagent trust. Given that at least some types of medical AI can reasonably be characterised as artificial agents (at least on some accounts of artificial agency), it seems more relevant to ask if medical AI is capable of being trustworthy on accounts of interagent trust, rather than on accounts of interpersonal trust.

Now consider point (2). Hatherley might object to what I said about (1) that even though Coeckelbergh's account of trust implies that medical AI is capable of being trustworthy, it should nevertheless be rejected. Hatherley could object that it rests on phenomenological and communitarian assumptions that are at least questionable, or that adopting a functionalist account of trust collapses trust into reliability. I remain agnostic about both of these objections. But even if they are true, it does not matter in the first place whether we can trust medical AI in the same way that we can trust a human clinician. Hatherley's first claim is vulnerable to what we might call a 'so what' objection. If what sets reliability and trust apart is that trust entails that the trustee is capable of having certain intentions or motives, then it is a trivial and unproblematic implication that medical AI cannot be trustworthy. What we care mostly about is whether the algorithm is reliable, that is, that the outputs of the algorithm are accurate, and that the algorithm performs consistently.[i] If the algorithm performs consistently better than human practitioners do, all things considered, then practitioners have obligations to follow the recommendations of the algorithm. Bjerring and Busch put it this way: 'if a practitioner knows of an epistemic source that is more knowledgeable, more accurate and more reliable in decision-making, she should treat it as an expert and align her verdicts with those of the source'.[1]

To put it bluntly, criticising medical AI for not being trustworthy is like criticising your beloved vintage sports car for not loving you back. It is simply not capable of doing what you are criticising it for not doing. However, it was unreasonable to expect—and perhaps unreasonable to desire—that it would be capable of doing so in the first place. If we were talking about a highly developed conscious AI, it would probably be reasonable to ask if we can trust it. But at this point, we are talking about a deep learning algorithm which is—after all—just an advanced statistical model.

Finally, consider point (3). Hatherley explains that on some accounts of trust, moral obligations are generated when someone trusts someone else. Hatherley affirms this claim and says that if a patient trusts a clinician, then it generates a moral obligation on behalf of the clinician to do what she is entrusted to do (Bjerring JC et al: p. 480).[1] Hatherley recognises that trust does not always generate moral obligations. He writes: 'There are some important limitations to this claim, for example, in circumstances where the trust

Aarhus Universitet, Aarhus, Denmark

Correspondence to Dr Jakob Thrane Mainz, Aarhus Universitet, Aarhus, Midtjylland, Denmark; jakob-mainz@hotmail.com

[i]However, we should also care about other important issues, such as biases in the outputs of the algorithm.

that one has in another is misguided or unwelcome. Suppose, for instance, that one were to place their trust in a friend who is a dermatologist to remove their wisdom teeth. Trusting the dermatologist for this procedure would appear quite mistaken, given that the dermatologist does not have the expertise or competency to perform this task. Nor, presumably, would the dermatologist welcome this trust in any way.' (Bjerring JC et al: p. 480).[1] Similarly, Coeckelbergh suggests that no moral obligations are generated if the trustee does not know that she is being trusted by the trustor (Coeckelbergh M: p. 55).[2] However, even with these caveats in place, the claim is false. Trusting someone does not—in itself—generate any moral obligations, even in a clinician–patient relation.

Many counterexamples to Hatherley's second claim come to mind. Here is one: Jones is a clinician, and Smith is his patient. They are both Jehovah's Witnesses. Clara is married to Smith, and they have a daughter. Clara is an atheist. Jones tells Smith that he will survive his surgery only if he gets a blood transfusion from the daughter. Smith declines the offer. He tells Jones that he trusts him not to tell Clara about the blood transfusion option. Jones welcomes Smith's trust and stays silent. Smith dies.

Hatherley's claim implies that Jones has an obligation not to tell Clara about the blood transfusion. This is at least a very controversial implication. Jones may have a legal obligation to comply with the rules of doctor–patient confidentiality, but it seems implausible that Jones has a moral obligation not to mention the blood transfusion to Clara. Trust does not in itself generate any moral obligations. Faulkner has recently put it this way: 'What the trusted should do in the trust situation is determined by how things are in the world rather than by the attitudes of the trusting party'.[3] (Hatherley JJ: 342).[4] If an account of trust implies that trust generates moral obligations, then so much the worse for that account.

Funding This study was funded by Carlsbergfondet (CF20-0257).

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

To cite Mainz JT. *J Med Ethics* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jme-2023-109414

Received 7 July 2023
Accepted 15 July 2023

J Med Ethics 2023;**0**:1–2.
doi:10.1136/jme-2023-109414

REFERENCES

1 Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. *Philosophy & Technology* 2020;141.
2 Coeckelbergh M. Can we trust robots *Ethics Inf Technol* 2012;14:53–60.
3 Faulkner P. The moral obligations of trust. *Philosophical Explorations* 2014;17:332–45.
4 Hatherley JJ. Limits of trust in medical AI. *J Med Ethics* 2020;46:478–81.