Cambridge Quarterly of Healthcare Ethics (2022), 1–10
doi:10.1017/S0963180122000445

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Misplaced Trust and Distrust: How Not to Engage with Medical Artificial Intelligence

Georg Starke and Marcello Ienca*

College of Humanities, EPFL, 1015 Lausanne, Switzerland
*Corresponding author. Email: marcello.ienca@epfl.ch

Abstract

Artificial intelligence (AI) plays a rapidly increasing role in clinical care. Many of these systems, for instance, deep learning-based applications using multilayered Artificial Neural Nets, exhibit epistemic opacity in the sense that they preclude comprehensive human understanding. In consequence, voices from industry, policymakers, and research have suggested trust as an attitude for engaging with clinical AI systems. Yet, in the philosophical and ethical literature on medical AI, the notion of trust remains fiercely debated. Trust skeptics hold that talking about trust in nonhuman agents constitutes a category error and worry about the concept being misused for ethics washing. Proponents of trust have responded to these worries from various angles, disentangling different concepts and aspects of trust in AI, potentially organized in layers or dimensions. Given the substantial disagreements across these accounts of trust and the important worries about ethics washing, we embrace a diverging strategy here. Instead of aiming for a positive definition of the elements and nature of trust in AI, we proceed *ex negativo*, that is we look at cases where trust or distrust are misplaced. Comparing these instances with trust expedited in doctor–patient relationships, we systematize these instances and propose a taxonomy of both misplaced trust and distrust. By inverting the perspective and focusing on negative examples, we develop an account that provides useful ethical constraints for decisions in clinical as well as regulatory contexts and that highlights how we should *not* engage with medical AI.

Keywords: AI medicine; ethics; decision-making; trust; distrust

Trust in Medical Artificial Intelligence

Trust is fundamental to human interactions. It helps us efficiently navigate complex social situations and offers a strategy to act under uncertainty. But are we justified to trust machines, let alone opaque machines used in medicine? In light of recent advancements in the field of artificial intelligence (AI) and the increasing integration of AI systems into clinical practice and medical technology, the question of their trustworthiness seems more pressing than ever. Accordingly, the past years have seen a vast expansion of the ethical literature scrutinizing medical AI, and the notions of trust and trustworthiness often feature prominently in these discussions. Attention to the topic of trust was further fuelled by the EU Commission's High-Level Expert Group on AI, whose 2019 *Ethics Guidelines for Trustworthy AI* provided an important cue for academic debate.

Despite the large corpora of academic research on the subject, there remain fundamental disagreements on the nature of trust.[1,2] To some extent, this seems little surprising given the many diverging accounts of trust theory developed for different contexts by sociologists, (moral) philosophers, neuroscientists, and psychologists working on the topic.[3,4,5,6] Yet, since the bioethical debate about medical AI strives not only for improvements in theory but should also provide practical guidance to healthcare professionals, developers, and regulatory bodies, this state of affairs seems unsatisfactory.

In this article, we, therefore, address the issue from the opposite direction focusing on the negative, that is, on misplaced trust and distrust. In doing so, we refrain from proposing a comprehensive, novel account of trust that covers the many different facets of a complex phenomenon but offers a limiting framework that serves to highlight how we *should not* engage with medical AI. Our article proceeds in four steps. We commence with a brief overview of the existing literature and some of the most prominent positions in the field, which allows us to arrive at a working definition of trust for the purpose of this article. In a second step, we advance two distinctions that are crucial to our argument, namely a semantic and conceptual clarification of trust and trustworthiness as well as the distinction between internal and external trustworthiness, loosely based on the work of Jacovi et al.[7] In a third step, we then discuss and systematize ways where trust or mistrust are misplaced, resulting from a mismatch between trusting actions and the AI's trustworthiness, highlighting each instance with a practical example. We conclude our article with an outlook on how focusing on misplaced trust can guide action, advance warranted trust by focusing on an AI system's components of trustworthiness, and how delineating shortfalls can help addressing warranted concerns about ethics washing of AI.

We take it that this approach is commensurate with an understanding of medical ethics that does not pretend to offer comprehensive, catch-all instructions on how one should act in every single instance but instead aims to provide a general framework that leaves the peculiarities of the particular case to individual discernment. Such a framework providing a mere outline can nevertheless guide action. For as Aristotle put it in the Nicomachean Ethics, "the proper procedure is to begin by making a rough sketch, and to fill it in afterwards. If a work has been well laid down in outline, to carry it on and complete it in detail may be supposed to be within the capacity of anybody" (1098a).[8]

The Many Forms of Trust in AI

In March 2019, the EU Commission's High-Level Expert Group on Artificial Intelligence published their widely received *Ethics Guidelines for Trustworthy AI.*[9] These guidelines set out to formulate conditions for lawful, ethical, and robust AI taken to be crucial for trustworthy AI systems.[10] The recommended conditions for ethical AI are structured around four principles: respect for human autonomy, prevention of harm, fairness, and explicability. These four principles were, in turn, largely derived from the AI4people framework,[11] that synthesized 47 recommendations from six international regulatory suggestions into five principles, representing the classical principles of bioethics, beneficence, nonmaleficence, respect for autonomy, and justice,[12] complementing them with an AI-specific principle of explicability.[13]

Following the lead of the EU guidelines, trust has taken center stage in academic debates about the ethics of AI. In particular, this is true for the case of medical AI, given that trust plays an undeniably vital role in the medical domain, whether in private between patients and their healthcare providers,[14] or in public contexts with a view to trust in health care systems.[15] Trust is therefore often ascribed a dual value in the medical domain: an instrumental value, as it, for example, enables patients to openly share their symptoms with their physicians, and an intrinsic value, reflecting for instance a physician's fiduciary obligations toward their patients.[16]

With a particular view on medical AI, many authors strongly oppose the notion of trust though. They argue that trust in medical AI is too anthropomorphist a notion that lends itself to ethics-washing, and hold that it constitutes an error of categories to talk about trust outside of human–human interaction.[17,18,19] These concerns have been mirrored by others with particular focus on medicine, shedding further light onto the specific risks involved in medical contexts and the often-complex interactions between healthcare professionals, patients, and AI systems.[20,21]

On the other side of the debate are authors who stress aspects of the conceptually rich notion of trust that are commensurate with an application to medical AI. Juan Durán and Karin Jongsma for instance have defended trust in opaque medical AI based on "computational reliabilism"[22] that rests on four criteria, namely on verification and validation methods, robustness analysis, the history of (un)successful implementations, and expert knowledge. In a convincing reductive approach that draws on normative

https://doi.org/10.1017/S0963180122000445 Published online by Cambridge University Press

theories of trust,[23,24] Philip Nickel has suggested a model that understands trust in medical AI as "giving discretionary authority" to an AI, that is, to vest an AI with discretion to answer medically important questions such as diagnosis or treatment decisions.[25]

Both approaches embrace a normative notion of trust that differs importantly from the kind of trust that Andrea Ferrario, Michele Loi, and Eleonora Viagnò have called "simple trust," which they describe as a noncognitive willingness to rely on a specific agent.[26,27] Going beyond this basic form of trust modeled on reliance, the authors further describe additional ways of trusting that can be seen as incremental layers: "reflective trust" goes beyond simple trust insofar as it only cedes control over a decision after cognitively assessing the reliability of an agent to perform a specific task. Following this model, "paradigmatic trust," finally, brings together elements of simple and reflective trust since it constitutes a disposition to rely on a specific agent without assessing grounds for this reliance further, but only after such trust has been vindicated successfully through past experience in the form of reflective trust.

It is apparent that these different forms of trust mirror substantive and long-standing disagreements in trust theory more general. For instance, they draw on debates whether trust should be understood as cognitive or noncognitive,[28] whether it requires presupposing some form of goodwill on the side of the trusted,[29,30] or whether trust arises from a rational assessment that the trusted agent has encapsulated your interest, that is, that they have an interest themselves to act as expected by the trusting party.[31,32]

It would be surprising if these disagreements were to be settled on the grounds of AI ethics. We, therefore, believe that our approach to start with examples of failed trust and distrust can contribute to advancing the debate. However, before we can turn to the question why and when trust fails, we briefly need to address the property that justifies warranted trust, namely trustworthiness.

Trust and Trustworthiness

As we have shown in the previous sections, there are many forms and models of trust. In the following, we will adhere to a rather modest, yet normatively substantial understanding of trust in line with Nickel's account.[33] Following the standard account of trust as a tripartite notion of the form that X trusts Y with regard to Z,[34,35] we take trust in medical AI to be a disposition by an AI user to provide an AI system with discretionary power with view to a medically relevant task, based on a cognitive assessment of said system. With view to ethical deliberation, one, therefore, needs to examine the properties of an AI system that render trust justified. For if trust is to be ethically meaningful, it needs to be bound to certain conditions of *trustworthiness*.

Similar to the many models of trust, one can also discern between many different aspects of trustworthiness and systematize them in different ways. The principles of the EU guidelines for trustworthy AI or the proposed criteria of computational reliabilism by Durán and Formanek can be seen as attempts to grasp trustworthiness.[36] Similarly, drawing on the Daniel Dennett's model of three stances, Starke et al.[37] have recently suggested a dimensional model that can help disentangle different aspects of trustworthiness with view to reliability, competence, and intentions. In the context of medical AI, these aspects help scrutinize a system asking whether it works under the specified conditions, how well it performs in a particular task, and whether the "intentions" of the system, as imbued in it by the developers, are compatible with the intentions of the AI user.

A further, complementary distinction of trustworthiness can be derived from the model of trust in AI that has been developed by Alon Jacovi et al.[38] While the authors do not specifically focus on medicine, they still provide a helpful distinction between two different ways of bringing about *trust* in AI, namely an intrinsic and an extrinsic way. Their account seems related to a distinction Mark Coeckelbergh has proposed in the context of robotics, between direct and indirect trust,[39] and can be made fruitful to understand how fulfilling requirements of trustworthiness can foster warranted trust.[40] Reformulating the considerations of Jacovi et al. with a view to *trustworthiness* allows to sort out our ensuing discussion of warranted and unwarranted trust with view to intrinsic and extrinsic factors (see Figure 1).

https://doi.org/10.1017/S0963180122000445 Published online by Cambridge University Press

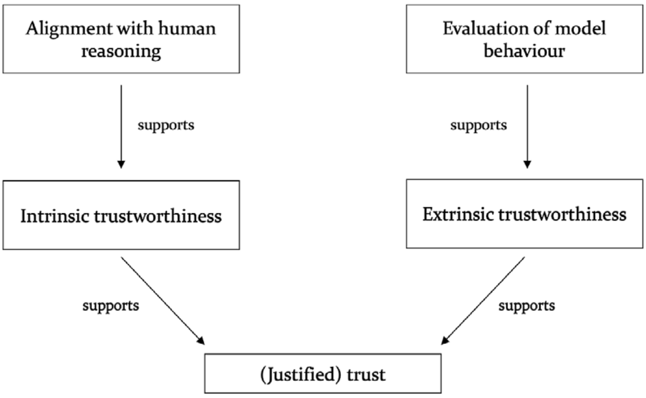
Alignment with human reasoning
supports
Intrinsic trustworthiness
supports
Evaluation of model behaviour
supports
Extrinsic trustworthiness
supports
(Justified) trust

Figure 1. Two kinds of AI trustworthiness.

On the one hand, trustworthiness can rely on *internal* factors, especially when the reasoning process of the ML model is explainable, whether ex-ante or ex-post, and aligns with human reasoning.[41] If indeed a model can increase its trustworthiness through alignment with human reasoning, as seems plausible following recent simulation-based surveys,[42,43] a model's perceived internal trustworthiness may not only be fostered by technical explainability but could also be supported by approaches that focus on understanding.[44] Internal trustworthiness is therefore not merely dependent on full disclosure but presupposes successful communication.[45] Such focus on communication in turn can help avoid dangers of using explainability as a cover for poorly trained models or as a means to mask unethical decisions of a biased AI system.[46,47]

On the other hand, *extrinsic* trustworthiness relies on observations not of the inner workings of an opaque algorithm, but of its behavior in a specific context.[48] Assuming that the black-box nature of the model cannot be circumvented, this type of trustworthiness is supported by methodological rigor in the evaluation process, and by proper regulation. In the same vein, scrutinizing a clinical AI system's fairness primarily with view to its outcome in different protected groups would foster such external trustworthiness.[49]

Both intrinsic and extrinsic trustworthiness contribute to justified trust. But what about instances where the trustor is aware that the trustee does not fulfill criteria of trustworthiness, that is, that they are untrustworthy? Here, they would be justified to embrace a stance of *distrust*. Following Russell Hardin, we understand distrust here not as a mere absence of trust but as a cognitive disposition that mirrors the tripartite structure of trust in that X distrusts Y with regard to Z, based on an assessment of the trustee's untrustworthiness.[50]

Before we proceed, it should be noted that in light of these many different aspects of trustworthiness, some authors have rightly admonished that trust should not be reduced to formulaic checklists.[51] We also do not intend to provide a comprehensive list here but will apply the distinctions raised above to cases where trust and distrust are clearly *not* justified.

Misplacing Trust and Distrust in Medicine

As we have seen, highlighted for instance by the model of "simple trust" by Ferrario and colleagues, trusting dispositions do not necessarily consider questions of trustworthiness on a cognitive level. Yet, if we aim for an ethically justified form of trust, users of medical AI should only trust trustworthy systems and deny trust to untrustworthy systems.[52] At the same time, given the potentially high cost of unwarranted distrust, for instance by foregoing an accurate diagnostic tool, AI users should only distrust untrustworthy AI systems. Put differently, we should only trust the trustworthy and distrust the untrustworthy and be able to support our trust with relevant and justifiable reasons.

https://doi.org/10.1017/S0963180122000445 Published online by Cambridge University Press

The reality of trust, however, is unfortunately more complicated and messier than the simple exhortation to trust the trustworthy may indicate.[53,54,55] The reason for that stems from the fact that it is logically and factually possible for an agent to trust a trustworthy system (or other agent) based on erroneous and/or ethically unjustified beliefs or motivations. Vice versa, it is equally possible to distrust an untrustworthy system (or agent) for a wrong and/or unjustifiable reason. In epistemology, the so-called "tripartite analysis of knowledge" asserts that three conditions are necessary and sufficient to knowledge: truth, belief, and justification. While the tripartite analysis of knowledge has been challenged by a landmark philosophical problem known as the Gettier problem,[56] it nonetheless offers valuable guidance on the necessary (albeit perhaps not sufficient) conditions to knowledge.[57] By applying such epistemological precepts to ethical analysis, we argue that it is equally crucial from a normative point of view to discern whether human agents trust an AI system (or agent) due to its external or internal trustworthiness, or whether they base their expenditure or withholding of trust on some other, potentially erroneous or unjustified beliefs. In other words, we claim that there is, ethically speaking, an unjustified way to trust the trustworthy just as there is an unjustified way to distrust the untrustworthy. To support this claim, we suggest a taxonomy of trust and distrust in the following (Figure 2), each of which can be related to different hypothetical case scenarios.

Our taxonomy highlights that there are more ways in which we can misplace trust or distrust than there are correct ones. To highlight possible shortcomings, let us look at each of the eight scenarios separately by comparing the case of hypothetical medical AIs with the case of trust placed in or denied to a human agents.

To Trust or Not to Trust

Let us first consider the paradigmatic case of justified trust in the trustworthy before turning to three logically possible ways of failed trust. Under ideal circumstances, a patient may trust their physician with regard to a therapeutic recommendation, based on an evaluation of their education, their clinical

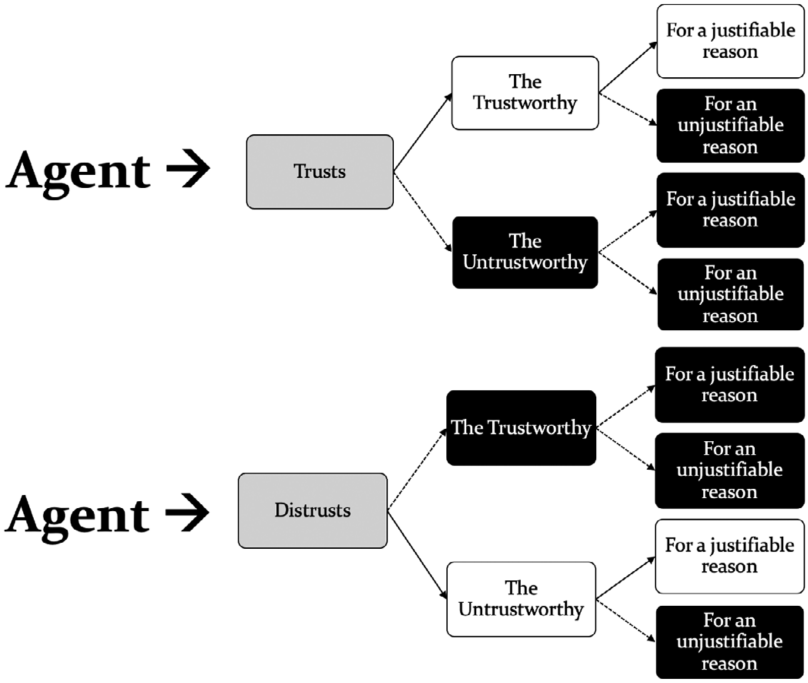

Figure 2. Taxonomy of misplaced trust and distrust. The lighter boxes with dark font denote a morally adequate placement of, respectively, trust and distrust. In contrast, the darker boxes with white font denote misplaced forms of trust.

https://doi.org/10.1017/S0963180122000445 Published online by Cambridge University Press

experience, and minimal formal criteria such as their medical license. Analogously, a patient may also trust a trustworthy medical AI with view to a recommended medication, based on an evaluation of the AI according to what Duràn and Jongsma have proposed as computational reliabilism, that is, the system's past performance history, its proper validation, its robustness, and its alignment with expert knowledge.[58] Given that such assessment may lie beyond the practical possibilities of the trusting party, they may also refer to a purely formal criteria such as approval by the FDA or EMA based on suitably shaped randomized-controlled trials.[59]

However, with view to the very same physician or AI system, trust may also be expedited erroneously. For instance, let us suppose that the patient trusts the ideal physician from the case above with a therapeutic recommendation, but based merely on the fact that said physician is a remote friend of their sister-in-law. In the case of the AI, one could easily envision a similar case, in which trust is placed based on a trustworthy system to come up with a suggested medication, but that this judgment is made not based on the trustworthiness of the system but merely on the convenience of its use. The obvious question in both cases is, why would such trust be problematic? After all, the likely result would be identical to the positive outcome of paradigmatic trust, and the patient would benefit from a reasonable drug recommendation. The important divergence, we believe, lies therefore not in its outcome but in the path to the trusting relationship. With a view to the individual case, the beneficial outcome may obscure the difference but with view to the general approach toward the trustworthiness of physicians or AI systems, the two cases differ fundamentally, as highlighted by Berkeley's formulation of rule-consequentialism: "The rule is framed with respect to the good of mankind; but our practice must be always shaped immediately by the rule."[60] Similarly, approving trust in cases where the outcome happens to be positive but without subjugating the trusted party to a proper investigation of their trustworthiness seems highly problematic since without cognitive scrutiny, such trusting behavior is likely to be extended to the untrustworthy, with potentially devasting consequences.

As a third and fourth instance in our taxonomy, let us consider cases in which trust is expedited to the untrustworthy. For the sake of the argument, let us suppose that a patient was to trust an untrustworthy conspiracy theorist charlatan with an important medical question, such as a treatment suggestion for their lung carcinoma. Again, there could be two ways in which this trust could be motivated, a right one and a wrong one. In the first case, trust would ensue based on the fact that the recommendation of the untrustworthy healthcare professional coincided, for once, with general treatment suggestions from the European Society for Medical Oncology. In contrast, in the second, worse case the patient would trust the dangerously wrong treatment recommendation of said charlatan based on aggressive online advertisements. Similarly, one could envision parallel cases for AI, namely an AI system which provides an (intrinsically) trustworthy treatment recommendation for a particular patient, aligning with human judgment, but does not fulfill evaluation standards of external trustworthiness and would be dangerous to trust for other patients. For instance, one may think of the widely-used example of Watson for Oncology here, which suggested too aggressive treatments for cancer patients.[61] Here, the end-user may again expedite trust based on a good reason, in contrast to cases where trust in the same, untrustworthy system would be brought about by an aggressive marketing strategy from the developers.

As cases 3 and 4 highlights, there are important differences at stake here. In the latter case, the health of the trusting patient may be endangered, whereas in the third case this may not happen immediately. Nevertheless, also in the third scenario, it would clearly be a fallacy to *trust* the system. Strictly speaking, if the end-user does indeed check all recommendations of the system against current guidelines, this instance in itself would not fall under a discretionary definition of trust at all, which entails ceding power of judgment to the AI system. However, one could very well envision that an end-user decides to trust the system based on the one decision in line with current guidelines—which may gravely endanger their physical well-being if the AI is in fact untrustworthy. So, in light of all these potential pitfalls, may it be wiser to distrust medical AI systems in general?

https://doi.org/10.1017/S0963180122000445 Published online by Cambridge University Press

How (Not) to Distrust Medical AI

Unfortunately, distrusting correctly seems almost as difficult as trusting correctly. Following our proposed taxonomy, let us consider the following instances, in which the trustworthy are erroneously distrusted. Two ways stand out here. First, in the case of human interaction, one may envision a well-trained, experienced, licensed physician. In certain instances, it may seem reasonable to distrust their guidance, for instance, if their therapeutic recommendation in a particular case does not conform to agreed guidelines. Analogously, a well-trained and—tested, approved AI system that can be considered trustworthy in general may provide dangerous, outlying advice in a particular instance. While not following the guidance in these cases may prove to be beneficial in this individual instance, it would overall still prove to be disadvantageous to distrust said physician or system, and disregard their suggestions altogether.

Distrusting the trustworthy may also be based on wrong judgments. The very same physician of the previous example may, for example, be mistrusted based on the fact that the physician happens to be Black, due to pervasive racist prejudices. In the same vein, a trustworthy AI system may be falsely mistrusted due to an end user's antisemitism who believes that the CEO of the tech company behind AI, who happens to be Jewish, aims to subdue the entire world to a Jewish world conspiracy—neither of which examples is as far-fetched as it should be.

The paradigmatic, justified scenario of distrust could be conceptualized as an instance in which a patient distrusts a charlatan practitioner who promises to cure cancer, but the patient, being aware of the charlatan's lack of education and the stark deviation from professional treatment guidelines, Similarly, a patient from an ethnic minority group would be justified to distrust an AI to provide an accurate diagnosis because they are aware that the AI has been trained exclusively on data from different populations and is therefore likely to provide misleading recommendations.

In contrast, a patient may also erroneously distrust the same untrustworthy agents, whether human or AI, based on unjustified reasons, for instance, because the untrustworthy health practitioner happens to be Black and the patient is moved by racist assumptions, or the end-user may distrust an untrustworthy AI to provide an accurate treatment recommendation based on the unfounded assumption that the CEO of the company developing the AI is part of a secret maleficent cabal.

Again, there are important differences between these cases. For instance, if a trustworthy AI is falsely distrusted, the gravest harm is likely to befall the overly suspicious trustor who needlessly foregoes a potentially helpful tool for their medical care. In the case of the Black physician, however, not only does the distrusting patient harm themselves but they also wrong the healthcare professional with racist behavior.

Implications of Failed Trust and Distrust

There are at least reasons why we believe that a clear taxonomy of different kinds of failed trust and distrust matters. First, it can help stress that there are many different ways in which trusting relationships can go wrong, but that not all are equally problematic from a normative point of view. Trusting the untrustworthy who happens to give good advice, for instance, seems dangerous with view to the general rule but is unlikely to harm the trusting person in the particular case. Also, assuming that an AI system cannot be harmed (yet), it helps stress differences between cases in which a falsely distrusted human party would be wronged and where a distrusted AI would not.

Second, by focusing more directly on reasons why agents factually trust or distrust a system, our approach can help build a stronger bridge between empirical literature from human–machine interaction studies and psychology and the bioethical literature on trust. As we have discussed with view to different models of explainability, it is far from clear that the ethically most trustworthy model of an AI actually gains the strongest trust from humans.[62] Instead, as Alam and Mueller have shown in simulations, it is commonly the context and the way in which explanations are provided, determining one aspect of an AI system's trustworthiness, that shapes patient satisfaction, and trust.[63]

https://doi.org/10.1017/S0963180122000445 Published online by Cambridge University Press

Third, by calling attention to the fact that the trust given to a system and its trustworthiness are not necessarily linked, our account aims to shift the ethical debate back toward being centered around the conditions of external and internal trustworthiness. Such refocusing in turn may help counter the much-criticized tendencies of ethics washing in this context,[64,65,66] by calling out past and future breaches of trust.

Conclusion and Normative Implications

Our analysis showed that it is logically and factually possible for an agent to trust a trustworthy system (or other agents) based on erroneous and/or ethically unjustified beliefs or motivations. Vice versa, it showed that it is equally possible to distrust an untrustworthy system (or agent) for a wrong and/or unjustifiable reason. This implies that, respectively, trusting the trustworthy and distrusting the untrustworthy are not sufficient conditions for ensuring ethically justifiable trust relations between humans and AI systems. We, therefore, argue that it is crucial from a normative-ethical point of view to discern whether human agents trust an AI system (or another human agent) due to its external or internal trustworthiness, or whether they base their attribution or withholding of trust on some erroneous or unjustified beliefs. As our taxonomy highlights, there are more ways in which we can misplace, respectively, trust or distrust than ways in which that placement of either trust or distrust is epistemically and ethically justified.

This analysis has both conceptual and normative implications. From a conceptual perspective, it can help elucidate cases where the placement of either trust or distrust in an AI system, albeit *prima facie* reasonable, is actually unjustified. From a normative perspective, it can help decision-makers navigate the complexity of trust dynamics in human–AI interaction and promote policies that prevent the misplacement of trust in AI systems under conditions where such placement of trust is unwarranted (or, vice versa, policies that prevent the misplacement of distrust under conditions where such skeptical attitude is unwarranted). We call producers of AI ethics guidelines at all levels to incorporate these considerations into their body of normative guidance.

Funding.

This work has been supported by the ERA-NET NEURON project HYBRIDMIND (Swiss National Science Foundation 32NE30_199436).

Notes

1. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine* 2020;1–2:100001.
2. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 2019;**1**(9):389–99.
3. Tallant J. You can trust the ladder, but you should not. *Theoria* 2019;**85**(2):102–18.
4. McLeod C. Trust. In: Zalta EN, ed. *The Stanford Encyclopedia of Philosophy*. Stanford: Metaphysics Research Lab, Stanford University; 2015.
5. Luhmann N. *Trust and Power*. English ed. Chichester: Wiley; 1979.
6. Krueger F, Meyer-Lindenberg A. Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends in Neurosciences* 2019;**42**(2):92–101.
7. Jacovi A, Marasović A, Miller T, Goldberg Y. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery; 2021:624–35.
8. Rackham AH. *Ethica Nicomachea*. Cambridge, MA: Harvard University Press; 1934.

https://doi.org/10.1017/S0963180122000445 Published online by Cambridge University Press

9. European Commission. *High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI*. Brussels: EU Commission; 2019.

10. European Commission. *Proposal for a Regulation of the European Parliament and of the Council of Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. Brussels: European Commission; 2021.

11. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, *et al.* AI4People-an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 2018;**28**(4):689–707.

12. Beauchamp TL, Childress JF. *Principles of Biomedical Ethics*. 7th ed. New York: Oxford University Press; 2013.

13. See note 11, Floridi et al. 2018.

14. O'Neill O. *Autonomy and Trust in Bioethics*. Cambridge: Cambridge University Press; 2002.

15. Gille F, Smith S, Mays N. Why public trust in health care systems matters and deserves greater research attention. *Journal of Health Services Research & Policy* 2015;**20**(1):62–4.

16. Hatherley JJ. Limits of trust in medical AI. *Journal of Medical Ethics* 2020;**46**(7):478–81.

17. Metzinger T. Ethics washing made in Europe. *Der Tagesspiegel* 2019 Apr 8.

18. Bryson J. AI & Global Governance: No one should trust AI. United Nations University Centre for Policy Research. *AI & Global Governance* 2018 Nov 13.

19. Ryan M. In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics* 2020;**26**(5):2749–67.

20. See note 16, Hatherley 2020.

21. DeCamp M, Tilburt JC. Why we cannot trust artificial intelligence in medicine. *The Lancet Digital Health* 2019;**1**(8):e390.

22. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 2021;**47**(5):329–35.

23. Baier A. Trust and antitrust. *Ethics* 1986;**96**(2):231–60.

24. Hawley K. Trust, distrust and commitment. *Noûs* 2014;**48**(1):1–20.

25. Nickel PJ. Trust in medical artificial intelligence: A discretionary account. *Ethics and Information Technology* 2022;**24**(1):1–10.

26. Ferrario A, Loi M, Viganò E. In AI we trust incrementally: A multi-layer model of trust to analyze human–artificial intelligence interactions. *Philosophy & Technology* 2020;**33**(3):523–39.

27. Ferrario A, Loi M, Viganò E. Trust does not need to be human: It is possible to trust medical AI. *Journal of Medical Ethics* 2021;**47**(6):437–8.

28. Becker LC. Trust as noncognitive security about motives. *Ethics* 1996;**107**(1):43–61.

29. Baier A. What is trust? In: Archard D, Deveaux M, Manson NC, Weinstock D, eds. *Reading Onora O'Neill*. Oxford: Routledge; 2013:175–85.

30. O'Neill O. Trust before trustworthiness? In: Archard D, Deveaux M, Manson NC, Weinstock D, eds. *Reading Onora O'Neill*. Oxford: Routledge; 2013:237–8.

31. Hardin R. Trustworthiness. *Ethics* 1996;**107**(1):26–42.

32. Hardin R. *Trust and Trustworthiness*. New York: Russell Sage Foundation; 2002.

33. See note 25, Nickel 2022.

34. See note 4, McLeod 2015.

35. Baier A. *Trust, The Tanner Lectures on Human Values*. Princeton, NJ: Princeton University Press; 1991.

36. Durán JM, Formanek N. Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines* 2018;**28**(4):645–66.

37. Starke G, van den Brule R, Elger BS, Haselager P. Intentional machines: A defence of trust in medical artificial intelligence. *Bioethics* 2022;**36**(2):154–61.

38. See note 7, Jacovi et al. 2021.

39. Coeckelbergh M. Can we trust robots? *Ethics and Information Technology* 2012;**14**(1):53–60.

40. Ferrario A, Loi M (January 28, 2022). How Explainability Contributes to Trust in AI. 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Available at SSRN 2022.

https://doi.org/10.1017/S0963180122000445 Published online by Cambridge University Press

41. See note 7, Jacovi et al. 2021.
42. Alam L, Mueller S. The myth of diagnosis as classification: Examining the effect of explanation on patient satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics and Decision Making* 2021;**21**:178.
43. Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association* 2020;**27**(4):592–600.
44. Starke G, Poppe C. Karl Jaspers and artificial neural nets: On the relation of explaining and understanding artificial intelligence in medicine. *Ethics and Information Technology* 2022;**24**:26.
45. Starke G. The emperor's new clothes? Transparency and trust in machine learning for clinical neuroscience. In: Friedrich O, Wolkenstein A, Bublitz C, Jox RJ, Racine E, eds. *Clinical Neurotechnology Meets Artificial Intelligence: Philosophical, Ethical, Legal and Social Implications*. Cham: Springer; 2021:183–96.
46. Starke G, Schmidt B, De Clercq E, Elger B. Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry. *AI and Ethics* 2022.
47. John-Mathews J-M. Some critical and ethical perspectives on the empirical turn of AI interpretability. *Technological Forecasting and Social Change* 2022;**174**:121209.
48. See note 7, Jacovi et al. 2021.
49. Starke G, De Clercq E, Elger BS. Towards a pragmatist dealing with algorithmic bias in medical machine learning. *Medicine, Health Care and Philosophy* 2021;**24**:341–9.
50. Hardin R. *Distrust*. New York: Russell Sage Foundation; 2004.
51. Braun M, Bleher H, Hummel P. *A leap of faith: Is there a formula for* "trustworthy" AI? *Hastings Center Report* 2021;**51**(3):17–22.
52. See note 30, O'Neill 2013.
53. See note 14, O'Neill 2002.
54. O'Neill O. *A Question of Trust. The BBC Reith Lectures 2002*. Cambridge, MA: Cambridge University Press; 2002.
55. Hartmann M. *Vertrauen: Die unsichtbare Macht*. Frankfurt am Main: Fischer; 2020.
56. Gettier EL. Is justified true belief knowledge? *Analysis* 1963;**23**(6):121–3.
57. Boghossian P. *Fear of Knowledge: Against Relativism and Constructivism*. Oxford: Clarendon Press; 2007.
58. See note 22, Durán, Jongsma 2021.
59. Grote T. Randomised controlled trials in medical AI: Ethical considerations. *Journal of Medical Ethics* 2021.
60. Berkeley G. Passive obedience. In: Wright GN, ed. *The Works of George Berkeley*. London: Tegg; 1712:17.
61. Ross C, Swetlitz, I. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Stat News* 2018.
62. See note 47, John-Mathews 2022.
63. See note 42, Alam, Mueller 2021.
64. See note 16, Hatherley 2020.
65. See note 17, Metzinger 2019.
66. Crawford K. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press; 2021.

Cite this article: Starke G and Ienca M (2022). Misplaced Trust and Distrust: How Not to Engage with Medical Artificial Intelligence. *Cambridge Quarterly of Healthcare Ethics*: 1–10, doi:10.1017/S0963180122000445