Why we should talk about institutional (dis)trustworthiness and medical machine learning

De Proost, Michiel; Pozzi, Giorgia

DOI
10.1007/s11019-024-10235-6

Publication date
2024

Document Version
Final published version

Published in
Medicine, Health Care and Philosophy

Medicine, Health Care and Philosophy
https://doi.org/10.1007/s11019-024-10235-6

SCIENTIFIC CONTRIBUTION

Why we should talk about institutional (dis)trustworthiness and medical machine learning

Michiel De Proost[1] · **Giorgia Pozzi**[2]

Accepted: 30 October 2024

Abstract
The principle of trust has been placed at the centre as an attitude for engaging with clinical machine learning systems. However, the notions of trust and distrust remain fiercely debated in the philosophical and ethical literature. In this article, we proceed on a structural level ex negativo as we aim to analyse the concept of "institutional distrustworthiness" to achieve a proper diagnosis of how we should *not* engage with medical machine learning. First, we begin with several examples that hint at the emergence of a climate of distrust in the context of medical machine learning. Second, we introduce the concept of institutional trustworthiness based on an expansion of Hawley's commitment account. Third, we argue that institutional opacity can undermine the trustworthiness of medical institutions and can lead to new forms of testimonial injustices. Finally, we focus on possible building blocks for repairing institutional distrustworthiness.

Keywords Trust · Institutional distrustworthiness · Institutional opacity · Medical machine learning · Epistemic injustice · AI ethics

Introduction

According to testimonials, by the mid-1980s, there were significant concerns about the prevalent deterioration of trust in the clinical relationship (Sherlock 1986). Pellegrino and Thomasma (1993; p. 65) mention "an ethics of distrust [that] has been gathering force" based on the rise of participatory democracy and the neutralisation of traditional paternalism. A positive kind of distrust was installed to overcome the doctor-knows-best paradigm. In relation to new medical technologies that have the potential to disrupt moral relationships (Baker 2013), one can notice a more ambiguous climate of distrust. Currently, there is a specific area of technological advancements in which the relationship with trust is fraught with tension: artificial intelligence (AI) systems. Particularly, machine learning (ML) systems, a subcategory of AI systems, seem to displace physicians from their authoritative position, and communication difficulties may emerge because they may deprive human agents of central epistemic goods, such as understanding (Burrell 2016). Medical ML seems to upend the patterns of our business-as-usual lives and disrupt trust relationships, which are usually invisible and mediate our daily interactions. As a consequence, increasing initiatives are installed to promote trust in ML technologies that are framed as a crucial step towards patient empowerment (Segers and Mertes 2022).

Over the past few years, there has been a significant focus on medical ML within ethical literature, with the concepts of trust and trustworthiness frequently being highlighted in these debates. For instance, the EU Commission's High-Level Expert Group on AI, whose 2019 *Ethics Guidelines for Trustworthy AI* set the stage for this ever-growing debate. In creating accounts of trust in medical AI, inspiration is drawn from the normative frameworks presented by Baier (1986) and Hawley (2014), which are influential contributions to the standard philosophical discussions on trust. Nickel (2022), for instance, develops a normative account of trust based on the concept of "discretionary authority" to explain the interconnections between the expectations of users, the invitation of trust within user interfaces, and the commitments of AI practitioners. In contrast and adopting a

✉ Michiel De Proost
michiel.deproost@ugent.be

[1] Bioethics Institute Ghent, Department of Philosophy and Moral Sciences, Ghent University, Blandijnberg 2, Ghent 9000, Belgium

[2] Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands

Published online: 13 November 2024

sceptical attitude toward the possibility of trusting AI systems, Hatherley (2020; p. 480) claims that "AI systems lack the right kind of motivation for trust—either in the form of encapsulated interest or a sense of goodwill—since they lack motivation entirely". Despite the large attention in academic circles on whether it is possible to directly trust medical AI, fundamental disagreements remain on the foundations of trust. Moreover, the role medical ML plays in mediating trust relationships in an institutional context has been largely neglected.

Skepticism may arise as to whether the trust discourse that currently holds sway in the realm of machine learning is anything more than a mere nod to the importance of trust. Some authors are now talking about "ethics washing" to describe the pervasiveness of trust talk (Freiman 2023). Others wonder whether it would be better to regard this relationship as one of reliance rather than trust (Holland et al. 2022). In response to the dominance of the trust rhetoric, such heightened scrutiny of concepts is desirable, as we do not want ethical debates about the acceptability of innovative medical technologies to be held in terms of empty labels. Moreover, features of healthcare institutions and medical ML have been described that scaffold testimonial and hermeneutical injustices (Pozzi 2023). According to Medina (2020), epistemic injustices deepen the erosion of trust and perpetuate dysfunctional patterns of trust. This lack of trust may not entail only epistemic and ethical mistreatment but could be facilitated by political mistreatment (Medina 2013). Although some scholars point to occurrences of distrust in the medical ML literature (Braun et al. 2021; Freiman 2023; Laux 2023; Starke and Ienca 2022; Wolkenstein 2024), its political dimension is given less frequently philosophical centre stage.

In this article, we proceed on a structural level ex negativo as we aim to analyse the concept of "institutional distrustworthiness" to achieve a proper diagnosis of how we should *not* engage with medical machine learning. The paper proceeds as follows. In section one, we start with several examples that hint at the emergence of distrust in the context of medical ML. In section two, we introduce the distrust theory of Hawley (2014, 2017). In section three, we show how it can be fruitful to expand Hawley's commitment account, going beyond the interpersonal level of trust relationships to account for trust and distrust pertaining to institutions. In section four, we argue that institutional opacity can undermine the trustworthiness of medical institutions and how new testimonial injustices can occur. In the final section, we focus on repairing institutional distrustworthiness in medical ML and, in turn, potential pathways for building trust.

Before starting with our argumentation, let us make a brief linguistic clarification. The philosophical debate on trust focuses on the distinction of two concepts that are often used interchangeably in everyday language: trust and reliance. These concepts are often used as synonyms, so I might happen to say that I trust my computer or that I rely on my general practitioner. However, in the philosophical literature on trust, this concept is mostly used in the context of interpersonal interactions. In contrast, reliability is generally used to refer to our relationship with inanimate objects. The issue at the heart of the debate is thus to evaluate whether AI systems, as inanimate entities, can be trusted in a morally relevant sense that goes beyond mere reliance (Zanotti et al. 2023). It is important to clarify that it is not our aim to argue "for" or "against" trust and/or reliance in medical ML. We agree that something is at stake in ethical terms when the role of trust is debated in the praxis of medicine. Let us also point out that we leave in the middle the question of whether ML systems could possess human-like attributes such as motivations, will, and moral obligations that are usually seen as central to interpersonal trust relationships. We maintain that taking a stand on these disputed issues is not necessary to advance an ethical analysis pertaining to the institutionalised distrustworthiness of medical ML.

A growing climate of distrust?

In several cases of so-called medical ML, many people indicate a sceptical attitude of distrust. A study published by Obermeyer et al. (2019) in *Science* showed that an algorithm widely used in US hospitals to allocate healthcare resources to patients with complex health conditions had a considerably lower rate of referrals for Black patients compared to white patients. When possible assumptions were examined that could explain the rampant racism in decision-making software, the scientists speculated "that this reduced access to care is due to the effects of systemic racism, ranging from distrust of the healthcare system to direct racial discrimination by healthcare providers" (Ledford 2019). According to Benjamin (2019), the context of structural and interpersonal racism in healthcare cannot be overlooked: as the author argues, "a 'lack of trust' on the part of Black patients is not the issue; instead, it is a lack of trustworthiness on the part of the medical industry". Likewise, Graham (2022; p. 198) describes another study of Obermeyer on a new ML system for more objective pain measures and how "such a system seems to replicate long-standing patterns of clinical distrust of Black pain".

A similar observation is made around the use of the NarxCare algorithms that are supposed to deliver an accurate estimation of the likelihood of opioid misuse: "The problem that really infuses the NarxCare discussion is that the environment in which it is being used has an intense element of

law enforcement, fear, and *distrust* of patients" (Szalavitz 2021; emphasis added). Robertson et al. (2023) find variations in preferences. Specifically, Black respondents were less likely to choose diagnostic AI systems. Finally, we could speculate that recent findings on the underrepresentation of certain groups in training data can create or sustain institutional distrust. The reason for this is that, arguably, the use of biased algorithms in medical care has a bearing on the (dis)trust patients attribute to the medical institution that integrates them (and thus justifies their use) into medical practice. Sadly, instances of algorithmic bias abound in the literature. For instance, algorithms for skin cancer detection might have worked in the testing phase. However, as they were implemented, they bore the harmful potential of producing discriminatory patterns (Davis 2021).

Against this background, the following question might emerge: Why don't Black people (and other marginalised groups) trust ML in the medical context? Implicit in the question, as Wilson argues (2022), is a pathologising of people—the idea that there is something wrong with them about their inability to properly trust rather than with the conditions within which they exist and make an attitude of trust unjustified. This sense of wrongness not only further disadvantages them but also ignores the role that healthcare institutions play in fuelling climates of distrust. In a similar vein, Newman (2022) argues that implicitly focusing on the mistrust of marginalised populations toward certain institutions is a corrective attitude to change this "deviant" behaviour and align it to the standard attitude of trust recognisable among privileged social groups. According to the author, "[m]istrust places the scrutiny on the mistrustful, instead of focusing on the provider, medical institution, or health care system that fails to provide a context within which a patient can be empowered and feel comfortable in making a decision" (Newman 2022; p. 271).

We endorse Newman's conclusion that a necessary step to overcome a situation, in which the experience of privileged white people is the implicit norm against which the experiences of socially disadvantaged groups are measured and evaluated, is to decentre the analytic lens from privileged populations as a group of reference. However, we maintain that there is a need to understand not only how mistrust as an attitude of people towards institutions emerges but also how distrustworthiness as a *property* of institutions themselves and as a whole manifests (see section on institutionalised opacity). Our particular focus will be on the role of ML systems in medicine in fostering the (dis)trustworthiness of healthcare institutions.[1]

Generally, discussions about issues of distrust in medical ML start from the assumption that the problem takes the following form: people do not trust trustworthy actors. Based on this perspective, interventions aimed at improving transparency are generally targeted to encourage an attitude of trust from the side of trustors. However, such arguments overlook social factors that influence individual decisions to trust in clinical contexts. To better grasp the complexity of these issues, we need an account of institutional distrustworthiness. We provide this in the following sections.

Trust and distrust as commitment: the need to move beyond interpersonal relations

Filling in the details of trust is complicated as scholars disagree on the nature of the concept. Most authors agree that reliance is a basic component of trust but that some extra element is needed in addition (Hawley 2014). This is intuitively the case when we think of which reactions are usually in place when our reliance is not upheld compared to when trust, understood in a morally rich sense, is breached. If someone relies on the proper functioning of their dishwasher, one can be disappointed if the device suddenly stops working, but one does surely not feel betrayed by it. Differently, when someone trusts a trustee to, say, take care of their pets while they are on holiday and the trustee does not uphold this trust relationship, a feeling of betrayal and the demand for an apology would be suitable responses. That is to say, a breach of trust brings about morally loaded reactive attitudes that a failed reliance does not.

More problematic is the specification of this extra element that characterises proper trust in contrast to mere reliance. An overview of the philosophical literature on trust exceeds the scope of this paper,[2] for our discussion aims explicitly at targeting a working notion of institutional (dis)trustworthiness. In this paper, we thus limit our focus to an account of trust that is, we maintain, functional in tackling institutional distrustworthiness, i.e., the *commitment account* advanced by Hawley (2014, 2017). Let us first briefly reconstruct the main characteristics of Hawley's account in its original formulation as an account of interpersonal trust.

Hawley understands trust in terms of both commitment and motive.[3] When trusting X to perform a specific task T,

[1] Of course, our analysis of institutional distrustworthiness cannot be decoupled from understanding how (dis)trust mechanisms arise since these two concepts are necessarily intertwined.

[2] The philosophical literature on trust is vast but McLeod (2015) provides an overview. Durán and Pozzi (under review) also offer a review of the literature on trustworthy AI tailored to the analytic distinction between reliance and some "extra factor" largely adopted in the standard philosophical literature on trust.

[3] It is worth noting that Hawley's account is different from motive-based accounts of trust in that the motivation of the trustee to uphold the trust relationship is not based on goodwill (Jones 1996) or the fact that they want to maintain or strengthen their relationship to the trustor (Hardin 2002). According to Hawley's account the motivation of the

we do not simply assume that X will perform T (this would amount to mere reliance). Rather, we take X to have a commitment and to be motivated by that commitment in ways that make her worthy of our trust. Naturally, Hawley's account implies the trustee's awareness that a commitment is in place. The notion of commitment in Hawley's account is broad enough to accommodate explicit commitments, such as promises, and implicit ones, such as commitments that go along with certain roles (e.g., the commitment of a parent to care for their child) or emerge in connection with shared social conventions (Hawley 2014). The sense of commitment that Hawley invokes is therefore not necessarily psychological – lacking a certain intention will not eliminate the commitment.

Just as trust, also distrust is normatively relevant so that mere non-reliance does not automatically amount to distrust. Misplaced distrust, i.e., distrusting someone who is actually worthy of our trust, brings about different reactive attitudes compared to a situation in which one mistakenly fails to rely on someone or something. If someone distrusts their friend to keep their promise and it later turns out that the friend was to be trusted, it seems suitable to feel remorse and offer the friend an apology. Differently, if we fail to rely on an instrument that turns out to be reliable after all, remorse would be out of place. This is the case because, similarly to cases of trust, distrust entails a normative and morally loaded dimension. In Hawley's formulation, "[t]o distrust someone to do something is to believe that she has a commitment to doing it, and yet not rely upon her to meet that commitment" (Hawley 2014; p. 10). Distrust thus entails a "moral criticism" that is not in place when we simply have low expectations of the trustee (Hawley 2017; pp. 70–71).

The emphasis on commitment seems thus attractive when we attempt to extend the picture to include an account of distrust. In fact, the commitment account provides us with ample guidance on how to place our (dis)trust (Hawley 2014). This is because, according to Hawley, maintaining trust or legitimately distrust depends on the ability to uphold others' normative expectations exclusively when these are grounded in the commitment of a trustee toward the trustor. Crucially, explicitly relating normative expectations to commitments excludes those expectations that may be irrational or largely inappropriate, that is to say, those that might not be pertinent to either trust or distrust. To make this point more graspable consider this hypothetical case discussed by Hawley (2014). Assume that a colleague regularly offers me the leftovers of her lunch not because I explicitly entrusted her to provide for my lunch (and she agreed) or because she promised she would share it with me. Rather, this happens simply because she is unable to properly quantify how much food she will actually need for lunch. If this happens regularly, I can get to rely on her to share her food with me. However, if I developed a normative expectation that she *should* bring food for me to the office and she failed to do so, an attitude of distrust from my side would be largely inappropriate. More precisely, in this case, neither trust nor distrust seem to be suitable attitudes. Cases similar to this elucidate a crucial strength of the commitment account: it provides a solid theoretical basis to distinguish between appropriate and inappropriate normative expectations precisely by making reference to the nature of the commitment that underlies a relationship of trust. So, the goal of Hawley's commitment account is not only to provide the theoretical tools to identify situations in which trust or distrust are suitable but to indicate when neither is suitable. Under this heading, distrust amounts to non-reliance and, in addition to this, the belief that the trustee has a commitment to uphold. Respectively, trust amounts to reliance plus the belief that the trustee has a commitment. In both instances, trust or distrust are appropriate attitudes only when a belief in commitment is in place.

For our purposes in this paper, it is important to point out that Hawley restricts her commitment account of trust and distrust to interpersonal relationships. In fact, the author is in favour of abandoning the trust–reliance distinction in collective contexts, thus suggesting that institutional (dis) trustworthiness and institutional (non) reliability should be treated synonymously (Hawley 2017; p. 4). According to Hawley, "we need to address structural problems and collective contexts if we are to combat injustice and create better institutions. However, we do not need the distinction between trustworthiness and mere reliability at the group level in order to pursue these projects" (Hawley 2017; p. 245). Following Baier, Hawley conceptualises trust as directed at individuals and the interpersonal dimension because of the intimate connections between trust and the reactive attitudes emerging when trust is breached, as previously pointed out. However, it seems fruitful to take a closer look at the dimension of distrust at the institutional level through the lens of Hawley's commitment account (even if this is not directly envisaged by the author).

Towards an account of institutional distrustworthiness

Demir-Doğuoğlu and McLeod (2023) formulated a critique of Hawley's account, defending the view that interpersonal (dis)trust should not be seen as entirely separated from institutional (dis)trust. By criticising the commitment account from a feminist perspective, the authors conceptualise the institutional distrust that oppressed groups may encounter.

trustee to fulfil the trust relation comes from the commitment itself (Hawley 2014).

Let us reconstruct both points of critique in turn in order to show how a slightly modified view of the commitment account can be useful to make sense of the distrust (and not mere non-reliance) experienced by members of disadvantaged social groups in collective contexts. This notion of institutional distrust will provide us with ample guidance on how to better understand the considerations related to distrust in medical ML.

The first critique advanced by Demir-Doğuoğlu and McLeod hinges on Hawley's definition of distrust as non-reliance plus belief in commitment, thus presupposing that proper distrust requires an attitude of non-reliance to be in place. So, for example, if someone does not rely on their neighbour to water their plants while they are gone amounts to distrust only if the neighbour made a commitment to, in fact, take care of the plants. If the person does not rely on the neighbour *despite* their commitment (because, say, the neighbour already forgot to uphold this commitment in the past), then proper distrust is in place. However, as Demir-Doğuoğlu and McLeod point out, there are cases in which reliance and distrust coexist and that are not taken into consideration by Hawley's account. This holds particularly true for members of disadvantaged social groups who happen to find themselves in situations of reliance on people or institutions that they (have good reasons to) distrust.

For example, it is a well-known and empirically grounded fact that Black people in the US often experience worse health treatment due to racial biases compared to other population groups (Curry 2020). So, one could say that an attitude of distrust in physicians or even in the healthcare system as a whole can emerge due to disparities ingrained in how the delivery of healthcare is experienced by members belonging to disadvantaged social groups. However, the distrust in one's physician does not necessarily mean that one can decide not to rely on them when in need of health support. In cases similar to these, one finds oneself in a situation of *having to* rely on someone (or, more generally, a social institution) that one distrusts, which shows that reliance and distrust can, in fact, coexist. An essential point that Hawley's account misses from the picture is that not relying on whom we distrust is an exercise of social power because it implies the possibility of *deciding* on whom to rely. This possibility often remains precluded to members of disadvantaged social groups due to systemic inequalities and pervading social injustices. That is to say, the possibility to avoid relying on whom one distrusts does not exclude proper distrust only because, out of conditions of practical necessity, one is bound to rely on distrusted people and institutions. To our mind, this fundamental critique advanced by Demir-Doğuoğlu and McLeod to Hawley's commitment account expresses a condition constitutive of institutional distrustworthiness. Let us dub this the *distrust-despite-reliance* condition.

Let us now turn to the second critique. This amounts to the fact that an absence of belief in commitment does not make the experienced distrust of oppressed people inappropriate. There are many examples in which members of oppressed groups feel distrust in public institutions and find the commitments these have made unbelievable. Demir-Doğuoğlu and McLeod point out the distrust that Black people in the US can have with respect to the veracity of the commitment of the police to racial equality in view of many cases of police violence and brutality targeted at people of color. That is to say, one could doubt that the commitment of the police only amounts to window dressing, i.e., doubt that it is believable. As Specker Sullivan (2023; p. S36) points out, "our assumptions about what we can expect from others and whether we can believe what they tell us influences our decisions to accept vulnerability to them and depend on them." If the experiences one makes in medical encounters do not show an explicit commitment to a good delivery of health support, then one is apt to decide to withhold trust and reduce dependency to a minimum (other than in cases of practical necessity).

In the face of these considerations, Demir-Doğuoğlu and McLeod suggest adding a "believability condition" to the commitment account. To illustrate what Demir-Doğuoğlu and McLeod mean by a believable commitment they refer to Hawley's own example of a friend making a promise to attend a birthday party. A person may find this promise unbelievable if she knows that her friend is overwhelmed by work or caring obligations. All these possible reasons, which influence whether or not the person can and will keep her promise, may create the impression that the promise is unbelievable.

However, we think there is also a further, albeit related, interpretation that could complement the commitment account for cases similar to the one described and that is further constitutive of institutional distrustworthiness. One could justifiably distrust not only in the case that a commitment is not believable but also if there is, in fact, no (perceived) commitment and one has good reasons to believe that there *should be* such commitment. Let us call this the *absence-of-commitment* condition. For example, if someone, as a member of an oppressed group, has reasons to believe that the healthcare system does not commit to caring for their health situation appropriately, this also amounts to proper distrust (even if they have no choice but to rely on the system for healthcare delivery). This is the case because, implicitly, we take that the healthcare system should have and clearly endorse said commitment in a way that emerges in medical encounters. So, on occasion, exactly due to the

absence of a commitment, one is justified to distrust (instead of merely not relying on) social institutions.

What has been said so far and in particular the two conditions for institutional distrustworthiness we spelled out (i.e., the *distrust-despite-reliance* and the *absence-of-commitment* conditions) supports two points of great relevance for the paper's overall goal. First, a revised version of Hawley's commitment account can allow us to make sense of distrust, even at the institutional level. Second, it is, contra Hawley, normatively relevant to distinguish between non-reliance and distrust as well as beyond the boundaries of interpersonal relations to make sense of structural injustices and issues pertaining to the unequal distribution of social power. In support of the second point mentioned, Fricker (2023; p. 8) suggests that the synonymous relationship between the trustworthiness of individuals and commitment should be resisted. Some institutions can be more reliable and have commitments and obligations, and when they do, the register of trust is in order. The author argues that "we do need to theorize trustworthiness in organizations, for there are some institutional bodies and processes whose dysfunctionality and moral status we can only fully understand if we pay attention to *ethos*, and the potential betrayal of individuals and groups that depend on them" (Fricker 2023; p. 741). In other words, an institution could have joint commitments explicitly made to values that comprise their ethos, or have commitments to joint decisions, actions, policies, and processes that embody those values. As Walker (2006; p. 84) confirms, it seems that what we often trust is not an individual person but "the reliable good order and safety of an environment." Above and beyond this theoretical perspective, there is plenty of empirical support for the view that people do trust (or distrust) organisations (Holland et al. 2022).

According to Fricker, there are three main reasons to make institutional trustworthiness a distinct value. First of all, to speak of institutional trustworthiness has a distinctive *functional* value: "insofar as institutional bodies act on commitment-based reasons that involve a responsiveness to our dependence on them, our display of trust in them (perhaps just by showing up and asking for a service) is a way of enlisting institutional agency to help us make things happen" (Fricker 2023; p. 736). Secondly, having some institutions that generally act reliably for commitment-based reasons has a special ethical-political value as citizens have many standing dependencies on the procedures of institutional bodies. For instance, our dependency on public transport networks makes it valuable that such institutions should be responsive to our dependency. Finally, accounting for institutional trustworthiness allows us to diagnose institutional *dis*trustworthiness where it may occur. Distinct features of distrust could be displayed in spades, such as being driven by a faulty ethos or relations of dependence that create betrayal and call for accountability of the institution itself instead of individual actors. The advantage of modelling a sort of stabilised motivational set in an institutional body is thus that it calls upon the mechanics of collective agency and responsibility. As Davidson and Satta indicate (2021; p. 23) "we have a collective responsibility to become more trustworthy instead of focusing on changing the minds of those who exhibit justified social distrust".

Institutionalised opacity

After explaining what we mean by institutional distrustworthiness and why it is relevant in the medical realm in the previous section, we will now argue that institutional opacity can undermine the trustworthiness of medical institutions deploying ML systems. The different forms of opacity that medical ML can take, including epistemic, methodological, and semantic, have been widely discussed in the literature (Bjerrin and Busch 2021; Burrell 2016; Creel 2020; Durán and Jongsma 2021). These discussions suggest that there are considerable concerns about opacity and accountability pertaining to medical ML in clinical decision-making (Smith 2021). Nonetheless, such debates mostly focus on technical or individual concerns and take restrictions of opacity as a dyadic arrangement of human to machine. As a consequence, the social structure within which healthcare decisions are made is neglected. So, while such dyadic considerations of opacity are important, we maintain that this framework is not broad enough to account for the significance of the power of social structures and institutional frameworks, and how these contribute to shaping people's decision-making process (Ho 2008).

Carel and Kidd (2021; p. 481) describe the concept of "institutional opacity" as "a general tendency within large-scale and internally complex institutions to increasingly become resistant to forms of assessment and understanding". Although some degree of opacity is unavoidable in medical institutions because such institutions are large and often hierarchical and compartmentalised, these features are further complicated by periodic restructurings that involve changes to the redefinition of roles and other practical or structural changes. The introduction of medical ML has the potential to alter relationships in the clinical environment, adding a further layer of opacity that cumulates with other forms of opacity as many new parties are involved in the design, procurement, and use of medical ML. One can, of course, ask the question of what makes an opacity connected to ML at the institutional level different from the one pertaining to standard medical practice.

To illustrate this difference, consider the example described by Fricker (2023; p. 736): "We show up at A&E

with a broken leg, and things happen; care is delivered, as best they can under stretched circumstances, because the Accident and Emergency service is committed to giving appropriate care, and because its staff respond to our manifest, acute dependence on them. If the service is trustworthy-in-general, then it reliably acts on precisely this sort of commitment-based, and dependence-responsive, motive." However, there may be more complex situations in which the cause of a certain health issue is not so easily accessible for healthcare providers, as it happens in many cases for patients suffering from chronic syndromes or psychosomatic diseases. Under this heading, what happens when a patient shows up at a hospital not with a broken bone (i.e., an objectively recognisable health issue) but with a condition where credibility questions are at stake? In these cases, the testimonial offerings of patients play a crucial role in allowing healthcare professionals to grasp the nature of the condition and take possible remedial actions.

Introducing ML as an authoritative epistemic entity in clinical care can change the role of physicians, particularly in cases where patients' credibility needs to be assessed (such as in pain management). Medical ML solutions are often framed as cost-saving measures. However, these characteristics could potentially lead to institutions becoming more opaque, particularly if there is variation in rules and procedures across different areas of the institution and if these are influenced by different biases and other epistemic issues such as lack of information on who is responsible for which task. In an opaque institution, determining the appropriate testimonial offerings and their role in informing medical decisions can be challenging. It becomes unclear what statements could have an impact, what inquiries would yield essential information, and what suggestions would align with the procedures that remain inaccessible to (often) non-expert patients.

If we look at the NarxCare case that is discussed in section one, it has been pointed out that ML systems used to predict patients' risk of misusing opioids are often considered, by default, more credible than patients' testimony (Pozzi 2023). Likewise, Graham (2022; p. 152) indicates that technologies, such as ML systems, "infer pain from physiological processes risk valorising expert assessment over patient report". This hinders the epistemic participation of patients in medical institutions and leads physicians to neglect or unjustifiably dismiss their contributions. In these situations, the mediating role of ML systems can exacerbate institutional opacity. In fact, critical decisions affected by ML systems risk no longer being properly explained to the patient or are accompanied by explanations delivered in a haze of jargon, thus hampering genuine understanding. Intuitively, these opaque mechanisms can fuel an attitude of distrust from the side of patients toward the medical institution as a whole. If medical delivery is compromised, as in the NarxCare case, patients justifiably develop a sceptical attitude, doubting the commitment of medical institutions to provide suitable health support for all. Medical situations mediated by ML similar to the one just described thus show that a central condition for institutional distrustworthiness previously discussed is in place. This is the case because these categories of patients have (justifiably) reasons to believe that there is an *absence of commitment* from the side of the medical institution using the ML system to deliver just medical care.

Let us point out that the problematic kind of opacity we aim to tackle pertains not only to the technical features of the particular ML system in question (i.e., its black box nature). Rather, the opacity we want to problematise is more encompassing and difficult to counteract. As Carel and Kidd (2021; p. 485) describe "bureaucracy, complexity, hierarchy, jargon, negative stereotyping", and we would add to this list opaque ML systems, "can together obstruct the practical and epistemic agency of persons". Since there is no effective way for patients to clear their records or understand the relevant factors that got them ranked as being at a high risk of drug misuse in the NarxCare case, it is factually impossible for them to seek redress, critically question their situation, and receive explanations from healthcare providers. Arguably, this creates a situation of epistemic vulnerability that can deflate patients' epistemic confidence and limit their epistemic agency, thus creating a further disadvantage for them (Carel and Kidd 2021). The justified distrust that emerges from what becomes an opaque medical institution can lead patients to rely on it and the ML systems it deploys when seeking medical support in spite of the fact that they (have reasons to) distrust it. As previously discussed, this can be the case due to conditions of practical necessity rather than the belief that the medical institution is worthy of being trusted. Thereby, also the other previously identified condition for institutional distrustworthiness (i.e., reliance-despite-distrust) is satisfied, also due to institutionalised opacity.

Yet, the growing institutional opacity should not lead to fatalism and present the ascription of responsibility as no longer possible – this is often referred to as the "many hands" problem (Coeckelbergh 2020). With the focus on institutional distrustworthiness, we can still highlight the need to hold institutions accountable for being unworthy of our trust. This enables the distribution of responsibility across a diverse network of both human and artificial agents. Humans are still running institutions that apply ML systems, which are complex but still embedded in institutional contexts that call for accountability. Fricker provides another example of a hospital that is held accountable for its caring services: "Just as an individual doctor, in acting on

her commitment to give proper care to a patient, is thereby displaying a responsiveness to the patient's dependence on her, so is a hospital that provides care for its patients for commitment-based reasons displaying a responsiveness to their dependence on *it*" (Fricker 2023; p. 736). The same logic can be applied to hospitals that use ML tools. We can and should direct our natural reactive attitudes of accountability, such as a sense of betrayal and feelings of resentment, towards the organisation itself rather than towards the individuals acting under the auspices of the institution.

The first building blocks for repairing institutional distrustworthiness by addressing epistemic justice

So far, we have offered an analysis of institutional distrustworthiness (based on two conditions, i.e., the *distrust-despite-reliance* and the *absence-of-commitment* conditions) and arising, among others, from institutional opacity. Against this background, it is an open question whether we should try to "fix" possible distrust in medical ML. The value of distrust is itself morally ambivalent and shaped by the positionality of people: it can be functional to resisting oppression just as it can enable it. Depending on a person's social situatedness, an attitude of distrust can point to a justifiably critical position that refuses to accept what dominant social groups might impose as a universal truth. On the other hand, for an agent in a position of social power, illegitimately distrusting another agent due to stigma and prejudices connected to their social identity can perpetuate oppressive patterns and systemic injustices. Concerning the first case mentioned, Demir-Doğuoğlu and McLeod (2023; p. 137) argue that "institutional distrust *itself* can have positive effects for members of oppressed groups, as the attitude fundamentally aims to protect them from further institutional harm and violence". Likewise, Krishnamurthy (2015) has argued that distrust can be a strategic tool to safeguard oppressed communities against tyranny. Under this heading, vigilance and wariness about patients and healthcare entities may, in fact, be healthy responses to the history of racism and discrimination in medicine and healthcare.

With these positive connotations of distrust in mind, we can identify some strategies that public health institutions can adopt to ameliorate the two conditions of distrustworthiness. In this way, we avoid addressing an attitude people can (justifiably) hold but rather a central property of relevant institutions. A crucial step – the minimal condition that must be met to begin with – is to acknowledge and take seriously existing relationships of distrust that emerge from different forms of epistemic injustice that are overly experienced by members of socially disadvantaged groups and move from this location. Rectifying injustice and, consequently, addressing issues that determine and justify attitudes of distrust requires acknowledgement, specifically acknowledging one's actions as redress for one's wrongdoing (Walker 2006). An acknowledgment of one's responsibility for perpetrating epistemic injustice does not complete the process of epistemic repair nor, on its own, license renewed trust. It is an important step, nonetheless, without which victims of unacknowledged epistemic injustice are otherwise deprived of the considerable practical and epistemic benefits enabled by functional trust relationships.

Other ways of encouraging the creation and maintenance of an institutional ethos of testimonial justice amount to "tailoring institutional norms and values" and "cultivating institutional appreciation of human diversity" (Carel and Kidd 2021). The first strategy aims to broaden institutional conceptions of flourishing and care and calls upon "institutions to tailor and relativise their norms to individual people, seen within their context" (Carel and Kidd 2021; p. 489). It is crucial that institutions acknowledge a flexible and inclusive use of medical ML and reflect on the individual patients' needs, desires, and values. Medical institutions have the tendency, often for good reasons of cost efficiency, to standardise and thus assume a certain degree of uniformity throughout patients' groups. However, this needs to be balanced against other values like accessibility, flexibility, and inclusion.

Notably, the ability to interact with medical ML assumes that some resources, broadly conceived, are already within the reach of persons. If a person receives, for instance, an ML-mediated result from the doctor as an economically privileged, health-literate, tech-savvy person, one can contextualize this information and know promptly how to act on it. If a person had fewer privileges and resources, the interaction might have made them feel powerless. In other words, subjects have different starting points from which the institutional world - especially when it is mediated by ML systems - unfolds. This points to the need to embrace the diversity of individuals and groups who interact with the institution or are served by it, which in turn requires understanding the importance of diversity to an institutional ethos of testimonial justice. The explicit manifestation of such an ethos would amount, in its most effective form, to concrete practices that recognize and support the value of what individuals and their needs can actively contribute to more inclusive policies. This involves an authentic appreciation of the different types of epistemic access needs, as well as individuals who fit within the broader moral landscape of the institution.

Determining this alignment necessitates at least two components: users must comprehend the organisation's structure, and they need to be able to speak effectively about what their potential roles within it might be. One can imagine how persons with disabilities do not find spaces, for instance, within institutions corrupted by ableist prejudices. Likewise, one could argue that algorithms can be designed that are discrimination-aware and

embrace the diversity of access needs (Cirillo et al. 2020). However, these approaches —such as many approaches towards algorithmic fairness available in the literature — are often too restrictive as they advance a decontextualised analysis of the "algorithm itself" (Hull 2023). We maintain that without efforts to tackle systemic injustices such as institutional racism, these initiatives are unlikely to succeed, and inequalities remain. Therefore, medical ML might reiterate the current status quo in healthcare, where very few dominant groups are privileged to the detriment of others.

Conclusion

The trust rhetoric in medical ML can be a double-edged sword. On the one hand, it can play a significant role in bridging the gap that exists between trustors and trustees by increasing transparency and overall quality of healthcare provision. However, on the other hand, it can also exacerbate and perpetuate existing epistemic injustices by ignoring existing patterns of distrust. While the literature has started to highlight the importance of integrating distrust considerations in medical ML, there is still much to explore about its *structural* implications, which we elaborated on in this paper.

Given the risks of medical ML when it comes to reproducing and exacerbating existing epistemic injustices, we offered a new perspective in the medical ML debate by introducing the concept of "institutional distrustworthiness". We suggested that there is a need to understand not only how mistrust as an attitude of people towards the use of medical ML emerges but also how distrustworthiness, as a property of institutions themselves and as a whole, manifests. Therefore, we developed an account of institutional trustworthiness based on the work of Hawley and Fricker. We further argued that institutional opacity can undermine the trustworthiness of medical institutions and how new testimonial injustices can occur with the use of medical ML. We concluded by offering some, albeit initial, ways of addressing this potent juncture of injustice and focused on repairing and ameliorating institutional distrustworthiness.

While this paper is limited to providing a broad conceptual sketch, it offers a relevant starting point for further discussions on how the ethos of institutions and attitudes of distrust are taking shape in specific medical ML practices. We raised awareness of the magnitude of this problem and only scratched the surface of how to ameliorate institutional opacity and distrustworthiness. However, we encourage further research working towards dismantling unjust structural mechanisms instead of rushing the deployment of medical ML in institutions that do not account for diversity, risking providing insufficient healthcare support for already disadvantaged populations.

Acknowledgements MDP has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement no. 949841-DIME). GP's contribution to this work was supported by the European Commission through the H2020-INFRAIA-2018-2020/H2020-INFRAIA-2019-1 European project "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (Grant Agreement no. 871042).

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Baier, A. 1986. Trust and antitrust. *Ethics* 96(2): 231–260.

Baker, R. 2013. *Before bioethics: A history of American medical ethics from the colonial period to the bioethics revolution*. Oxford: Oxford University Press.

Benjamin, R. 2019. Assessing risk, automating racism. *Science* 366(6464): 421–422.

Bjerring, J. C., and J. Busch. 2021. Artificial intelligence and patient-centered decision-making. *Philosophy & Technology* 34: 349–371.

Braun, M., H. Bleher, and P. Hummel. 2021. A leap of faith: Is there a formula for trustworthy AI? *Hastings Center Report* 51(3): 17–22.

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1). https://doi.org/10.1177/2053951715622512

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science* 87(4): 568–589. https://doi.org/10.1086/709729

Carel, H., and I. J. Kidd. 2021. Institutional opacity, epistemic vulnerability, and institutional testimonial justice. *International Journal of Philosophical Studies* 29(4): 473–496.

Cirillo, D., S. Catuara-Solarz, C. Morey, E. Guney, L. Subirats, S. Mellino, and N. Mavridis. 2020. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digital Medicine* 3(1): 1–11.

Coeckelbergh, M. 2020. Artificial Intelligence, responsibility attribution, and a relational justification of Explainability. *Science and Engineering Ethics* 26: 2051–2068.

Curry, T. J. 2020. Conditioned for death: Analysing black mortalities from Covid-19 and police killings in the United States as a syndemic interaction. *Comparative American Studies an International Journal* 17(3–4): 257–270.

Davidson, L. J., and M. Satta. 2021. Justified social distrust. In *Social Trust*, eds. Kevin Vallier and Michael Weber, 122–148. New York: Routledge.

Davis, N. 2021. AI skin cancer diagnoses risk being less accurate for dark skin – study. The Guardian. https://www.theguardian.com/society/2021/nov/09/ai-skin-cancer-diagnoses-risk-being-less-accurate-for-dark-skin-study. Accessed 17 March 2024.

Demir-Doğuoğlu, H., and C. McLeod. 2023. Toward a feminist theory of distrust. In *The moral psychology of trust*, eds. David Collins, Iris Vidmar Jovanović, and Mark Alfano, 125–143. London: Lexington Books.

Durán, J. M., and K. R. Jongsma. 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 47(5): 329–335.

Durán, J. M., and G. Pozzi. (under review). What is trustworthy AI?.

European Commission. 2019. Ethics guidelines for trustworthy AI. High-level expert group on artificial intelligence. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed: May 16, 2024.

Freiman, O. 2023. Making sense of the conceptual nonsense 'trustworthy AI'. *AI and Ethics* 3(4): 1351–1360.

Fricker, M. 2023. Diagnosing institutionalized 'Distrustworthiness'. *The Philosophical Quarterly* 73(3): 722–742.

Graham, S. S. 2022. *The doctor and the algorithm: Promise, peril, and the future of health AI*. Oxford: Oxford University Press.

Hardin, R. 2002. *Trust and trustworthiness*. New York: Russell Sage Foundation.

Hatherley, J. J. 2020. Limits of trust in medical AI. *Journal of Medical Ethics* 46(7): 478–481.

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

Hawley, K. J. 2017. Trust, distrust and epistemic injustice. In *The Routledge handbook of epistemic injustice, eds. Ian James Kidd, José Medina and Gaile Pohlhaus Jr*, 69–78. New York: Routledge.

Ho, A. 2008. The individualist model of autonomy and the challenge of disability. *Journal of Bioethical Inquiry* 5: 193–207.

Holland, S., J. Cawthra, T. Schloemer, and P. Schröder-Bäck. 2022. Trust and the acquisition and use of public health information. *Health Care Analysis* 30: 1–17.

Hull, G. 2023. Dirty data labeled dirt cheap: Epistemic injustice in machine learning systems. *Ethics and Information Technology* 25(3): 38.

Jones, Karen. 1996. Trust as an affective attitude. *Ethics* 107(1): 4–25.

Krishnamurthy, M. ed. 2015. (White) Tyranny and the democratic value of distrust. *The Monist* 98(4): 391–406.

Laux, J. 2023. Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the European Union AI act. *AI & Society*. https://doi.org/10.1007/s00146-023-01777-z

Ledford, H. 2019. Millions affected by racial bias in health-care algorithm. *Nature* 574(31): 2.

McLeod, C. 2015. Trust. The stanford encyclopedia of philosophy. https://plato.stanford.edu/archives/fall2015/entries/trust/. Accessed 16 May 2024.

Medina, J. 2013. *The epistemology of resistance: Gender and racial oppression, epistemic injustice, and the social imagination*. Oxford: Oxford University Press.

Medina, J. 2020. Trust and Epistemic Injustice. In *The Routledge handbook of trust and philosophy*, eds. Ian James Kidd, José Medina and Gaile Pohlhaus Jr, 52–63. New York: Routledge.

Newman, A. M. 2022. Moving beyond mistrust: Centering institutional change by decentering the white analytical lens. *Bioethics* 36(3): 267–273.

Nickel, P. J. 2022. Trust in medical artificial intelligence: A discretionary account. *Ethics and Information Technology* 24(1): 7.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464): 447–453.

Pellegrino, E. D., and D. C. Thomasma. 1993. *The virtues in medical practice*. New York: Oxford University Press.

Pozzi, G. 2023. Testimonial injustice in medical machine learning. *Journal of Medical Ethics* 49(8): 536–540.

Robertson, C., A. Woods, K. Bergstrand, J. Findley, C. Balser, and M. J. Slepian. 2023. Diverse patients' attitudes towards Artificial Intelligence (AI) in diagnosis. *PLOS Digital Health* 2(5): e0000237.

Segers, S., and H. Mertes. 2022. The curious case of trust in the light of changing doctor–patient relationships. *Bioethics* 36(8): 849–857.

Sherlock, R. 1986. Reasonable men and sick human beings. *The American Journal of Medicine* 80(1): 2–4.

Smith, H. 2021. Clinical AI: Opacity, accountability, responsibility and liability. *AI & Society* 36(2): 535–545.

Specker Sullivan, L. 2023. Climates of distrust in medicine. *Hastings Center Report* 53: S33–S38.

Starke, G., and M. Ienca. 2022. Misplaced trust and distrust: How not to engage with medical artificial intelligence. *Cambridge Quarterly of Healthcare Ethics*. https://doi.org/10.1017/S0963180122000445

Szalavitz, M. 2021. The pain was Unbearable. so why did doctors turn her away. *Wired*. https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/. Accessed 16 May 2024.

Walker, M. U. 2006. *Moral repair: Reconstructing moral relations after wrongdoing*. New York: Cambridge University Press.

Wilson, Y. 2022. Is Trust Enough? Anti-black racism and the perception of Black Vaccine Hesitancy. *Hastings Center Report* 52: S12–S17.

Wolkenstein, A. 2024. Healthy mistrust: Medical Black Box Algorithms, Epistemic Authority, and Preemptionism. *Cambridge Quarterly of Healthcare Ethics*. https://doi.org/10.1017/S0963180123000646

Zanotti, G., M. Petrolo, D. Chiffi, and V. Schiaffonati. 2023. Keep trusting! A plea for the notion of trustworthy AI. *AI & Society*. https://doi.org/10.1007/s00146-023-01789-9