

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

2

3 **Abstract**

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

13 **Keywords: trust, institutional distrustworthiness, institutional opacity, medical machine learning,**
14 **epistemic injustice, AI ethics**

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29 **Introduction**

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

43

44 Over the past few years, there has been a significant focus on medical ML within ethical literature, with the
45 concepts of trust and trustworthiness frequently being highlighted in these debates. For instance, the EU
46 Commission’s High-Level Expert Group on AI, whose 2019 *Ethics Guidelines for Trustworthy AI* set the stage
47 for this ever-growing debate. In creating accounts of trust in medical AI, inspiration is drawn from the normative
48 frameworks presented by Baier (1987) and Hawley (2014), which are influential contributions to the standard
49 philosophical discussions on trust. Nickel (2022), for instance, develops a normative account of trust based on the
50 concept of “discretionary authority” to explain the interconnections between the expectations of users, the
51 invitation of trust within user interfaces, and the commitments of AI practitioners. In contrast and adopting a
52 sceptical attitude toward the possibility of trusting AI systems, Hatherley (2020; p. 480) claims that “AI systems
53 lack the right kind of motivation for trust—either in the form of encapsulated interest or a sense of goodwill—
54 since they lack motivation entirely”. Despite the large attention in academic circles on whether it is possible to
55 directly trust medical AI, fundamental disagreements remain on the foundations of trust. Moreover, the role
56 medical ML plays in mediating trust relationships in an institutional context has been largely neglected.

57

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

70

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

80

81 Before starting with our argumentation, let us make a brief linguistic clarification. The philosophical debate on
82 trust focuses on the distinction of two concepts that are often used interchangeably in everyday language: trust
83 and reliance. These concepts are often used as synonyms, so I might happen to say that I trust my computer or
84 that I rely on my general practitioner. However, in the philosophical literature on trust, this concept is mostly used
85 in the context of interpersonal interactions. In contrast, reliability is generally used to refer to our relationship with
86 inanimate objects. The issue at the heart of the debate is thus to evaluate whether AI systems, as inanimate entities,
87 can be trusted in a morally relevant sense that goes beyond mere reliance (Zanotti et al. 2023). It is important to

88 clarify that it is not our aim to argue “for” or “against” trust and/or reliance in medical ML. We agree that
89 something is at stake in ethical terms when the role of trust is debated in the praxis of medicine. Let us also point
90 out that we leave in the middle the question of whether ML systems could possess human-like attributes such as
91 motivations, will, and moral obligations that are usually seen as central to interpersonal trust relationships. We
92 maintain that taking a stand on these disputed issues is not necessary to advance an ethical analysis pertaining to
93 the institutionalised distrustworthiness of medical ML.

94

95 **A growing climate of distrust?**

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

107

108 A similar observation is made around the use of the NarxCare algorithms that are supposed to deliver an accurate
109 estimation of the likelihood of opioid misuse: “The problem that really infuses the NarxCare discussion is that the
110 environment in which it is being used has an intense element of law enforcement, fear, and *distrust* of patients”
111 (Szalavitz 2021, emphasis added). Robertson et al. (2023) find variations in preferences. Specifically, Black
112 respondents were less likely to choose diagnostic AI systems. Finally, we could speculate that recent findings on
113 the underrepresentation of certain groups in training data can create or sustain institutional distrust. The reason
114 for this is that, arguably, the use of biased algorithms in medical care has a bearing on the (dis)trust patients
115 attribute to the medical institution that integrates them (and thus justifies their use) into medical practice. Sadly,
116 instances of algorithmic bias abound in the literature. For instance, algorithms for skin cancer detection might

117 have worked in the testing phase. However, as they were implemented, they bore the harmful potential of
118 producing discriminatory patterns (Davis 2021).

119

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

131

132 We endorse Newman’s conclusion that a necessary step to overcome a situation, in which the experience of
133 privileged white people is the implicit norm against which the experiences of socially disadvantaged groups are
134 measured and evaluated, is to decentre the analytic lens from privileged populations as a group of reference.
135 However, we maintain that there is a need to understand not only how mistrust as an attitude of people towards
136 institutions emerges but also how distrustworthiness as a *property* of institutions themselves and as a whole
137 manifests (see section 3 on institutionalised opacity). Our particular focus will be on the role of ML systems in
138 medicine in fostering the (dis)trustworthiness of healthcare institutions.¹

139

140 Generally, discussions about issues of distrust in medical ML start from the assumption that the problem takes the
141 following form: people do not trust trustworthy actors. Based on this perspective, interventions aimed at
142 improving transparency are generally targeted to encourage an attitude of trust from the side of trustors. However,
143 such arguments overlook social factors that influence individual decisions to trust in clinical contexts. To better

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.

144 grasp the complexity of these issues, we need an account of institutional distrustworthiness. We provide this in
145 the following sections.

146

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

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

157

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

164

165 Hawley understands trust in terms of both commitment and motive.³ When trusting X to perform a specific task
166 T, we do not simply assume that X will perform T (this would amount to mere reliance). Rather, we take X to
167 have a commitment and to be motivated by that commitment in ways that make her worthy of our trust. Naturally,
168 Hawley's account implies the trustee's awareness that a commitment is in place. The notion of commitment in
169 Hawley's account is broad enough to accommodate explicit commitments, such as promises, and implicit ones,

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

170 such as commitments that go along with certain roles (e.g., the commitment of a parent to care for their child) or
171 emerge in connection with shared social conventions (Hawley 2014). The sense of commitment that Hawley
172 invokes is therefore not necessarily psychological – lacking a certain intention will not eliminate the commitment.

173

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

184

185 The emphasis on commitment seems thus attractive when we attempt to extend the picture to include an account
186 of distrust. In fact, the commitment account provides us with ample guidance on how to place our (dis)trust
187 (Hawley 2014). This is because, according to Hawley, maintaining trust or legitimately distrust depends on the
188 ability to uphold others’ normative expectations exclusively when these are grounded in the commitment of a
189 trustee toward the trustor. Crucially, explicitly relating normative expectations to commitments excludes those
190 expectations that may be irrational or largely inappropriate, that is to say, those that might not be pertinent to
191 either trust or distrust. To make this point more graspable consider this hypothetical case discussed by Hawley
192 (2014). Assume that a colleague regularly offers me the leftovers of her lunch not because I explicitly entrusted
193 her to provide for my lunch (and she agreed) or because she promised she would share it with me. Rather, this
194 happens simply because she is unable to properly quantify how much food she will actually need for lunch. If this
195 happens regularly, I can get to rely on her to share her food with me. However, if I developed a normative
196 expectation that she *should* bring food for me to the office and she failed to do so, an attitude of distrust from my
197 side would be largely inappropriate. More precisely, in this case, neither trust nor distrust seem to be suitable
198 attitudes. Cases similar to this elucidate a crucial strength of the commitment account: it provides a solid
199 theoretical basis to distinguish between appropriate and inappropriate normative expectations precisely by making

200 reference to the nature of the commitment that underlies a relationship of trust. So, the goal of Hawley's
201 commitment account is not only to provide the theoretical tools to identify situations in which trust or distrust are
202 suitable but to indicate when neither is suitable. Under this heading, distrust amounts to non-reliance and, in
203 addition to this, the belief that the trustee has a commitment to uphold. Respectively, trust amounts to reliance
204 plus the belief that the trustee has a commitment. In both instances, trust or distrust are appropriate attitudes only
205 when a belief in commitment is in place.

206

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

218

219 **Towards an account of institutional distrustworthiness**

220 Demir-Doğuoğlu and McLeod (2023) formulated a critique of Hawley's account, defending the view that
221 interpersonal (dis)trust should not be seen as entirely separated from institutional (dis)trust. By criticising the
222 commitment account from a feminist perspective, the authors conceptualise the institutional distrust that oppressed
223 groups may encounter. Let us reconstruct both points of critique in turn in order to show how a slightly modified
224 view of the commitment account can be useful to make sense of the distrust (and not mere non-reliance)
225 experienced by members of disadvantaged social groups in collective contexts. This notion of institutional distrust
226 will provide us with ample guidance on how to better understand the considerations related to distrust in medical
227 ML .

228

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

238

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

253

254 Let us now turn to the second critique. This amounts to the fact that an absence of belief in commitment does not
255 make the experienced distrust of oppressed people inappropriate. There are many examples in which members of
256 oppressed groups feel distrust in public institutions and find the commitments these have made unbelievable.
257 Demir-Doğuoğlu and McLeod point out the distrust that Black people in the US can have with respect to the
258 veracity of the commitment of the police to racial equality in view of many cases of police violence and brutality

259 targeted at people of color. That is to say, one could doubt that the commitment of the police only amounts to
260 window dressing, i.e., doubt that it is believable. As Specker Sullivan (2023; p. S36) points out, “our assumptions
261 about what we can expect from others and whether we can believe what they tell us influences our decisions to
262 accept vulnerability to them and depend on them.” If the experiences one makes in medical encounters do not
263 show an explicit commitment to a good delivery of health support, then one is apt to decide to withhold trust and
264 reduce dependency to a minimum (other than in cases of practical necessity).

265

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

272

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

283

284 What has been said so far and in particular the two conditions for institutional distrustworthiness we spelled out
285 (i.e., the *distrust-despite-reliance* and the *absence-of-commitment* conditions) supports two points of great
286 relevance for the paper's overall goal. First, a revised version of Hawley’s commitment account can allow us to
287 make sense of distrust, even at the institutional level. Second, it is, contra Hawley, normatively relevant to
288 distinguish between non-reliance and distrust as well as beyond the boundaries of interpersonal relations to make

289 sense of structural injustices and issues pertaining to the unequal distribution of social power. In support of the
290 second point mentioned, Fricker (2023; p. 8) suggests that the synonymous relationship between the
291 trustworthiness of individuals and commitment should be resisted. Some institutions can be more reliable and
292 have commitments and obligations, and when they do, the register of trust is in order. The author argues that “we
293 do need to theorize trustworthiness in organizations, for there are some institutional bodies and processes whose
294 dysfunctionality and moral status we can only fully understand if we pay attention to *ethos*, and the potential
295 betrayal of individuals and groups that depend on them” (Fricker 2023; p. 741). In other words, an institution
296 could have joint commitments explicitly made to values that comprise their ethos, or have commitments to joint
297 decisions, actions, policies, and processes that embody those values. As Walker (2006; p. 84) confirms, it seems
298 that what we often trust is not an individual person but “the reliable good order and safety of an environment.”
299 Above and beyond this theoretical perspective, there is plenty of empirical support for the view that people do
300 trust (or distrust) organisations (Holland et al. 2022).

301

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

316

317 **Institutionalised opacity**

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

329

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

339

340 To illustrate this difference, consider the example described by Fricker (2023; p. 736): “We show up at A&E with
341 a broken leg, and things happen; care is delivered, as best they can under stretched circumstances, because the
342 Accident and Emergency service is committed to giving appropriate care, and because its staff respond to our
343 manifest, acute dependence on them. If the service is trustworthy-in-general, then it reliably acts on precisely this
344 sort of commitment-based, and dependence-responsive, motive.” However, there may be more complex situations
345 in which the cause of a certain health issue is not so easily accessible for healthcare providers, as it happens in
346 many cases for patients suffering from chronic syndromes or psychosomatic diseases. Under this heading, what
347 happens when a patient shows up at a hospital not with a broken bone (i.e., an objectively recognisable health

348 issue) but with a condition where credibility questions are at stake? In these cases, the testimonial offerings of
349 patients play a crucial role in allowing healthcare professionals to grasp the nature of the condition and take
350 possible remedial actions.

351

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

361

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

377

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

393
394 Yet, the growing institutional opacity should not lead to fatalism and present the ascription of responsibility as no
395 longer possible – this is often referred to as the “many hands” problem (Coeckelbergh 2020). With the focus on
396 institutional distrustworthiness, we can still highlight the need to hold institutions accountable for being unworthy
397 of our trust. This enables the distribution of responsibility across a diverse network of both human and artificial
398 agents. Humans are still running institutions that apply ML systems, which are complex but still embedded in
399 institutional contexts that call for accountability. Fricker provides another example of a hospital that is held
400 accountable for its caring services: “Just as an individual doctor, in acting on her commitment to give proper care
401 to a patient, is thereby displaying a responsiveness to the patient’s dependence on her, so is a hospital that provides
402 care for its patients for commitment-based reasons displaying a responsiveness to their dependence on *it*” (Fricker
403 2023; p. 736). The same logic can be applied to hospitals that use ML tools. We can and should direct our natural
404 reactive attitudes of accountability, such as a sense of betrayal and feelings of resentment, towards the organisation
405 itself rather than towards the individuals acting under the auspices of the institution.

406

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

408 So far, we have offered an analysis of institutional distrustworthiness (based on two conditions, i.e., the *distrust-*
409 *despite-reliance* and the *absence-of-commitment* conditions) and arising, among others, from institutional opacity.
410 Against this background, it is an open question whether we should try to “fix” possible distrust in medical ML.
411 The value of distrust is itself morally ambivalent and shaped by the positionality of people: it can be functional to
412 resisting oppression just as it can enable it. Depending on a person’s social situatedness, an attitude of distrust can
413 point to a justifiably critical position that refuses to accept what dominant social groups might impose as a
414 universal truth. On the other hand, for an agent in a position of social power, illegitimately distrusting another
415 agent due to stigma and prejudices connected to their social identity can perpetuate oppressive patterns and
416 systemic injustices. Concerning the first case mentioned, Demir-Doğuoğlu and McLeod (2023; p. 137) argue that
417 “institutional distrust *itself* can have positive effects for members of oppressed groups, as the attitude
418 fundamentally aims to protect them from further institutional harm and violence”. Likewise, Krishnamurthy
419 (2015) has argued that distrust can be a strategic tool to safeguard oppressed communities against tyranny. Under
420 this heading, vigilance and wariness about patients and healthcare entities may, in fact, be healthy responses to
421 the history of racism and discrimination in medicine and healthcare.

422
423 With these positive connotations of distrust in mind, we can identify some strategies that public health institutions
424 can adopt to ameliorate the two conditions of distrustworthiness. In this way, we avoid addressing an attitude
425 people can (justifiably) hold but rather a central property of relevant institutions. A crucial step – the minimal
426 condition that must be met to begin with – is to acknowledge and take seriously existing relationships of distrust
427 that emerge from different forms of epistemic injustice that are overly experienced by members of socially
428 disadvantaged groups and move from this location. Rectifying injustice and, consequently, addressing issues that
429 determine and justify attitudes of distrust requires acknowledgement, specifically acknowledging one’s actions as
430 redress for one’s wrongdoing (Walker 2006). An acknowledgment of one’s responsibility for perpetrating
431 epistemic injustice does not complete the process of epistemic repair nor, on its own, license renewed trust. It is
432 an important step, nonetheless, without which victims of unacknowledged epistemic injustice are otherwise
433 deprived of the considerable practical and epistemic benefits enabled by functional trust relationships.

434
435 Other ways of encouraging the creation and maintenance of an institutional ethos of testimonial justice amount to
436 “tailoring institutional norms and values” and “cultivating institutional appreciation of human diversity” (Carel
437 and Kidd 2021). The first strategy aims to broaden institutional conceptions of flourishing and care and calls upon

438 “institutions to tailor and relativise their norms to individual people, seen within their context” (Carel and Kidd
439 2021; p. 489). It is crucial that institutions acknowledge a flexible and inclusive use of medical ML and reflect on
440 the individual patients’ needs, desires, and values. Medical institutions have the tendency, often for good reasons
441 of cost efficiency, to standardise and thus assume a certain degree of uniformity throughout patients’ groups.
442 However, this needs to be balanced against other values like accessibility, flexibility, and inclusion.

443

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

456

457 Determining this alignment necessitates at least two components: users must comprehend the organisation's
458 structure, and they need to be able to speak effectively about what their potential roles within it might be. One can
459 imagine how persons with disabilities do not find spaces, for instance, within institutions corrupted by ableist
460 prejudices. Likewise, one could argue that algorithms can be designed that are discrimination-aware and embrace
461 the diversity of access needs (Cirillo et al. 2020). However, these approaches —such as many approaches towards
462 algorithmic fairness available in the literature — are often too restrictive as they advance a decontextualised
463 analysis of the “algorithm itself” (Hull 2023). We maintain that without efforts to tackle systemic injustices such
464 as institutional racism, these initiatives are unlikely to succeed, and inequalities remain. Therefore, medical ML
465 might reiterate the current status quo in healthcare, where very few dominant groups are privileged to the detriment
466 of others.

467

468 **Conclusion**

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

475
476 Given the risks of medical ML when it comes to reproducing and exacerbating existing epistemic injustices, we
477 offered a new perspective in the medical ML debate by introducing the concept of “institutional
478 distrustworthiness”. We suggested that there is a need to understand not only how mistrust as an attitude of people
479 towards the use of medical ML emerges but also how distrustworthiness, as a property of institutions themselves
480 and as a whole, manifests. Therefore, we developed an account of institutional trustworthiness based on the work
481 of Hawley and Fricker. We further argued that institutional opacity can undermine the trustworthiness of medical
482 institutions and how new testimonial injustices can occur with the use of medical ML. We concluded by offering
483 some, albeit initial, ways of addressing this potent juncture of injustice and focused on repairing and ameliorating
484 institutional distrustworthiness.

485
486 While this paper is limited to providing a broad conceptual sketch, it offers a relevant starting point for further
487 discussions on how the ethos of institutions and attitudes of distrust are taking shape in specific medical ML
488 practices. We raised awareness of the magnitude of this problem and only scratched the surface of how to
489 ameliorate institutional opacity and distrustworthiness. However, we encourage further research working towards
490 dismantling unjust structural mechanisms instead of rushing the deployment of medical ML in institutions that do
491 not account for diversity, risking providing insufficient healthcare support for already disadvantaged populations.

492

493 **References**

- 494 Abraham, J. M., and V. Rajasekaran. 2024 Emotions of endometriosis in clinical encounters: An analysis of
495 women's experiences of health care. *Journal of Evaluation in Clinical Practice*. Doi:10.1111/jep.1397
- 496 Baier, A. 1986. Trust and antitrust. *Ethics* 96(2): 231–260.
- 497 Benjamin, R. 2019. Assessing risk, automating racism. *Science* 366(6464): 421–422.

- 498 Baker, R. 2013. *Before bioethics: a history of American medical ethics from the colonial period to the bioethics*
499 *revolution*. Oxford: Oxford University Press.
- 500 Bjerring, J.C., and J. Busch. 2021. Artificial Intelligence and Patient-Centered Decision-Making. *Philosophy &*
501 *Technology* 34: 349–371.
- 502 Braun, M., Bleher, H., and P. Hummel. 2021. A leap of faith: is there a formula for “Trustworthy” AI?. *Hastings*
503 *Center Report* 51(3): 17–22.
- 504 Bryson, J. 2018. AI & Global Governance: No one should trust AI. *United Nations Centre for Policy Research*.
505 <https://unu.edu/cpr/blog-post/ai-global-governance-no-one-should-trust-ai>. Accessed: May 16, 2024.
- 506 Burrell, J. 2016. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big data &*
507 *society* 3(1): 2053951715622512
- 508 Carel, H., and I. J. Kidd. 2021. Institutional opacity, epistemic vulnerability, and institutional testimonial justice.
509 *International Journal of Philosophical Studies* 29(4): 473–496.
- 510 Coeckelbergh, M. 2020. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of
511 Explainability. *Science and Engineering Ethics* 26, 2051–2068.
- 512 Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., ... and N. Mavridis. 2020. Sex and
513 gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ digital medicine* 3(1):
514 1–11.
- 515 Curry, T. J. 2020. Conditioned for death: analysing black mortalities from Covid-19 and police killings in the
516 United States as a syndemic interaction. *Comparative American Studies An International Journal* 17(3-4): 257–
517 270.
- 518 Davidson, L. J., and M. Satta. 2021. Justified social distrust. In *Social Trust*, eds. Kevin Vallier and Michael
519 Weber, 122–148. New York: Routledge.
- 520 Davis, N. 2021. AI skin cancer diagnoses risk being less accurate for dark skin – study. *The Guardian*.
521 [https://www.theguardian.com/society/2021/nov/09/ai-skin-cancer-diagnoses-risk-being-less-accurate-for-dark-](https://www.theguardian.com/society/2021/nov/09/ai-skin-cancer-diagnoses-risk-being-less-accurate-for-dark-skin-study)
522 [skin-study](https://www.theguardian.com/society/2021/nov/09/ai-skin-cancer-diagnoses-risk-being-less-accurate-for-dark-skin-study). Accessed: March 17, 2024.
- 523 Demir-Doğuoğlu, H., & McLeod, C. 2023. Toward a Feminist Theory of Distrust. In *The Moral Psychology of*
524 *Trust*, eds. David Collins, Iris Vidmar Jovanović, and Mark Alfano, 125–143. London: Lexington Books.
- 525 Durán, J. M., and K. R. Jongsma. 2021. Who is afraid of black box algorithms? On the epistemological and ethical
526 basis of trust in medical AI. *Journal of Medical Ethics* 47(5): 329–335.
- 527 Durán, J. M. and G. Pozzi. (under review). What is Trustworthy AI?
- 528 European Commission. 2019. Ethics guidelines for trustworthy AI. High-level expert group on artificial
529 intelligence. *European Commission*. [https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-](https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai)
530 [trustworthy-ai](https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai). Accessed: May 16, 2024.
- 531 Freiman, O. 2023. Making sense of the conceptual nonsense ‘trustworthy AI’. *AI and Ethics* 3(4): 1351–1360.
- 532 Fricker, M. 2023. Diagnosing Institutionalized ‘Distrustworthiness’. *The Philosophical Quarterly*, 73(3): 722–
533 742.
- 534 Graham, S. S. 2022. *The doctor and the algorithm: Promise, peril, and the future of health AI*. Oxford: Oxford
535 University Press.
- 536 Grasswick, H. 2017. Epistemic injustice in science. In *The Routledge handbook of epistemic injustice*, eds. Ian
537 James Kidd, José Medina and Gaile Pohlhaus Jr, 313–323. New York: Routledge.
- 538 Hardin, R. 2002. *Trust and Trustworthiness*. New York: Russell Sage Foundation.
- 539 Hatherley, J. J. 2020. Limits of trust in medical AI. *Journal of medical ethics* 46(7): 478–481.

- 540 Hawley, K. 2014. Trust, distrust and commitment. *Noûs* 48(1): 1–20.
- 541 Hawley, K. J. 2017. Trust, distrust and epistemic injustice. In *The Routledge handbook of epistemic injustice*, eds.
542 Ian James Kidd, José Medina and Gaile Pohlhaus Jr, 69–78. New York: Routledge.
- 543 Ho, A. 2008. The individualist model of autonomy and the challenge of disability. *Journal of Bioethical Inquiry*
544 5: 193–207.
- 545 Holland, S., Cawthra, J., Schloemer, T., and P. Schröder-Bäck. 2022. Trust and the acquisition and use of public
546 health information. *Health Care Analysis*, 1-17.
- 547 Hull, G. 2023. Dirty data labeled dirt cheap: epistemic injustice in machine learning systems. *Ethics and*
548 *Information Technology* 25(3): 38.
- 549 Jacobs, N., and J. Evers. 2023. Ethical perspectives on femtech: Moving from concerns to capability-sensitive
550 designs. *Bioethics* 37(5): 430–439.
- 551 Jones, Karen. 1996. Trust as an Affective Attitude. *Ethics* 107(1): 4–25.
- 552 Khan, M., and C. Ewuoso. 2024. Epistemic (in)justice, social identity and the Black Box problem in patient care.
553 *Medicine Health Care and Philosophy* 27: 227–240.
- 554 Krishnamurthy, M. 2015. (White) Tyranny and the democratic value of distrust. *The Monist* 98(4): 391–406.
- 555 Laux, J. 2023. Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design
556 of AI governance under the European Union AI Act. *AI & Society*. doi:10.1007/s00146-023-01777-z
- 557 Ledford, H. 2019. Millions affected by racial bias in health-care algorithm. *Nature* 574(31): 2.
- 558 Newman, A. M. 2022. Moving beyond mistrust: Centering institutional change by decentering the white analytical
559 lens. *Bioethics* 36(3): 267–273.
- 560 Nickel, P. J. 2022. Trust in medical artificial intelligence: a discretionary account. *Ethics and Information*
561 *Technology* 24(1): 7.
- 562 Medina, J. 2013. *The epistemology of resistance: Gender and racial oppression, epistemic injustice, and the social*
563 *imagination*. Oxford: Oxford University Press.
- 564 Medina, J. 2020. Trust and Epistemic Injustice. In *The Routledge handbook of trust and philosophy*, eds. Ian
565 James Kidd, José Medina and Gaile Pohlhaus Jr, 52–63. New York: Routledge.
- 566 McLeod, C. 2015. Trust. *The Stanford encyclopedia of philosophy*. [https:// plato.stanford. edu/archives/](https://plato.stanford.edu/archives/fall2015/entries/trust/)
567 [fall2015/entries/trust/](https://plato.stanford.edu/archives/fall2015/entries/trust/). Accessed: May 16, 2024.
- 568 Obermeyer, Z., Powers, B., Vogeli, C., and S. Mullainathan. 2019. Dissecting racial bias in an algorithm used to
569 manage the health of populations. *Science* 366(6464): 447–453.
- 570 Pellegrino, E. D., and D. C. Thomasma. 1993. *The virtues in medical practice*. New York: Oxford University
571 Press
- 572 Pozzi, G. 2023. Testimonial injustice in medical machine learning. *Journal of medical ethics* 49(8): 536–540.
- 573 Robertson, C., Woods, A., Bergstrand, K., Findley, J., Balsler, C., and M. J. Slepian. 2023. Diverse patients’
574 attitudes towards Artificial Intelligence (AI) in diagnosis. *PLOS Digital Health* 2(5): e0000237.
- 575 Scully, J. L. 2014. Disability and vulnerability: On bodies, dependence, and power. In *Vulnerability: New essays*
576 *in ethics and feminist philosophy*, eds. Catriona Mackenzie, Wendy Rogers, and Susan Dodds, 204–221. Oxford:
577 Oxford University Press.
- 578 Segers, S., & Mertes, H. 2022. The curious case of “trust” in the light of changing doctor–patient relationships.
579 *Bioethics* 36(8): 849–857.
- 580 Sherlock, R. 1986. Reasonable men and sick human beings. *The American journal of medicine* 80(1): 2–4.

- 581 Smith, H. 2021. Clinical AI: opacity, accountability, responsibility and liability. *AI & Society* 36(2): 535–545.
- 582 Specker Sullivan, L. 2023. Climates of Distrust in Medicine. *Hastings Center Report* 53: S33–S38.
- 583 Starke, G., and M. Ienca. 2022. Misplaced Trust and Distrust: How Not to Engage with Medical Artificial
584 Intelligence. *Cambridge Quarterly of Healthcare Ethics*. doi:10.1017/S0963180122000445
- 585 Szalavitz, M. 2021. The pain was Unbearable. so why did doctors turn her away. *Wired*.
586 <https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/>. Accessed: May 16, 2024.
- 587 Walker, M. U. 2006. *Moral repair: Reconstructing moral relations after wrongdoing*. New York: Cambridge
588 University Press.
- 589 Wilson, Y. 2022. Is Trust Enough? Anti-Black Racism and the Perception of Black Vaccine “Hesitancy”. *Hastings*
590 *Center Report* 52: S12–S17.
- 591 Wolkenstein, A. 2024. Healthy Mistrust: Medical Black Box Algorithms, Epistemic Authority, and
592 Preemptionism. *Cambridge Quarterly of Healthcare Ethics*. doi:10.1017/S0963180123000646
- 593 Zanotti, G., Petrolo, M., Chiffi, D., and V. Schiaffonati. 2023. Keep trusting! A plea for the notion of Trustworthy
594 AI. *AI & SOCIETY*. doi:10.1007/s00146-023-01789-9