Minds and Machines (2026) 36:10
https://doi.org/10.1007/s11023-026-09762-3

Epistemic Justice as a Condition for Meaningful Human Control Over Medical AI

Giorgia Pozzi[1] · **Filippo Santoni de Sio**[2]

Received: 27 March 2025 / Accepted: 16 January 2026

Abstract

AI technologies are increasingly deployed in medical care and decision-making, and efforts geared toward conceptualizing how human control over AI systems can be *meaningful*, i.e., sufficient to preserve the relevant human agency and responsibility, are mounting. However, a suitable conceptualization of Meaningful Human Control (MHC) explicitly tailored to AI-mediated clinical practice is still underdeveloped. This paper addresses this research gap in two ways. First, it applies the framework of Meaningful Human Control as reason-responsiveness to the medical field. Second, it shows that considerations of epistemic (in)justice ought to be included in efforts toward securing MHC in medical care. MHC demands that the moral reasons of relevant agents be made available to the socio-technical system in which the AI operates. However, this requirement can be compromised by epistemic injustices, i.e., when patients' and clinicians' epistemic offerings to the medical discourse are unduly limited. The paper argues that epistemic justice is an important enabler for MHC, and, when properly understood, MHC is a crucial element in a strategy to promote a more just medical AI. Since epistemic injustice depends on power asymmetries and systemic inequalities, achieving epistemic justice and MHC over medical AI requires addressing power and justice issues in the development and use of (new) medical AI.

Keywords Meaningful human control · Medical AI · Epistemic injustice · Ethics of AI · Epistemology of AI

✉ Giorgia Pozzi
G.Pozzi@tudelft.nl

Filippo Santoni de Sio
f.santoni.de.sio@tue.nl

[1] Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands

[2] Eindhoven University of Technology, De Zaale, Eindhoven, The Netherlands

Published online: 06 February 2026

1 Introduction

Artificial intelligence-based (AI) systems are powerful epistemic technologies. As Alvarado argues, AI "is primarily designed, developed *and* deployed to be used in epistemic contexts such as inquiry, it is explicitly deployed in such contexts to manipulate epistemic content such as data, *and* it manipulates such content specifically through epistemic operations such as inferences, predictions or analysis." (Alvarado, 2023). Moreover, AI techniques and methods are always deployed as part of broader socio-technical systems, which include data and the people who produce them, technical interfaces that allow and mediate the interaction between people and technologies (computers, software, applications), and social and institutional structures within which these interactions take place (rules, practices, laws) (Santoni de Sio, 2024, p. 4). When introduced as support to healthcare professionals in central tasks pertaining to diagnostic and patient treatment, among many others, AI systems become important mediators in the production of medical knowledge and clinical decision-making that may crucially shape the interaction between doctors and patients. While it is still unclear what the role of emerging AI techniques like Generative AI and Large Language Models (LLMs) will be in healthcare, some older forms of AI, namely machine learning (ML), are currently implemented to take up a variety of tasks in medicine and healthcare. High levels of accuracy have been, for instance, achieved in the analysis of images in radiology and skin cancer screening (Esteva et al., 2017; Topol, 2019). Risk assessment tools are also increasingly used, for example, to predict patients' likelihood of developing sepsis in intensive care units (Ross, 2021).

One central general ethical question with the introduction of AI systems in healthcare is how AI can create shifts in moral agency and control. The first problem in this regard concerns the responsibility of clinicians. Since medical decisions often occur under considerable uncertainty and time limitations, the reliance of clinicians on AI systems risks limiting their agential space, that is, their possibility to maintain sufficient power and control over the final clinical decision. When doctors' power and control are unduly reduced, they may be less able to take moral (and even legal) responsibility for a particular course of medical action. This concern is related to the widely discussed problem of the responsibility gap with AI, which, in general, designates morally problematic situations in which someone is wronged by the decision of an AI system, and no human agent fulfills the central conditions needed for responsibility attribution (Matthias, 2004). One often-discussed version of the responsibility gap is that concerning culpability, that is, the question of who is to *blame* or should be held *legally liable* when someone is wronged by a (medical) decision taken by or with the decisive support of an AI system (Beck et al., 2024). However, the responsibility gap problem has broader moral significance in medicine and elsewhere. When doctors lose control over the AI system they interact with, they may also be put in a moral corner when it comes to, among others, justifying their decisions to patients[1]. They may also be less able to live up to their professional standards and identity[2].

[1] This is what Santoni de Sio & Mecacci (2021) call the accountability gap.

[2] This is what Santoni de Sio & Mecacci (2021) call the active responsibility gap.

A second problem explicitly concerns the moral agency of patients and clinicians. Not only clinicians but also patients (should) play an important role in medical interactions. Think of patients' role in collaborating with clinicians to get an accurate diagnosis by presenting and discussing their symptoms and experiences, as well as taking part in the decision-making process about their best treatment. As it were, patients should also remain *agents* able to actively contribute to the medical process. When the use of AI systems unduly reduces the patients' (or clinicians') capacity to offer their knowledge and agency within healthcare practices, two potential related moral harms may occur: patients or clinicians are wronged by not being sufficiently respected in their capacity as knowers, i.e., they are victims of epistemic injustice, *and* the quality of healthcare may be negatively affected by their exclusion.

One important philosophical question raised by these responsibility and agency issues is which conception of human control should be used in the context of AI. Given the above-mentioned epistemic and systemic nature of AI, human control cannot be understood as merely having a "human in the loop" with the opportunity to causally influence the AI system's behavior, as in the paradigmatic case of someone sitting in the driver's seat of a car. First, since AI is an epistemic technology, some relevant epistemic conditions must also be realized for human control to be meaningful. That is, the relevant human agents must possess some relevant knowledge about the AI system involved in clinical decision-making. Second, since AI is part of complex socio-technical systems, human control can be distributed across a range of different agents, which, in the case of medical AI, includes the AI's developers, the doctors who directly interact withit as well as the patients who provide the data to the AI system, the managers who decide if and how doctors must use the AI and others. Third, and relatedly, human control over technology does not only concern the relation between one person and one technical system but also the power relations between different people and institutions involved in technology development and use.

The philosophical concept of Meaningful Human Control (MHC) as reason-responsiveness has been introduced as a measure to counter the emergence of responsibility gaps with AI and, more generally, to make sense of and protect human moral agency within AI systems (Santoni de Sio & van den Hoven, 2018). The concept has been discussed from different disciplinary perspectives and in relation to various domains of applications (Mecacci et al., 2024), including the medical context (Beck et al., 2024; Braun et al., 2021; Hille et al., 2023). In this paper, we propose what we consider to be an important integration and deepening of the analysis of the concept of MHC in the medical domain. Following up and elaborating on an argument made by Santoni de Sio (2024) in relation to AI systems generically conceived, we argue that MHC over medical AI requires *epistemic justice*. To support this claim, we proceed as follows.

In Sect. 2, we provide a brief review of the applications of MHC in medicine currently available in the literature and motivate why we adopt and tailor to the medical context the account of MHC advanced by Santoni de Sio & van den Hoven (2018). In Sect. 3, we consider the case of an AI system used to support clinicians in pain management. Through the analysis of this case, we show that patients and clinicians may be victims of epistemic injustice in their interaction with AI systems when (new)

technical and social norms and practices do not sufficiently allow their morally relevant offerings to be reflected in its functioning. We claim that when this happens, patients and clinicians may be wronged in their capacity as knowers (Pozzi, 2023b, 2023a). To make the connection between epistemic injustice in medical AI and MHC explicit, we further tackle, in Sect. 4, three main mechanisms of epistemic injustice in AI that hamper the fulfillment of requirements for securing MHC as reason-responsiveness. As we further show in Sect. 5, the focus on epistemic justice as a precondition for control reveals the need for a more detailed analysis of patients' participation beyond mere involvement in shared decision-making practices in AI-supported medical delivery (Bjerring & Busch, 2021; Salloch & Eriksen, 2024). We should not underestimate how underlying power dynamics and structures of inequality can make interactions taking place within the patient-physician-AI triad unjust. In fact, the participation paradigm may fall short if the voices of relevant agents do not receive appropriate social recognition and uptake to be effectively integrated into the medical discourse. As proposed in the literature on design justice (Costanza-Chock, 2020), a superficial application of the participation model may even backfire and allow for subtle forms of exploitation.

Directing attention to justice also sheds light on, more generally, the importance of addressing social structures of power to allow people to remain in control of technology. In fact, as we will see, the introduction of AI systems in healthcare may pose a threat to human control of technology not only for patients but also for clinicians, that is, highly educated and trained specialists with an otherwise well-recognized epistemic, social, and professional role. One broader lesson is that while AI usually reinforces existing positions of power and social vulnerabilities, it sometimes introduces new forms of power and creates new vulnerabilities, for instance, among professionals (Pozzi, 2023a).

Finally, *if*, as we argue, (a) to protect their moral agency, patients and clinicians must be given MHC over AI systems; (b) MHC requires, among other things, addressing epistemic injustice; and (c) epistemic injustice typically depends on social structures of power; *then*: it is these social structures of power which, among other things, we must address to prevent a morally relevant loss of control and power by patients and clinicians in relation to AI. As we maintain in Sect. 6, this ultimately requires a systemic approach, which addresses the technical, socio-technical, socio-economic, organizational, and political conditions under which AI is currently developed and introduced in the medical sector.[3] Our analysis thus shows that it is paramount to start thinking about alternative ways to design, develop, and introduce AI technologies in the medical domain that promote MHC *through* epistemic justice.

2 Meaningful Human Control in Medical AI

Before turning to the analysis of how forms of epistemic injustice represent a challenge for Meaningful Human Control (MHC) and how designing for MHC can contribute to making healthcare more just, in this section, we provide a brief overview

[3] Compare van Wynsberghe and Li (2019) for a similar proposal.

of the literature discussing MHC in the context of medical AI, and we introduce our preferred conception of MHC, i.e., MHC as reason-responsiveness (Santoni de Sio, 2024; Santoni de Sio & van den Hoven, 2018).

Starting from the analysis of different normative challenges about medical decision-support systems (e.g., conditions of trustworthiness and shared agency), Braun et al. (2021) refer to several aspects that, according to these authors, are central to an account of MHC in medical AI. The first pertains to the legal dimension of ensuring a human being is in the loop and can be rendered accountable for AI decision-making. The second concerns data sovereignty, i.e., who can access and process data, so control is understood as having a say about how and for what purpose one's data are used. And finally, the third is related to clinicians' epistemic and decisional authority. Here, being in control refers to the fact that it should be in the hands of human clinicians to properly decide when to defer to an AI recommendation and when to follow an independent course of action. All these aspects are relevant to control issues in AI-mediated scenarios, and the last one coincides with one of the concerns discussed in this paper. However, due to its general nature, the approach advanced by Braun and colleagues remains at a quite general level of analysis and does not provide an actionable framework to identify and address the conditions under which MHC can be secured in specific clinical interactions among patients, clinicians, and AI. For example, the framework does not provide concrete guidance on how to assess whether, in a particular use of a medical AI system, clinicians are, in fact, in *meaningful* control of it.

Similarly, Hille et al. (2023) provide a review that captures the state of the art of discussions revolving around MHC in medicine. These authors focus on the issue of *which agents* must be in control of medical AI systems and rightfully argue that, above and beyond designers and clinicians, patients are also relevant agents who need to be granted meaningful control.[4] This claim is one that underlies also the arguments advanced in this paper. Since patients are major stakeholders impacted by the introduction of AI systems in medical decision-making, we submit that they are also entitled to relevant forms of control. This is an important matter of principle, in view of the shift from paternalistic patterns of care to the model of shared decision-making in medicine, according to which patients should not be relegated to the role of passive recipients of care in view of the authority of clinicians grounded in their epistemic status as experts. Rather, it is increasingly recognized that patients have a right to determine their own care in an autonomous way, thus making choices in collaboration with clinicians that mirror their personal values and preferences (Lorenzini et al., 2023). In view of the increasingly relevant AI systems are likely to acquire in medical practices, it is paramount to subscribe to forms of MHC that aim at preserving not only clinicians' but also the intrinsic value of patients' agency and responsibility. What kind of, and how much control by patients is necessary and sufficient to protect their moral agency *and* that of the professionals is the ethical and philosophical question to be addressed.

[4] For an analysis of MHC in connection with legal responsibility in medical decision-making involving AI systems, see Beck et al. (2024). Also these authors point out the need to ensure that patients, as relevant stakeholders, have MHC.

By providing an analysis of the available literature, Hille and colleagues point out factors that enable MHC and identify agents in control, along with evaluators of MHC. However, this paper does not specify under which conditions MHC would be secured, specifically in the context of medical AI. Ultimately, while we agree with Hille and colleagues that also patients should be seen as agents in control and that "(w)hat kind of control is meaningful is (…) not just a technical question, but a social one" (Hille et al., 2023, p. 8), their analysis does not contextualize the social dimension of MHC in existing social epistemological frameworks to identify and make sense of underlying social mechanisms of power and inequality preventing the realization of MHC.

In the remainder of this paper, we aim to add to the social component that impairs MHC by referring to issues of epistemic injustice. However, we need a more specific account of MHC before considering these factors. For this, we tailor the account advanced by Santoni de Sio & van den Hoven (2018), i.e., MHC understood as *reason-responsiveness*, specifically to the context of medical AI.

In their seminal 2018 paper, philosophers Filippo Santoni de Sio and Jeroen van den Hoven proposed a generic account of MHC over AI systems grounded in the theory of *guidance control* developed by John Martin Fischer and Mark Ravizza (1998) in the context of the philosophical debates on moral responsibility. Santoni de Sio and van den Hoven developed and integrated Fischer's and Ravizza's theory of control in two ways. First, they expanded the idea of guidance control from the individual, intrapsychic domain (i.e., control by an individual human agent over one's own everyday actions) to the domain of (distributed) human interaction with AI technology (i.e., control by (a group of) human agents over the behavior of a socio-technical AI system of which they are part). Second, in line with the ideal of Value-Sensitive Design, they started casting the conditions for control so that these, with the support of further scientific, social, and technical research, could be designed into a socio-technical system.[5] Two conditions or requirements for MHC are eventually identified. Let us briefly elucidate both conditions in turn.

The first condition requires that the socio-technical system in which AI operates must remain robustly *responsive* to the *relevant reasons* of the *relevant human agents* in the system. This reason-responsiveness requires that the system's behavior reflects and dynamically co-varies with the relevant intentions, values, norms, or principles of, ideally, all the relevant human agents in the system. Reason-responsiveness, also called the *tracking condition* for MHC, will be the focus of this paper. The second condition requires that some relevant human agents in the socio-technical system – at least one but ideally more - maintain at the same time the capacity to understand the real capabilities of the technical system they interact with and the moral awareness of their moral responsibility for its behavior. This is called the *tracing condition* for MHC. While it is also crucial to avoid responsibility gaps with AI, the latter will not be discussed further in this paper. Following Santoni de Sio (2024), we will call his and van den Hoven's approach "MHC as reason-responsiveness" to distinguish it

[5] See Cavalcante Siebert et al. (2023) for a further attempt to specify these general conditions in technical terms.

from the other approaches present in the literature (Mecacci et al., 2024; Robbins, 2023).

We take this conceptualization of MHC to be particularly fitting for addressing questions of control connected to the use of AI in the medical domain and highlighting its social justice dimension for two main reasons. First, the conditions of tracking and tracing represent normative requirements that can be implemented from the design stages of technological development with the purpose of securing control and responsibility. The availability of specific requirements that can be translated at the level of technological design is particularly relevant in high-stakes domains such as medicine and healthcare, in which responsibility attribution cannot be merely an afterthought. Second, the fact that the tracking condition requires us to tackle relevant moral reasons of relevant agents urges us to critically question whose reasons should be considered as relevant in a particular social context. This condition thus allows us to scrutinize the position of different agents who interact with or are affected by an AI technology and make room for their agency in shaping the impact of the technology in a socially desirable direction (Santoni de Sio, 2024). The latter point lends itself particularly well to the critical scrutiny of existing power imbalances and structural injustices that unduly hinder possibly relevant agents from offering their relevant (moral) reasons in the first place. In turn, due to the focus of MHC as reason-responsiveness on designing AI systems for control and responsibility, imperative questions of social justice can acquire a central stage in the phases of technological design and development. This offers designers concrete reasons to expand their focus beyond an AI system's technical functioning to include central issues of social (in)justice. For all these reasons, we take this conception of MHC as the foundation of the analysis advanced in the remainder of the paper.

3 Epistemic injustice in medical AI

AI systems are currently implemented to take up a variety of tasks in medicine and healthcare. High levels of accuracy have been, for instance, documented in the analysis of images in radiology and skin cancer screenings (Esteva et al., 2017; Topol, 2019). Moreover, risk assessment tools are increasingly used, for example, to predict patients' risk of developing sepsis in intensive care units (Ross, 2021). Recent Generative AI systems, such as Large Language Models (LLMs), also promise to present new ways to develop healthcare relationships mediated by artificial agents. All these systems arguably offer opportunities to improve the quality of healthcare practices, but they also raise questions about what kind of human control and responsibility should be maintained over them. While the issue of clinicians' control and responsibility for the behavior of AI systems has already been discussed to a greater extent (Beck et al., 2024), in this paper, we want to focus on uses of AI in medicine and healthcare that may also undermine the epistemic and moral agency of patients, alongside of discussing the issue of clinicians' control and responsibility for the behavior of AI systems. Therefore, we will do so by focusing on an example of the use of AI in healthcare in which, on the one hand, patients' lived experience and testimonial knowledge – albeit being central to inform medical decision-making - risk

remaining unduly unacknowledged. On the other hand, clinicians are severely hampered in their ability to exercise their critical scrutiny and expertise.

More precisely, we analyze an AI system used in the USA to support healthcare professionals in clinical decision-making by predicting patients' risk of opioid addiction or drug abuse (Siegel, 2022; Szalavitz, 2021, 2024)[6]. Against the background of an ongoing opioid crisis, AI systems are implemented with the goal of providing clinicians with relevant clinical information related to a patient's health status to inform their decision as to whether to prescribe opioid medication. The AI, called NarxCare, ascribes to patients a risk score retrievable for clinicians and pharmacists along with other patients' data in their medical records. While it is not entirely disclosed which proxies play a role in generating the risk scores, Oliva (2022) discusses how problematic metrics are used. For example, whether a patient has experienced sexual assault, has a criminal record, or has traveled long distances to reach a pharmacy or physician, these are all parameters that directly flow into the risk score of patients, thus paving the way for possibly misleading and outright discriminatory outcomes. In fact, traveling a long distance to reach a physician or pharmacy can have reasons that are completely unrelated to what is considered "doctor shopping behavior" (Oliva, 2022) and can indicate an opioid addiction. Patients living in rural areas usually have to travel longer to get the medical care they need compared to people living in densely populated urban areas. Moreover, it has been shown that NarxCare proxies tend to erroneously flag patients with complex medical needs: a study reported that 20% of patients targeted due to presumed doctor and pharmacy shopping needed additional medical care because they were diagnosed with cancer (Buonora et al., 2024; Delcher et al., 2021). These considerations substantiate the worry that inconclusive metrics are used, which can lead to discriminatory outcomes that particularly risk disadvantaging historically underrepresented population subgroups.

So understood, the problem partly amounts to a case of algorithmic bias that occurs when an AI system's output unjustifiably and systematically disadvantages certain social groups (and benefits others), paving the way to patterns of inequality and discrimination (Kilby, 2021; Kordzadeh & Ghasemaghaei, 2022; Panch et al., 2019). Under this heading, this case seems to be sadly in line with many widely discussed instances of algorithmic bias and unfairness. Consider, for example, the often-discussed case advanced by Obermeyer et al. (2019). These authors analyze how an algorithm widely deployed in the USA to decide on the allocation of additional healthcare resources used, as a proxy for eligibility, patients' past health costs. This algorithmic system turned out to be biased against Black patients due to systemic

[6] While this is a case that features specifically in the US, we maintain that it is helpful to highlight some general features of the introduction of AI systems in medical care that may realize also in European or other countries. These amount to issues related to the potential injustice coming from the unfair exclusion of vulnerable populations from decision-making processes and the pressure of professionals to adopt AI out of cost-efficiency or other output-driven metrics (see, e.g., the discussion on managed care advanced later in this section). Also, while this case has been previously scrutinized to spell out some issues of epistemic injustice (Pozzi, 2023a, 2023b), the analysis advanced in this paper complements previous efforts by showing how these issues intersect with concerns related to responsibility and control with the use of AI in healthcare, particularly in spelling out mechanism of epistemic injustice that can hamper the fulfillment of the two central conditions for MHC. We thank an anonymous reviewer for encouraging us to clarify these points.

disparities in accessing healthcare in the first place, thus perpetuating mechanisms of inequality and social exclusion (see also Benjamin, 2019). However, we maintain that there are at least two further aspects to the NarxCare case under consideration that render it problematic beyond the issue of algorithmic bias generically considered.

First, certain types of information can be brought into the medical discourse *exclusively* through patients' testimony and situated knowledge of their lived experience of illness. This issue is thus not solvable at the technical level by, say, adjusting possibly inconclusive metrics (Mittelstadt et al., 2016) and expanding the representation of certain patients' categories in the AI system's training data. Rather, when patients are prevented from offering their knowledge because they receive less credibility than they deserve in the face of an AI risk score, we have a clear source of *epistemic injustice* (Pozzi, 2023a). Epistemic injustice generally occurs when epistemic agents become (a) unable to express their situated knowledge due to a lack of relevant concepts, which constrains their possibility to shape collective moral and social life, thereby perpetuating their conditions of oppression and exclusion; and/or (b) are less able to exercise their epistemic capacities, because of identity prejudices held by members of dominant social groups. The first type of injustice is what Fricker (2007) refers to as hermeneutical injustice, which occurs in cases in which, due to a lack of conceptual resources, members of oppressed social groups cannot express their lived experiences. An example advanced by Fricker to illustrate this type of injustice is the experience of postpartum depression at a time in which this concept was not part of shared hermeneutical resources (and thus not recognized as a medical condition). The second type of injustice mentioned pertains to cases of testimonial injustice (Fricker, 2007). In these cases, the testimony of members of underprivileged social groups is attributed less credibility than they deserve due to illegitimately held prejudices. For example, this form of injustice occurs if a woman's report of the pain she experiences is not taken seriously because a (male) physician believes that women generally exaggerate symptoms (Kidd & Carel, 2017).

In relation to the NarxCare case under scrutiny, instances of epistemic injustice occur, for example, on occasions in which patients want to contradict the risk score they receive because it is not compatible with their own lived experience and health condition, i.e., with the knowledge they hold about themselves. However, they are impaired in doing so because their testimony is not granted credibility in the face of a high risk score (Pozzi, 2023b). Once the AI system has generated an output regarding patients' likelihood of opioid addiction, the information and knowledge they hold about themselves cannot be captured by it (say, contextual information regarding their lived experience) and risks remaining unacknowledged in the further decision-making process if their testimony is unconsidered (Szalavitz, 2021). Empirical research on the impact of the patient-physician relationship substantiates these concerns by showing that "(i)n response to worrisome PDMP [Prediction Drug Monitoring Programs] profiles with new patients, participants [clinicians using automated risk scores] reported declining to prescribe, except in the case of acute, verifiable conditions." (Leichtling et al., 2017, p. 1063) Thus, it has been argued that patients are victims of epistemic injustice because their risk scores are, *de facto*, used as the main markers to assess the trustworthiness of their testimony, which often results in its illegitimate dismissal (Pozzi, 2023b).

To be sure, the need to include patients' testimonial knowledge in medical decision-making varies according to the specificities of different clinical situations in which AI systems play a role. In the specific case under consideration, certain relevant information on patients' health status can *exclusively* be provided by patients actively (e.g., contextual information about their living circumstances that might be relevant for clinical decisions pertaining to pain management, which is thus central to having completeness of the required medical information). This might be different for other clinical uses of AI systems. For example, in radiology, it seems unproblematic that data can be passively extracted, and patients' testimony might play a subordinate role. We intentionally chose a case in which patients' input needs to be included in the medical discourse to shed light on the connection between forms of epistemic injustice and MHC that we will further develop in the remainder of this article.

A second reason why misjudgment by risk-assessment tools like NarxCare algorithms cannot be reduced to cases of algorithmic bias has to do with the role of clinicians who deploy them. NarxCare algorithms are used as decision-support systems, which means that physicians can ultimately decide whether a patient will be prescribed opioid medication or not. However, the higher-level institutional system grants these technologies much more power than one would consider justifiable for decision-*support* systems (Buonora et al., 2024). Oliva (2022) points out that clinicians deciding to prescribe opioids to a patient who has been flagged with a high-risk score make themselves legally liable and even risk losing their practicing authorization. This effectively means that medical professionals are expected to accept the AI system's recommendations to avoid serious legal consequences (Haselager & Mecacci, 2024).[7] Rather than a case of algorithmic bias, this seems closer to a case of automation bias, that is, a situation in which people tend to over-rely on AI systems without critically scrutinizing their output (Goddard et al., 2014; Khera et al., 2023). Crucially, in this case, the over-reliance is caused not only by some general human cognitive features but also by specific legal incentives operating on the clinicians deploying the AI tool.

So, on the one hand, the NarxCare case paradigmatically exemplifies how already disadvantaged categories of patients (e.g., those affected by the stigma of drug misuse) can be further burdened by a technical system that does not account for their lived experience of illness and leaves no space for their voices to be heard (Pozzi, 2023b, 2023a). Here, epistemic injustice seems to track social injustice. On the other hand, the perspective of clinicians shows how epistemic injustice can also depend on a different kind of injustice. The introduction of automatically produced risk scores aligns with the goal of providing clinicians with allegedly objective ways to assess patients' credibility and ultimately reducing the number of unjustified opioid prescriptions they issue. However, as Oliva (2022) points out, if the prescriptions issued decrease following the AI's risk scores, the AI is deemed efficient, irrespective of

[7] Let us clarify here that to satisfy the conditions for MHC, we need *relevant* agents to be in control (Santoni de Sio & van den Hoven, 2018). This means that it is not sufficient to be able to trace responsibility back to any human agents in the system development chain who are legally responsible for the role these systems play (say, lawmakers). Since they cannot account for the specific responsibilities that emerge in medical practice and play out in clinician-patient-AI triadic relationships, we need particular agents, i.e., physicians, to hold MHC and responsibility.

the consequences that medication withdrawal might have on patients' overall health (e.g., irrespective of whether they develop serious mental health conditions, commit suicide, or turn to illicit drugs) (Oliva, 2022).

The idea of prioritizing alleged (cost-)efficiency often at the price of reduced overall patient well-being is in line with general approaches in so-called "managed care", that is, roughly, an approach to healthcare management that sees it as a business rather than a care service (see, e.g., Kersbergen, 2000). This is problematic and unjust, in the terms introduced by the political philosopher Michael Walzer, to the extent that normative standards of assessment pertaining to different "spheres of life" are, unjustifiably, conflated (Walzer, 1983).[8] In this case, the quality of healthcare is mainly assessed through the standards of cost-efficiency rather than through the well-being of patients.

As we will argue in more detail in the next section, the fact that, for different reasons that go beyond the technical functioning of AI systems, relevant information and knowledge coming from patients and clinicians remains out of the medical discourse is problematic in the case under scrutiny since it impairs the fulfillment of the conditions needed to secure MHC over AI systems (Santoni de Sio, 2024; Santoni de Sio & van den Hoven, 2018). This case thus shows the need to consider underlying power dynamics and systemic inequalities and injustices that characterize how AI systems enter the medical field, therefore explicitly connecting issues of control and justice in medical AI. More specifically, we need to explicitly address the question of how socially vulnerable patient populations, but also professionals working under the pressure not to challenge the suggestions of an AI-driven system, are limited in their possibilities to exercise meaningful control over AI technologies. This motivates the need to expand the standard framework of MHC to include considerations related to social justice. This shall include an analysis of the conditions for epistemic justice for physicians vis-à-vis technology developers and management. At the same time, framing the problem in terms of control will allow us to envisage positive design and policies to improve the justice of the entire socio-technical system, as opposed to just exposing its injustice.

4 How Epistemic Injustice in AI Hampers MHC

As pointed out in Sect. 2, MHC as reason-responsiveness requires that an AI system is designed to be sensitive and responsive to the relevant reasons coming from relevant human agents involved in AI-supported decision-making. By definition, this MHC account is highly context-dependent, as what counts as a *relevant* reason of a *relevant* agent crucially depends on the normative context of application (Mecacci & Santoni de Sio, 2020). A general implication is that relevant reasons of relevant agents must, first of all, be available to the system for them to be successfully tracked.[9] In a nut-

[8] For further reflections on how AI systems impact the sphere of clinical care and healthcare professionals' moral and epistemic role in it see also Pozzi and Van den Hoven (2023).

[9] We take the idea that the system needs to be responsive to patients' relevant reasons (or conversely, that relevant agents need to see their reasons reflected in the system's behavior) as falling under the tracking

shell, this is why epistemic injustice can affect MHC in the medical context, to the extent that underlying power dynamics in AI-supported medical systems can strongly constrain which reasons are included in the AI system's decisions. In other words, the question of which, particularly *whose* reasons are deemed relevant, is closely tied to social (in)justice issues. As Santoni de Sio (2024) points out, under conditions of epistemic injustice, *relevant* reasons risk being unduly conflated with *dominant* ones (p. 188).

In order to show how forms of epistemic injustice can hamper MHC as reason-responsiveness, we identify, by referring to the NarxCare case introduced in Sect. 3, three epistemically unjust mechanisms that limit patients' possibility to make their reasons available to the medical discourse. We make this division for analytical purposes because, in practice, we take the three to intersect with each other. We refer to the first as the *objectification constraint*, the second as the *capacity constraint*, and the third as the *self-silencing constraint*. In the following, we address each one in turn.

The *objectification constraint* pertains to forms of epistemic objectification occurring in cases where patients are treated as *mere* sources of information instead of full-fledged epistemic agents (Fricker, 2007). Fricker recognizes in this phenomenon the core issue at the root of testimonial injustices, i.e., those emerging due to an unjustified denial of credibility of someone's testimony based on unfounded prejudices related to the speaker's social identity. In standard (i.e., not AI-supported) healthcare settings, this can occur in the case that physicians only consider information that can be indirectly extracted about a patient's health status as medically relevant due to prejudices related to patients' ability to meaningfully contribute to medical processes (Kidd & Carel, 2017). This typically means that only interactions in which patients are passive (say, clinicians obtain information about their health status through lab tests or scans that patients undergo) are considered medically valuable (Pozzi & Durán, 2024a). In turn, patients are treated as *mere* sources of information if the knowledge that exceeds what can be indirectly derived about their biological health but is nonetheless relevant for appropriate medical decision-making remains unacknowledged (Kidd & Carel, 2017). In cases of epistemic objectification, patients are not granted the possibility to *actively* contribute to the medical discourse through their testimonial report of their lived experience of illness (e.g., because this is considered too idiosyncratic to be valuable for medical action) (Kidd & Carel, 2017).

We argue that AI systems like NarxCare are likely to systematize patients' objectification if they are used as an *alternative* to patients' active reports of the knowledge they can convey through their testimony. Unfortunately, there are reasons to think that this is the case. For instance, Szalavitz (2021) reports the story of a woman suffering from endometriosis who was discharged from the hospital in the face of a high Narx Score without even being given the possibility to share relevant information that might support her eligibility for narcotics despite theAI's red flag. In her particular case, it later turned out that her pet's medication was included in her medical record and erroneously ended up informing her risk score. Above and beyond this

condition for MHC. We leave open the question of whether this would also be compatible with the tracing condition, as this does not compromise the arguments advanced in the remainder of the paper.

particular case, the idea behind these automated systems seems to be the need for an allegedly objective baseline to allocate opioid medication without having to rely on and assess the veracity of patients' input (Chiarello, 2021). However, this is likely to lead to a dangerous shift in medical practice in which patients' testimony remains unaccounted for, and their values and preferences, the consideration of which is central to the medical ideal of shared decision-making, cannot be upheld. Crucially, also morally relevant reasons, i.e., facts about one's medical condition, can fail to receive appropriate uptake, thus leading to unsuitable patient treatment.

Against this background, it becomes clear that some forms of epistemic objectification that risk becoming systematized through the use of AI systems stand in the way of fulfilling the conditions for MHC. Recall that the tracking condition for MHC, as it has been presented in Sect. 2, requires that the socio-technical system in which the AI is embedded is responsive to the relevant moral reasons of relevant agents. In the case considered in this paper, this also concerns clinicians and patients. However, if patients are epistemically objectified, they cannot *actively* convey morally relevant information (e.g., values and treatment preferences), and these cannot be rendered available to the AI system (Pozzi & Durán, 2024b). This may sometimes lead to a failure to include relevant reasons in the medical setup in which the system mediates medical practices, such as those in which an AI influences patients' treatment options, like in the case under scrutiny. We maintain that these forms of objectification are prone to emerge at a higher scale with the use of AI due to clinicians' tendency to over-rely on AI systems. As previously pointed out, empirical studies confirm the frequent occurrence of automation bias (Goddard et al., 2014; Khera et al., 2023). This might lead clinicians to neglect other relevant sources of information, such as what patients should be enabled to actively contribute to the medical discourse, thus perpetuating their objectification.

So understood, the objectification constraint represents an overarching issue that either healthcare professionals or other structural healthcare mechanisms subject patients to. These can amount to time pressures and a culture of efficiency, potentially leading to dismissive patterns of care, which are considerably exacerbated by a possible over-reliance on AI systems as alleged sources of objective information about patients. However, the other two constraints we identify find their home in how patients themselves react to (systematic) testimonial injustices and objectification that is likely to be further reinforced if AI systems are introduced in the medical context as an epistemically authoritative entity providing clinicians with actionable information about patients. Thus, the next two forms shed light on mechanisms by which patients *self-constrain* their own testimonial offerings in what is perceived to be a hostile epistemic environment.

The *capacity constraint* represents a further challenge for MHC. As Santoni de Sio (2024) points out, patients' capacity to offer their moral reasons can be impaired if people are repeatedly victims of epistemic injustice. This constraint highlights a psychological dimension of epistemic injustice that comes to light once its victims internalize their epistemic mistreatment and can no longer fully participate in the medical discourse in an epistemically substantial manner. This argument can be supported by considering how feminist literature problematizes certain attitudes women internalize due to the pressure to conform to established social standards. In "Throw-

ing Like a Girl", feminist philosopher Iris Marion Young tackles this issue by refuting the hypothesis that the different way girls throw a ball is biologically grounded, but rather compellingly relates to their internalization of gender-related social expectations (Young, 1980).

Similarly, when patients repeatedly find themselves in situations where healthcare professionals silence them in the face of a possibly inaccurate NarxCare score, and as a consequence, their testimonial offerings are dismissed and discredited as unworthy of attention, they can lose confidence in their capacities as conveyors of valuable information and knowledge. This can lead to their deskilling and ultimate loss of capacity to fully participate in medical interactions (Carel & Kidd, 2017). Also in this case, albeit due to different underlying mechanisms (i.e., a psychological one) compared to the *objectification constraint* previously elucidated, patients in practice are hindered in contributing to the epistemic discourse, and their moral reasons remain excluded from those available to the system, thus hampering the fulfillment of the tracking condition for MHC. Hostile communication patterns between clinicians relying on NarxCare risk scores and patients substantiate the likelihood of the AI instantiating what we have referred to as the capacity constraint. Hildebran et al. (2014, 2016) report that clinicians often close off communication with patients in the face of a problematic risk score or refrain from sharing this information in the first place. When confronted with situations in which their credibility is doubted due to an AI-generated risk score, it is easy to imagine that it might become increasingly difficult for patients to fight back against what is felt as an inaccurate portrayal of their medication consumption due to, for instance, the fear of being repeatedly turned down. These situations are conceivable in the face of the authoritative role that AI systems can play in medical interactions to the detriment of patients' epistemic standing and capacities. We ultimately maintain that power asymmetries created with the introduction of AI systems, particularly in cases similar to NarxCare, can fuel dynamics of self-silencing, effectively disempowering patients and their epistemic confidence in medical encounters (which, like in the case of pain management, can already be fraught with stigma and preconceptions related to possible drug misuse).

Let us now turn to the final constraint identified, the *self-silencing constraint*. Similarly to the previous one, this constraint also happens because patients themselves refrain from offering their testimony. However, unlike in cases pertaining to the *capacity constraint*, patients do not fail to contribute to medical interactions due to their inability to do so. Instead, they silence themselves through a sort of self-censoring, which is not necessarily to be tracked back to a failure to contribute to the discourse due to a loss of epistemic confidence and capacity (Dotson, 2011). This phenomenon amounts to what Dotson refers to as *testimonial smothering*. According to Dotson, instances of testimonial smothering amount to "the truncating of one's own testimony to ensure that the testimony contains only content for which one's audience demonstrates testimonial competence." (p. 244) While Dotson has in mind certain racial microaggressions perpetrated by a hearer to a speaker due to their social positioning, we maintain that this phenomenon can also take place in exchanges between patients and physicians. In the face of interactions with medical professionals in which patients perceive that the information clinicians seek to obtain is limited to what can be easily measured and quantified, patients might offer minimal testi-

mony pertaining to confirming or denying specific symptoms, thus restricting their role to epistemic agents in the minimal sense (Carel & Kidd, 2014). In doing so, they can refrain from sharing more substantial knowledge about their lived experience of disease. This can be the case because they realize that their testimony would remain unintelligible to clinicians because it would not be considered medically relevant, particularly if an AI system is part of the medical interaction as an opaque knowledge-generating entity about patients' health status. This form of self-censorship is driven by the fear that medical professionals would nullify the value of what patients are reporting, which might be connected to painful and intimate details of how they experience disease. So understood, we argue that these dynamics in clinical encounters can qualify as forms of testimonial smothering. Related to the NarxCare case, it is not too far-fetched to assume that patients might smother their own testimony under the assumption that the latter will not receive appropriate uptake in view of the decision-making power attributed to the AI system outputting their risk score.

Under this heading, we can conclude that this manifestation of epistemic injustice does not per se require the presence of a perpetrator that directly inflicts the injustice on the victim. So understood, this phenomenon is deeply ingrained in structural power dynamics that might be further reinforced by the introduction of AI systems like NarxCare in medical care. Testimonial smothering can happen even if patients are, in principle, capable of expressing their testimony. While related to the previous constraint, this expression of testimonial injustice acquires a further level of complexity because it entails the conscious decision to withhold one's testimony due to previous unjust testimonial encounters or the speaker's perceived inability (or lack of willingness) of their hearer to find their testimony intelligible. This problem cannot be solved at the individual level of analysis. It rather requires tackling underlying power dynamics and systemic inequalities in how knowledge and information are received and imparted in different social practices (Dotson, 2014).

From what has been said so far, it becomes clear that there is a dire need to consider underlying power dynamics when assessing the availability of relevant reasons to enable the fulfillment of conditions for MHC. In fact, we want technical systems that humans are meaningfully in control of and in which stakeholders' voices and relevant reasons, such as patients, are definitely not systematically and not pre-emptively excluded.

Let us also take notice of the fact that while in this section we have been focusing on the epistemic situation of patients (as intrinsically vulnerable epistemic subjects), clinicians and healthcare professionals, more generally, can also be epistemically disadvantaged by the introduction of AI systems in medical decision-making (Pozzi, 2023a). As pointed out in Sect. 3, systems like NarxCare algorithms may negatively affect the meaningful participation of physicians in the decision of whether to prescribe opioids to a patient who received a high risk score. Physicians are implicitly bound to follow the AI system's recommendation due to the regulatory setup in many US states (Leichtling et al., 2017). Therefore, it is unclear how much space they effectively have to express their own relevant moral reasons in the evaluation of a patient's health status and their possible need for pain medication. A recently published article expressing clinicians' worries related to the use of these AI systems substantiates these concerns: "As clinicians in this setting, we are concerned that

if our best judgment conflicts with fear of criminal liability, our ability to provide evidence-based, compassionate care to our patients may be compromised." (Buonora et al., 2024, p. 859).

Furthermore, the pressure not to challenge the decision of the AI system may discourage clinicians from claiming their expertise. Interestingly, while liability attribution should arguably be an incentive for (medical) professionals to take responsibility for their (clinical) decisions (also due to the fear of facing legal consequences in cases of medical misconduct), the introduction of AI may yield the opposite result. In fact, AI systems similar to NarxCare rather disincentivize doctors to take responsibility because the legal framework in which the technology is embedded entails that clinicians are 'on the safe side' by simply following the AI's recommendation. This shows that AI systems can bring to the table new forms of power imbalances in which otherwise epistemically privileged agents are also considerably constrained, albeit to a different extent and form (Pozzi, 2023a).

5 The Broader Picture: Addressing Control of AI Through Justice

The analysis advanced in the previous sections shows that introducing AI systems in medical decision-making raises new challenges related to the epistemic and moral status of relevant actors, particularly patients and medical professionals. From the NarxCare case, the need emerges to gear efforts toward safeguarding the physician-patient relationship in medical interactions mediated by AI systems. This starts by protecting physicians' moral accountability and active responsibility for the care relationship (Pozzi & Van den Hoven, 2023; Santoni de Sio & Mecacci, 2021). As for the patients, the need to focus on patients' practical ability to remain actively involved in the care relationship comes to the forefront. To this aim, mechanisms need to be in place that allow them to reclaim their position at the center of care and not lead them to be pushed to its margins due to the (epistemic) power that is often unduly attributed to AI within medical care.

Against this background, we maintain that framing the problem through the lens of MHC as reason-responsiveness allows us to formulate an important positive requirement to mitigate forms of epistemic injustice: *in the face of a restructuring of the medical system via the introduction of AI, we must protect and reinforce the responsiveness of these technologies to the reasons of the relevant agents.* While discussions on epistemic injustice predominantly spell out the nature of these injustices in a negative fashion, focusing on how to overcome the constraints highlighted in the previous section as a requirement to enable MHC provides an, albeit initial, but actionable way to mitigate those.

Moreover, while discussions revolving around MHC usually focus on the inadequate design of the human-machine interaction and/or lack of sufficient training or digital literacy of the users, we have highlighted a further component that has comparatively received limited attention. Through the analysis of different forms of epistemic injustice, we showed that social structures of power affecting the ability and willingness of clinicians and patients to have their reasons sufficiently reflected in the AI system's functioning is one relevant condition that may affect its capacity

to track their reasons (i.e., reduce the reason responsiveness of the AI system). This further helps redirect the attention of designers and developers concerned with securing MHC toward issues of epistemic injustice that could otherwise go unnoticed.

All in all, if, as we have argued, increasing MHC requires, among other things, reducing epistemic injustice and epistemic injustice is a form of social injustice grounded in social structures of power, then one crucial way to reduce epistemic injustice and increase MHC is a better understanding and addressing these social injustices. On the other hand, taking an MHC perspective may also allow us to take a more active attitude towards the creation of a more just AI-mediated healthcare system. The case in which relevant patients' voices and reasons are not sufficiently reflected and incorporated into the functioning of the AI system can be reframed as a case in which the introduction of AI systems reinforces and perpetuates forms of social exclusion and vulnerability. Patients affected by medical conditions *and* social exclusion are more likely to become unable to be seen as moral agents capable of contributing morally relevant reasons in the AI-mediated decision-making process. This is mainly reflected in the *capacity* and *self-silencing* constraints discussed in Sect. 4.

As for the objectification constraint, even though patients are not entirely excluded from the medical decision-making process, their participation happens only through the inclusion of some of their personal data (which are usually extracted in medical encounters in which they play a passive role) in the AI-assisted decision-making process, rather than enabling them to actively contribute to the deliberation and decision process (Pozzi & Durán, 2024b). In terms of MHC as reason-responsiveness, this is not sufficient for the system to track the patients' relevant reasons. The objectification constraint thus highlights a broader issue with justice and technology. In a nutshell, participation, even when optimally managed, does not guarantee justice or control. One could envisage optimal conditions for participation if all relevant stakeholders are involved in the deliberation process, they are all provided a (formal) opportunity to contribute to it, and the technology's designers do not have any explicit reasons to exclude any relevant stakeholder from participating. However, even in cases in which similar measures to increase participation as much as possible are taken, this does not ensure that conditions of justice and control are met due to underlying structural injustices. For instance, the self-silencing constraint previously mentioned shows how tacit and hard to spot certain expressions of epistemic injustice can be. Moreover, structural mechanisms of power imbalances might prevent people from expressing their reasons and/or being properly listened to. Also, stakeholders' reasons might effectively remain unacknowledged and excluded if they are expressed in terms that are not aligned with designers' (technical) language, or they may be discarded when they do not match designers' (implicit) problem-framing and other underlying assumptions.

In this sense, even progressive movements proposing to include so-called stakeholders in the technological design process, such as Value-Sensitive Design and Participatory Design, have been criticized for not sufficiently allowing stakeholders to present their reasons, crucially, in their own terms (Borning & Muller, 2012). Researchers, designers, and developers rather often utilize data and information gathered from them according to their own framing of the problem and to address their

own research or development agendas. The "design justice" movement (Costanza-Chock, 2020) has called these approaches "extractive" and has exposed them as a form of *unjust* participation of relevant stakeholders, which must be transformed.

At first sight, the case of epistemic injustice affecting clinicians can also be understood as an extension of traditional forms of social injustice, in this case, the oppression of the workforce by managers. In fact, it has been argued that the introduction of AI can be seen as a new and partly original chapter of an old story, that of the managerial control over the workforce mediated by technological systems. While the physical constraints experienced by the factory workers in the assembly line were the paradigmatic example of managerial control of the 20th century, the constraints and controls allowed for by AI systems may become the new tools for managerial control in the 21st century (Kellogg et al., 2020). Doctors complying with the "request" of the AI systems introduced by the management out of cost-efficiency reasons to avoid incurring the risk of being held liable or even fired seem to confirm this view of them as vulnerable employees subject to forms of managerial power mediated by technology that they cannot control. However, unlike what happens in other cases of employer-employee relations mediated by AI, what is violated here is not as much the doctors' rights as employees (e.g., their privacy, freedom, or psychological safety), but rather their (epistemic) agency as professionals, that is, their capacity to have their expertise matter in medical procedures.

As pointed out in Sect. 3, this issue can be best framed as a violation of a different kind of social justice; that is, in terms of political philosopher Michael Walzer, justice is the recognition and respect of the different values and principles that pertain to different "spheres of life" (Walzer, 1983).

The shift of values caused by AI technologies goes beyond the specific NarxCare case discussed in this paper. For instance, with specific reference to Walzer's theory, Tamar Sharon has recently spoken of the risk of the Googlization of healthcare research, caused by the increasingly prominent role of big technological companies with no previous experience in the healthcare sector or commitment to its traditional values, transforming the ethos of research in this domain with their mindset and practices (Sharon, 2016).

6 Prospects for Future Research and Concluding Remarks

MHC as reason-responsiveness is a context-dependent and normatively loaded concept. Its realization requires defining which relevant reasons of which relevant agents must be reflected in socio-technical systems that include AI systems. Medicine or healthcare covers a broad range of settings – diagnosis, treatment, surgery, rehabilitation, long-term care, and others – which may be integrated with different kinds of AI-driven technologies – decision-support systems, robots, exoskeletons, etc. In this paper, we have considered the specific case of an AI system used by doctors in the US as a decision-support system for the prescription of opioids (NarxCare algorithms), thus restricting our focus to ML applications implemented to make predictions for the purposes of patient treatment and the allocation of medical resources in the context of pain management. While we decided to focus exclusively on this particular case

to provide a detailed analysis of different mechanisms of epistemic injustice and how they are intertwined with issues of control, we think that these considerations can also be transferred to other cases. For instance, it can be plausibly anticipated that overfocusing on cost-efficiency-driven metrics might trump consideration of justice also in other healthcare applications, thus making our analysis a relevant starting point for critically scrutinizing possibly hard-to-notice issues of epistemic injustice and (lack of) control. While this paper aims to provide an initial theoretical underpinning to issues of justice and control in medical AI, further research is needed to apply these relevant considerations to other medical AI systems, tailoring them to their specificities.

Related to the NarxCare case, we have tentatively identified doctors and patients as the primary human agents involved and discussed one specific issue: to what extent the current AI system (fails to) track their relevant reasons (a condition for MHC) *due to epistemic injustice*. Future multidisciplinary research, including ethical, ethnographic, sociological, technical, and other, must develop concrete proposals to rethink the development and use of decision-support systems in the medical context in a way that promotes epistemic justice and MHC. A similar work must be done in relation to all other domains and technical applications.

The analysis of the epistemic injustice suffered by patients and professionals advanced in this paper suggests two further general considerations for future research. First, we have argued that MHC and epistemic justice go hand in hand since the former requires relevant agents – patients, healthcare professionals, and others – to be able to contribute with their relevant knowledge, experiences, and skills to the development and functioning of a socio-technical system. So understood, epistemic justice is an important enabler for MHC towards the fulfillment of the tracking condition.[10] From this claim follows that any future research in the development of medical technology must devise methodologies that assess to what extent these agents must not only be involved in the process by providing some data or information but by being more actively involved from a stronger position of epistemic power in the decisions and processes regarding the design, development, introduction, and use of new technologies. This is a big challenge that requires a radical rethinking of the research and innovation process and precedes concerns regarding how to guarantee shared decision-making in medical practices mediated by AI systems (Bjerring & Busch, 2021). Structural approaches that look at how to achieve epistemic justice at the institutional level are currently present in standard debates in social epistemology (Anderson, 2012). However, efforts along similar lines need to be explicitly tailored to face the challenges brought about by new technologies, for instance, by developing and applying the idea of "design justice" (Costanza-Chock, 2020).

Second, and relatedly, such a radical transformation requires specific social, economic, and political conditions. Social spaces and economic opportunities for experimenting with new innovation models must be created and protected, ideally through the collaboration of public universities, private industries, and local com-

[10] This is in line with what Santoni de Sio (2024) calls a precondition for MHC as reason-responsiveness.

munities.[11] Long-term research programs must be established and funded. Dominant visions about a future in which all problems will be solved by (AI) technology need to be resisted and contrasted, and the social and economic power of the big technological and political players supporting and pushing these limited visions must be regulated and governed. As the NarxCare case discussed in this paper shows, even the relatively small challenge of protecting justice and MHC in the use of one specific technological system cannot be addressed without considering the broader social, cultural, economic, and political conditions under which it is designed, developed, and introduced. Future research must also address this higher level of analysis and study the conditions under which new, more just forms of technological innovations in the medical sector can be realized.

Acknowledgements We would like to express our deepest gratitude to Deborah Forster, Patrik Hummel, Giulio Mecacci, and Philip Nickel for their extremely helpful comments on an earlier version of this paper. Their feedback helped us considerably to improve the quality of this work.

Author Contributions The authors contributed equally to the manuscript.

Funding Not applicable.

Data Availability Not applicable.

Declarations

Conflict of interest The authors confirm that there are no competing interests to disclose.

References

Alvarado, R. (2023). AI as an epistemic technology. *Science and Engineering Ethics*, *29*(5). https://doi.org/10.1007/s11948-023-00451-3

Anderson, E. (2012). Epistemic justice as a virtue of social institutions. *Social Epistemology*, *26*(2), 163–173. https://doi.org/10.1080/02691728.2011.652211

Beck, S., Gerndt, S., & Samhammer, D. (2024). Meaningful human control in shared medical decision making. In G. Mecacci, D. Amoroso, L. Cavalcante Siebert, D. Abbink, J. Van den Hoven, & F. Santoni de Sio (Eds.), Research Handbook on Meaningful Human Control of Artificial Intelligence Systems (pp. 131–147).

[11] This complex multidisciplinary and participatory approach to research and design is often referred to as "transdisciplinary" (see, e.g., Zaga et al., 2024).

Benjamin, R. (2019). Assessing risk, automating racism. *Science*, *366*(6464), 421–422. https://doi.org/10.1126/science.aaz3873

Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and Patient-Centered Decision-Making. *Philosophy & Technology*, *34*, 349–371. https://doi.org/10.1007/s13347-019-00391-6

Borning, A., & Muller, M. (2012). Next Steps for Value Sensitive Design. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1125–1134.

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2021). Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics*, *47*(12), e3. https://doi.org/10.1136/medethics-2019-105860

Buonora, M. J., Axson, S. A., Cohen, S. M., & Becker, W. C. (2024). Paths forward for clinicians amidst the rise of unregulated clinical decision support software: Our perspective on narxcare. *Journal of General Internal Medicine*, *39*(5), 858–862. https://doi.org/10.1007/s11606-023-08528-2

Carel, H., & Kidd, I. J. (2014). Epistemic injustice in healthcare: A philosophial analysis. *Medicine Health Care and Philosophy*, *17*(4), 529–540. https://doi.org/10.1007/s11019-014-9560-2

Carel, H., & Kidd, I. J. (2017). Epistemic injustice in medicine and healthcare. In I. J. Kidd, J. Medina, & G. Pohlhaus (Eds.), *The Routledge handbook of epistemic injustice* (pp. 336–346). Routledge.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G. J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI and Ethics*, *3*(1), 241–255. https://doi.org/10.1007/s43681-022-00167-3

Chiarello, E. (2021). Pharmacists should treat patients who have opioid use disorders, not Police them. *Journal of the American Pharmacists Association*, *61*(6), e14–e19. https://doi.org/10.1016/j.japh.2021.06.019

Costanza-Chock, S. (2020). *Design justice: Community-Led practices to build the worlds we need*. The MIT Press.

Delcher, C., Harris, D. R., Park, C., Strickler, G. K., Talbert, J., & Freeman, P. R. (2021). Doctor and pharmacy shopping: A fading signal for prescription opioid use monitoring? *Drug and Alcohol Dependence*, 221. https://doi.org/10.1016/j.drugalcdep.2021.108618

Dotson, K. (2011). Tracking epistemic Violence, tracking practices of Silencing. *Hypatia*, *26*(2), 236–257. https://www.jstor.org/stable/23016544?seq=1&cid=pdf-

Dotson, K. (2014). Conceptualizing epistemic oppression. *Social Epistemology*, *28*(2), 115–138. https://doi.org/10.1080/02691728.2013.782585

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Fischer, J., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Fricker, M. (2007). *Epistemic Injustice. Power & the ethics of knowing*. Oxford University Press.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics*, *83*(5), 368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Haselager, P., & Mecacci, G. (2024). Reflection machines and the proximity scale of reasons: Addressing accountability asymmetry. In G. Mecacci, D. Amoroso, L. Cavalcante Siebert, D. Abbink, Van den J. Hoven, F. Santoni de, & Sio (Eds.), *Research handbook on meaningful human control of artificial intelligence systems* (pp. 28–37). Edward Elgar Publishing Limited.

Hildebran, C., Cohen, D. J., Irvine, J. M., & Foley, C. (2014). & others. How Clinicians Use Prescription Drug Monitoring Programs: A Qualitative Inquiry. Pain Medicine, 15(7), 1179–1186. https://academic.oup.com/painmedicine/article/15/7/1179/1878292

Hildebran, C., Leichtling, G., Irvine, J. M., & Cohen, D. J. (2016). & others. Clinical Styles and Practice Policies: Influence on Communication with Patients Regarding Worrisome Prescription Drug Monitoring Program Data. Pain Medicine, 17(11), 2061–2066. https://doi.org/10.1093/pm/pnw019

Hille, E. M., Hummel, P., & Braun, M. (2023). Meaningful human control over AI for health? A review. *Journal of Medical Ethics*. https://doi.org/10.1136/jme-2023-109095

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, *14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Kersbergen, A. (2000). Managed care shifts health care from an altruistic model to a business framework. *Nursing and Health Care Perspectives*, *21*(2), 81–83.

Khera, R., Simon, M. A., & Ross, J. S. (2023). Automation bias and assistive AI risk of harm from AI-Driven clinical decision support. *International Journal of Medical Informatics*, *330*, 2255. https://pubmed.ncbi.nlm.nih.gov/38112824/

Kidd, I. J., & Carel, H. (2017). Epistemic injustice and illness. *Journal of Applied Philosophy*, *34*(2), 172–190. https://doi.org/10.1111/japp.12172

Kilby, A. (2021). Algorithmic fairness in predicting opioid use disorder using machine learning. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness Accountability and Transparency*, *272*. https://doi.org/10.1145/3442188.3445891

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Leichtling, G. J., Irvine, J. M., Hildebran, C., Cohen, D. J., Hallvik, S., & Deyo, R. (2017). Clinicians' use of prescription drug monitoring programs in clinical practice and Decision-Making. *Pain Medicine*, *18*(6), 1063–1069. https://doi.org/10.1093/pm/pnw251

Lorenzini, G., Arbelaez Ossa, L., Shaw, D. M., & Elger, B. S. (2023). Artificial intelligence and the doctor–patient relationship expanding the paradigm of shared decision making. *Bioethics*, *37*(5), 424–429.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*, 175–183. https://doi.org/10.1007/s10676-004-3422-1

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, *45*(3), 156–160. https://doi.org/10.1136/medethics-2018-105118

Mecacci, G., & Santoni de Sio, F. (2020). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology*, *22*(2), 103–115. https://doi.org/10.1007/s10676-019-09519-w

Mecacci, G., Amoroso, D., Cavalcante Siebert, L., Abbink, D., Van den Hoven, J., & Santoni de Sio, F. (Eds.). (2024). *Research handbook on meaningful human control of artificial intelligence systems*. Edward Elgar Publishing.

Mittelstadt, B. D., Allo, P., Taddeo, M., & Wachter, S. (2016). & others. The ethics of algorithms: Mapping the debate. Big Data and Society, 3(2), 1–21. https://doi.org/10.1177/2053951716679679

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting Racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Oliva, J. D. (2022). Dosing discrimination: Regulating PDMP risk scores. *California Law Review*, *110*(1), 47–115. https://doi.org/10.15779/Z38Z31NP8J

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: Implications for health systems. *Journal of Global Health*, *9*(2). https://doi.org/10.7189/jogh.09.020318

Pozzi, G. (2023a). Automated opioid risk scores: A case for machine learning-induced epistemic injustice in healthcare. *Ethics and Information Technology*, *25*(1). https://doi.org/10.1007/s10676-023-09676-z

Pozzi, G. (2023b). Testimonial injustice in medical machine learning. *Journal of Medical Ethics*, *49*, 536–540. https://doi.org/10.1136/jme-2022-108630

Pozzi, G., & Durán, J. M. (2024a). Social causes and epistemic (in) justice in medical machine Learning-Mediated medical practices. In I. Phyllis, & F. Russo (Eds.), *The Routledge handbook of causality and causal methods* (pp. 178–189). Routledge.

Pozzi, G., & Durán, J. M. (2024b). From ethics to epistemology and back again: Informativeness and epistemic injustice in explanatory medical machine learning. *AI & SOCIETY*. https://doi.org/10.1007/s00146-024-01875-6

Pozzi, G., & Van den Hoven, J. (2023). Physicians' professional role in clinical care: AI as a change agent. *The American Journal of Bioethics*, *23*(12), 57–59. https://doi.org/10.1080/15265161.2023.2272924

Robbins, S. (2023). The many meanings of meaningful human control. *AI and Ethics*, *4*, 1377–1388. https://doi.org/10.1007/s43681-023-00320-6

Ross, C. (2021). Epic's sepsis algorithm is going off the rails in the real world. The use of these variables may explain why. STAT. https://www.statnews.com/2021/09/27/epic-sepsis-algorithm-antibiotics-model/

Salloch, S., & Eriksen, A. (2024). What are humans doing in the loop? Co-Reasoning and practical judgment when using machine Learning-Driven decision aids. *The American Journal of Bioethics*, 1–12. https://doi.org/10.1080/15265161.2024.2353800

Santoni de Sio , F. (2024). *Human freedom in the age of AI*. Routledge.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*, 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 1–14. https://doi.org/10.3389/frobt.2018.00015

Sharon, T. (2016). The googlization of health research: From disruptive innovation to disruptive ethics. *Personalized Medicine*, *13*(6), 563–574. https://doi.org/10.2217/pme-2016-0057

Siegel, Z. (2022). In a world of stigma and Bias, can a computer algorithm really predict overdose risk? *Annals of Emergency Medicine*, *79*(6), A16–A19. https://doi.org/10.1016/j.annemergmed.2022.04.006

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/

Szalavitz, M. (2024). Say Hello to Your Addiction Risk Score-Courtesy of the Tech Industry. New York Times. https://www.nytimes.com/2024/04/20/opinion/addiction-risk-score-avertd-narxcare.html

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

van Wynsberghe, A., & Li, S. (2019). A paradigm shift for robot ethics: From HRI to human–robot–system interaction (HRSI). *Medicolegal and Bioethics*, *9*, 11–21. https://doi.org/10.2147/mb.s160348

Walzer, M. (1983). *Spheres of Justice. A defense of pluralism and equality*. Basic Books.

Young, I. M. (1980). Throwing like a girl: A phenomenology of feminine body comportment motility and spatiality. *Human Studies*, *3*, 137–156.

Zaga, C., Matos-Castaño, J., & van der Voort, M. (2024). Transdisciplinarity: Taking stock beyond buzzwords and outlining an agenda for design research, in Gray, C., Hekkert, P., Forlano, L., Ciuccarelli, P. (Eds.), *DRS2024: Boston*, 23–28 June, Boston, USA. https://doi.org/10.21606/drs.2024.1566