AI & SOCIETY
https://doi.org/10.1007/s00146-024-01875-6

MAIN PAPER

From ethics to epistemology and back again: informativeness and epistemic injustice in explanatory medical machine learning

Giorgia Pozzi[1] · **Juan M. Durán**[1]

Received: 12 September 2022 / Accepted: 10 January 2024

Abstract
In this paper, we discuss epistemic and ethical concerns brought about by machine learning (ML) systems implemented in medicine. We begin by fleshing out the logic underlying a common approach in the specialized literature (which we call the *informativeness account*). We maintain that the informativeness account limits its analysis to the impact of epistemological issues on ethical concerns without assessing the bearings that ethical features have on the epistemological evaluation of ML systems. We argue that according to this methodological approach, epistemological issues are *instrumental* to and *autonomous* of ethical considerations. This means that the informativeness account considers epistemological evaluation uninfluenced and unregulated by an ethical counterpart. Using an example that does not square well into the *informativeness account*, we argue for ethical assessments that have a substantial influence on the epistemological assessment of ML and that such influence should not be understood as merely informative but rather regulatory. Drawing on the case analyzed, we claim that within the theoretical framework of the informativeness approach, forms of epistemic injustice—especially *epistemic objectification*—remain unaddressed. Our analysis should motivate further research investigating the regulatory role that ethical elements play in the epistemology of ML.

Keywords Epistemology and ethics of ML · Informativeness · Medical ML · Explanatory ML · Epistemic injustice

1 Introduction

Artificial intelligence (AI) plays an increasingly morally relevant role throughout various domains, bearing the potential to significantly influence crucial decision-making processes that are usually reserved for human expertise. There are studies showing that AI-based methodologies, in particular machine-learning (ML) techniques, are paving the way for promising developments in high-stakes fields, such as medicine and healthcare (e.g., Topol 2019; Esteva et al. 2019).

Unfortunately, the excitement associated with these developments is not always justified. Epistemic limitations in connection with the way in which these systems operate give rise to serious ethical concerns. Central to the success of ML systems is their capacity to reconstruct sets of rules from large datasets, which in turn can reveal new patterns in the data (Alpaydin 2014). Due to the large amount of data processed by these systems and the complexity of the calculations, they become epistemically opaque to human enquirers (Humphreys 2009; Durán and Formanek 2018; Beisbart 2021).[1]

The consideration of how the epistemic limitations of ML systems lead to ethical issues has, justifiably, gained a central stage in current debates and has given rise to a wealth of literature on the topic. For example, scholars have pointed out that the epistemic opacity of ML algorithms is connected to ethically relevant problems that range from fairness-based concerns (e.g., Zarsky 2016) to questions of accountability (e.g., De Laat 2018) and the trust we are justified in attributing to these systems' outputs (e.g., Hatherley 2020). Other authors connect the epistemology and ethics of AI even more explicitly. For instance, Grote and Berens (2020) identify the epistemological pitfalls of ML systems implemented in medicine (e.g., issues of peer disagreement and epistemic

✉ Giorgia Pozzi
G.Pozzi@tudelft.nl

Juan M. Durán
J.M.Duran@tudelft.nl

[1] Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands

[1] In a less nuanced and more metaphorical way, these algorithms are known as "black-boxes".

Published online: 28 February 2024

uncertainty) as directly conducive to crucial ethical implications (e.g., problems of paternalism, patients' informed consent, and defensive medicine). Similarly, Bjerring and Busch (2021) recognize in the black-box nature of AI systems the concrete possibility that it undermines the ethical ideal of patient-centered medicine. Relatedly, Babushkina and Votsis (2022) consider primarily how epistemological constraints of ML systems in the context of medical diagnoses lead to ethical considerations in terms of epistemic responsibility.

Thus, the relevance of showing the bearings of epistemological issues on ethical concerns has been recognized and extensively analyzed. However, an analysis of the extent to which ethical considerations influence the epistemology of ML is still lacking, as Russo et al. (2023) point out in a recent paper. In addressing this research gap, these authors take an approach that aims to *explicitly* point out the interconnected nature of the ethics and epistemology of AI. They do so without presupposing that epistemological considerations are prior to ethical considerations, as is often assumed in the debate.[2] This is a position that we endorse in this paper. We share these authors' approach in seeing "the equal importance of the two fields [i.e., of ethics and epistemology] and their intertwinement" (ibid., 2). We are, in fact, committed to the same goal of overcoming a division in the ethics and epistemology of AI that is not tenable if we want to understand the impact of these technologies on society. However, our approach differs from theirs in at least three respects.

First, Russo et al. take a holistic approach that accounts for "the process of design, implementations, and assessment of AI that simultaneously considers ethics and epistemology, and the expertise of the actors that inquire into these two" (Russo et al. 2023, 2). We take a more fine-grained level of analysis than this high-level approach. In fact, we analyze the role that the intertwined dimensions of epistemology and the ethics of AI should play in a very concrete setting, that is, one in which the medical decision-making of an ML system is analyzed regarding its displacement of physicians from their epistemically authoritative position.

Second, by analyzing a concrete case, we provide more reasons as to why the available approach in the literature is limited. In Sect. 2, we dissect it in its parts, analyze its underlying logic, and show its shortcomings. In this sense, our analysis considerably expands on one of the fundamental premises made by Russo and colleagues—that is, the ethics and epistemology of AI are largely disconnected in the current debate.

Finally, in our analysis, we consider the epistemology of AI as a genuinely and inherently normative dimension and place a strong emphasis on this point throughout the entire article. This comes particularly to light in our concrete case analysis in Sect. 3, which underscores that what a physician *should* believe and the explanation she *should* accept is partially determined by an ethical feature of the particular situation in focus.[3] In making a distinction between epistemic (e.g., explainability) and normative (e.g., fairness) aspects, Russo et al. (2023, 10) do not seem to embrace this aspect, which is, however, central to our analysis.

We are convinced that the role that ethical features play in the epistemology of AI needs special attention, and it is the overall aim of this paper to lay the groundwork for a more explicit discussion of this important aspect. To effectively show the relevance of our argumentative goals, some considerations are in order, starting from the kind of ML systems that are the object of our analysis.

In this contribution, we focus on ML systems that displace or risk displacing physicians from the center of knowledge production. Here, their courses of action are dependent upon those indicated by the ML system involved in the decision-making process. Even though this scenario is surely undesirable, as we would expect these systems to remain under the ultimate control of experienced professionals—and particularly for ML systems implemented in medicine—it is, unfortunately, not too far-fetched. As we will show, some currently deployed ML systems dramatically disappoint this expectation. For example, algorithmic Prediction Drug Monitoring Programs (PDMPs) used to predict patients' likelihood of opioid misuse and currently implemented in the USA to inform clinician's decisions on a daily basis have been shown to be de facto replacing—instead of merely supporting—medical decision-making (cf. Szalavitz 2021; Oliva 2022). Furthermore, these ML platforms are opaque to their end users (i.e., physicians) in that they lack insight into how the algorithms classify patients as being at a high risk of opioid abuse (Szalavitz 2021). Lastly, the proxies used to determine patients' risk scores are not necessarily indicative of opioid misuse and can result in misleading ML outputs that do not represent a patient's actual drug consumption (Oliva 2022). Given that these are "law enforcement-developed digital surveillance systems" (ibid., 51), physicians are expected to act upon the outcomes generated, even though they lack any kind of understanding regarding how the system's results are obtained. In fact, due to these constraints, physicians are in no position to determine whether a patient is justifiably considered at risk of drug misuse or whether the systems establish disparate correlations that are not reliably connected to a person's drug consumption (Oliva 2022; Pozzi 2023a, b). Thus, these

[2] As we point out later in the paper, this is one of the limitations we recognize in the approach available in the literature.

[3] We clarify the nature of the moral and epistemic "should" and their relation in Sect. 3.

systems are incontestable and are clearly displacing physicians from their epistemic and moral authority, creating undesirable effects that have led to patient abandonment and denial of medication (Szalavitz 2021).

In the face of the harm that epistemically authoritative systems similar to ML-based PDMPs can generate, which theoretical approach can be functional in effectively addressing the epistemic and moral issues they bring about? This question motivates our analysis of the relationship between the epistemology and ethics of ML. We label approaches that consider the bearing of epistemological issues on ethical concerns but neglect the impact of ethical elements on epistemic features of situations involving ML as the *informativeness account*. We elaborate on the assumptions built into this account and analyze an example in the field of explanatory ML in healthcare that does not square well into it. We argue that in cases similar to the one under scrutiny, it is paramount to consider the role that ethical properties play in influencing and regulating epistemologically relevant aspects of ML (e.g., explanatory ML). We dedicate the main part of this contribution to the effort to make explicit the compelling nature of this claim.

With these considerations, we gain a purchase on how certain epistemic practices with ML in medicine (such as the ones illustrated in our case in Sect. 3) expose patients to diverse forms of epistemic injustice. We are particularly interested in showing how, following the logic of the informativeness account, ML algorithms *epistemically objectify* patients. The section on epistemic injustice aims to further substantiate the claim that we need an approach in the ethics and epistemology of ML that considers the impact of ethics on epistemology. Although these considerations strongly suggest the need to expand the informativeness approach, it is beyond the scope of this article to show how this is effectively done.

The remainder of this article proceeds as follows. We provide a description of what we define as the *informativeness account* (Sect. 2). We then substantiate our case through an example of explanatory ML in medicine that cannot be adequately accepted when considered within the framework of the informativeness account (Sect. 3). Finally, we consider how the situation experienced by the patient in our example leads to a case of epistemic injustice understood in terms of epistemic objectification (Sect. 4).

2 Defining informativeness

To advance claims regarding the suitability of a merely informative approach, we deem it useful to zoom out from the analysis of specific issues and consider the logic underlying the general relationship between the epistemology and ethics of ML, as it has been treated so far. To achieve this goal, we consider an often-cited overview of the debates revolving around the epistemology and ethics of ML, an article published a few years ago by Mittelstadt and colleagues: "The ethics of algorithms: Mapping the debate" (Mittelstadt et al. 2016).[4] This article has justifiably served as the basis for much good research on the epistemology and ethics of ML, providing a systematic organization of an otherwise fragmented debate. Although, on the one hand, we acknowledge the value of the contribution provided by these authors, on the other hand, we want to complement this general approach by taking into consideration specific aspects pertinent to the debate that have not been considered by the authors. We scrutinize this review because we see it as particularly clearly illustrating a general approach taken in the literature that is characterized by considering the epistemology of ML as serving an *informative* role in the ethics of ML.[5] To substantiate this claim, we dive deeper into Mittelstadt et al.'s (2016) contribution and analyze its underlying logic.

In their mapping review, Mittelstadt and colleagues provide a conceptual map that allows for the identification of ethical challenges related to the use of decision-making algorithms whose inner logic is cognitively inaccessible to humans. They "are interested in algorithms whose actions are difficult for humans to predict or whose decision-making logic is difficult to explain after the fact" (Mittelstadt et al. 2016, 3). To this category belong, among others, clinical decision support systems (CDSS) that recommend diagnoses and treatments to physicians in the field of healthcare (Morley et al. 2020). Following Mittelstadt et al.'s conceptual

[4] The approach taken by these authors has been restated and substantiated through more updated literature.

In a recent publication by Tsamados et al. (2021). In the latter article, the methodology adopted by Mittelstadt and colleagues in analyzing the relationship between epistemology and ethics in ML has been kept unchanged (ibid., p. 2). Moreover, Morley et al. (2020) recently provided a mapping review of the ethics of ML in healthcare, also adopting the methodology developed by Mittelstadt et al. (2016). Since we are interested in discussing and building upon the approach considered by these authors in accounting for ethical and epistemological issues, we mostly keep referring to Mittelstadt's contribution throughout this paper.

[5] As previously pointed out in the first part of this introduction, examples of approaches in the literature that recognize the bearing of epistemological issues on ethical concerns (but not the other way around, i.e., the bearing of ethical properties on epistemic matters) abound. We decide to consider, specifically, the approach advanced by Mittelstadt and colleagues because, from our perspective, it illustrates at best the dichotomy existing between epistemological and ethical aspects of ML in the general debate. This makes it more immediate to effectively show the extent to which approaches that investigate only the bearing of epistemological features on ethical concerns are limited in important ways. This does not imply that this paper's approach is the only one that can be labeled *informative* according to our definition.

map, the authors identify six different types of ethical and epistemological concerns raised by algorithmic mediation in decision-making processes. Three are classified as epistemic (i.e., inconclusive evidence, inscrutable evidence, and misguided evidence), two as normative (i.e., unfair outcomes and transformative effects), and a sixth (i.e., traceability) is understood as an overarching concern that, it is argued, can neither be considered entirely epistemic nor entirely normative (Mittelstadt et al. 2016, 4–5).[6] We will now show that their analysis of the general relationship between epistemology and ethics develops exclusively on the information-serving level.

There are two main dimensions that, as we see it, characterize what we define as the informativeness account, and that can be recognized in the approach underlying the authors' methodology in mapping the debate. That is, epistemological claims about algorithms are (1) *instrumental* to and (2) *autonomous* of ethical considerations.[7] Let us discuss each one in turn.

To see what we mean by *instrumentality*, consider the analysis of the first three kinds of epistemic concerns advanced by Mittelstadt and colleagues, which predicate the quality of the output (O) produced by ML algorithms. These are inconclusive evidence, inscrutable evidence, and misguided evidence (Mittelstadt et al. 2016, 4). The authors' analysis includes showing how these epistemological shortcomings, such as a lack of certainty in O, lead to ethical concerns related to O (ibid., 4). For example, epistemic limitations due to the difficulty of knowing whether connections within datasets are causal (or merely correlational) and the inaccessibility to the connection between the processed data and the conclusion reached by the algorithm lead to concerns about the (lack of) moral justification of actions taken in response to possibly inconclusive outcomes (ibid., 5).

A similar approach is taken in considering how identifying epistemic limitations, understood as the inscrutability of the evidence produced by ML algorithms, leads to ethical problems. The latter are related to, for instance, meaningful consent to data processing and how algorithmic opacity affects the autonomy of data subjects (ibid., 6–7). The authors also point out that a lack of transparency in how these algorithms operate can lead to a loss of trust from the side of lay data subjects in ML systems and in data controllers (ibid., 7). The same method of analysis also applies to the consideration of what they define as misguided evidence, i.e., the fact that due to technical constraints or flaws in the data that are unintentionally taken up by the algorithm. That is, biased outcomes can be traced back to epistemic limitations that characterize how ML algorithms operate.

From the reconstruction of the first part of Mittelstadt et al.'s conceptual map (ibid., 4), it becomes clear that the analysis and assessment of the epistemology are understood as being prior to claims regarding the ethical acceptability of the outputs of ML systems. In fact, epistemic limitations understood in terms of inconclusive, inscrutable, and misguided evidence not only temporally precede the recognition of ethical issues but are also taken as the very source of these concerns and as instrumental to their identification.

Thus understood, in Mittelstadt et al.'s analysis of epistemic concerns, the ethical assessment of ML is strongly related to and dependent upon its epistemological merits. As previously pointed out, this is a legitimate assumption underlying Mittelstadt et al.'s analysis. Unfortunately, the same dependence cannot be recognized in their assessment of the epistemology of ML, which remains decoupled from ethical considerations in Mittelstadt et al.'s approach. Relatedly, the second dimension that characterizes Mittelstadt et al.'s analysis is the visible degree of *autonomy* of the epistemological treatment of ML with respect to ethical assessments. The dimension of autonomy comes to light in the second part of their conceptual map, that is, the one related to normative concerns and "based on how algorithms process data to produce evidence and motivate action" (Mittelstadt et al. 2016, 4).

In particular, in their assessment of unfair outcomes, the authors leave out the consideration of epistemological factors altogether (implicitly assuming the suitability of the epistemology), stating that the "ethical evaluation of algorithms can also focus solely on the *action* itself" (ibid., 5). Here, the epistemology of ML no longer fulfills

[6] Since we want to explicitly address the relationship between ethics and epistemology as considered in the approach taken by Mittelstadt and colleagues, the consideration of traceability as an overarching ethical concern exceeds the purpose of our analysis. In fact, even though questions regarding responsibility attribution of actions in response to ML systems' outputs are of great importance, it is not our aim to discuss this problem in this contribution. Rather, we focus on the two parts of Mittelstadt's conceptual map (Mittelstadt et al. 2016, p. 4) in which both epistemic and normative concerns are explicitly addressed, since there we can most effectively show the informative nature of their general approach. It is, however, true that traceability could also be understood as an epistemic issue leading to an ethical concern. That is to say, the difficulty of accessing the inner workings of ML algorithms (an epistemological issue) constrains the possibilities of responsibility attribution (an ethical concern). Nevertheless, we limit our analysis to the parts of Mittelstadt and colleagues' map that they explicitly recognize as being either ethical or epistemological in nature. We thank an anonymous reviewer for suggesting this possible reading of Mittelstadt et al.'s traceability problem that further supports our interpretation in terms of an informative relation.

[7] Let us note that these two dimensions are not mutually exclusive. In fact, we take that instrumentality applies exclusively to the analysis advanced by the authors in the first part of their conceptual map (Mittelstadt et al. 2016, p. 4) (i.e., the one addressing *epistemic concerns*), while autonomy applies exclusively to the second part of the same map (i.e., the one addressing *normative concerns*). Whereas we see instrumentality as unproblematic, we consider autonomy to be the aspect of their approach that needs to be abandoned to enable a regulatory approach. We will make a case for this claim in Sect. 3.

an instrumental role; rather, it is completely left unconsidered and disconnected from the ethical analysis. The same applies to their analysis of transformative effects, in which the authors investigate the impact of algorithmic decision-making in terms of how they affect the autonomy of data subjects and the changes they cause to our understanding of privacy and to the concept of personal identity (ibid., 9–10). For example, Tsamados et al. (2021) point out that the increasing use of profiling algorithms substantially limits the control that data subjects have over their own information. The fact that users are unaware of how their data are processed can contribute to a decreasing level of personal autonomy (ibid., 9). This analysis is highly relevant in pointing out non-obvious ethical concerns related to how ML algorithms reshape our self-understanding and the way we perceive and interact with the world.

However, zooming out from the relevance of the particular issues addressed, it can be said that in analyzing the general relationship between the ethics and epistemology of ML, ethical considerations are treated as partly disconnected from epistemological issues since the former cannot influence the epistemic features of a given ML. Indeed, at no point in Mittelstadt et al.'s analysis of transformative effects do the authors refer back to the epistemology of ML systems, nor do they advance claims regarding the role that ethical considerations should play in regulating it to avoid the ethical issues they discuss.

Drawing on the consideration of these two dimensions, the epistemological treatment of ML emerges as either instrumental to its ethical assessment or autonomous from it. Thus, the epistemological assessment of O fulfills the informative role of identifying the scope and merits of different ethical concerns. The contrary—that is, including ethical considerations in epistemological assessments of ML—seems to be missing from the framework of the relevant literature they analyze in their mapping review. We will make such a view a centerpiece in this paper, showing, in the next section, the limitations of the informativeness account and the need for an approach capable of accounting for the conflating nature of the epistemology and ethics of ML. To make these considerations more graspable, in the next section, we zoom into specific issues that derive from the application of the logic underlying the informativeness approach to a concrete case.

Although the informativeness approach is correct in many respects and, as we pointed out in the previous section, is indeed the approach that has been mostly endorsed in the literature, the epistemological and the ethical assessment of ML emerges as partly decoupled. Instead, we aim to show that this way of seeing the relationship between the epistemology and the ethics of ML sidelines two central aspects. First, the fact that ethical features also exert influences on the epistemological counterpart. Second, and perhaps more importantly, this influence is not merely informative but regulatory of the epistemology of ML to the extent that an ethical feature of the situation should lead to, on occasion, the re-evaluation of central epistemic functions such as explanation.[8] With this analysis, we aim to point out some difficulties that emerge in connection with the general approach to the epistemology and ethics of ML. In particular, we intend to show the necessity of seeing the epistemology and ethics of ML as substantially intertwined, thereby reaffirming their mutually regulatory role.

Now that we have provided a brief characterization of the informativeness account by analyzing the logic of the general approach taken by Mittelstadt and colleagues according to the two dimensions identified, in the next section, we take into consideration an example of explanatory medical ML that does not square well into this general approach. This should be functional to show in a more tangible and compelling way the need to expand and build upon the informativeness approach, accounting for the fact that the ethics and epistemology of ML are not to be considered compartmentalized dimensions.

3 Beyond informativeness: a case for explanatory medical ML

In what follows, we focus our efforts on showing the shortcomings of the logic underlying the informativeness account as we reconstructed it in the previous section. This allows us to argue for the need to consider the mutually regulatory role of epistemological and ethical features of situations mediated by a medical ML.

When confronted with the output of an ML system, the human inquirer is prompted to form beliefs about the empirical world.[9] These beliefs are intended to be associated with and populate our system of knowledge and understanding of the world, broadly conceived. To see how these beliefs are formed, consider *med*+*ML*, a cancer-detection system that renders as output the following explanation: "the chances for melanoma for patient X are 89% given the analysis of the following characteristics: the image shows a mole that is 98% asymmetrical; the image shows a mole that is 8mm long (< 6mm considered no melanoma), etc." (Esteva et al. (2019) present such a system). Should the physician believe

[8] We make clear what we mean by a regulatory relation in the next section of the paper.

[9] An anonymous reviewer rightly pointed out that the human inquirer could also simply suspend their judgment. Even though we acknowledge this possibility, we consider the more relevant case in which a decision needs to be made following (or not) an ML output. This entails that beliefs need to be formed connecting the ML output with the empirical world to render the former actionable.

this output, then *med*+*ML* has induced a specific belief about the patient's mole, namely that it is carcinogenic. Let us note that this belief is based on specific biological markers about the patient that *med*+*ML* detects and analyses. These markers, along with any explanation of how they are obtained, populate the physician's body of knowledge about the patient's medical condition, potential treatments, and prognosis. Let us also note that the output of *med*+*ML* might also induce moral beliefs about the most suitable medical action,[10] the general principles that the physician must follow, and the like. In fact, based on the biological markers measured, the physician forms a moral judgment that will guide her actions: the best treatment for this patient is surgery, chemotherapy, or something else. Naturally, these decisions are not exclusively made by physicians but also depend on the values upheld by the medical department, the hospital, and the national health service.[11]

3.1 On the mutual dependence of the epistemology and ethics of ML

To illustrate the limitations of the informativeness account, consider the following situation for *med*+*ML*. After analyzing large amounts of data pertaining to a given patient *p*, along with other relevant medical information, theories, and data, *med*+*ML* classifies *p*'s image of a mole as melanoma. Suppose now that *med*+*ML* suggests chemotherapy as the most promising treatment for *p*'s melanoma. Consider further that *med*+*ML* also offers a *bona fide* explanation for this output (cf. Durán 2021). That is, the explanation is well-structured, answers *why*-seeking questions—as opposed to merely classifying the output—and delivers epistemic goods, such as understanding the output and coherence with a larger body of medical beliefs. For the sake of the argument, let us say that *med*+*ML* offers a reliable diagnosis and an accurate treatment.

Thus, the output of *med*+*ML* plays a critical role in forming the physician's epistemic attitude: the physician believes to possess medically relevant knowledge about *p* having melanoma, and that the best treatment is chemotherapy. Furthermore, having an explanation of the output also fosters a moral belief in the physician, one in which she is justified in administering chemotherapy to *p*. We frame it this way because, *ex hypothesi*, the physician is in no epistemic, cognitive, or moral position to confirm, contend, or opt out from believing the output of *med*+*ML*. As presented, the physician is epistemically justified in believing the output of *med*+*ML* and morally justified in subjecting *p* to chemotherapy treatment (Durán and Jongsma 2021).[12]

In light of this example, the physician is convinced that she holds a piece of knowledge about *p* and that she is compelled to accept the treatment suggested by the *med*+*ML* as likely the most suitable for *p*. In terms of the informativeness account, the physician is then morally justified in preparing and subjecting the patient to chemotherapy, as per the epistemically grounded recommendation of *med*+*ML*.

Consider two further developments. First, chemotherapy induces anemia as a consequence of blood loss, bone marrow infiltration with disruption of erythropoiesis, and functional iron deficiency as a consequence of inflammation. This is a frequent and unfortunate consequence that many patients must face during chemotherapy (Bryer and Henry 2018). For a number of reasons, depending on the medical and genetic conditions of *p*, anemia can be treated with a blood transfusion. Second, *p*'s personal values dictate that receiving a blood transfusion is an unacceptable form of treatment, and it must be unequivocally rejected.

In light of the new information, one could argue that the physician can reject the output of *med*+*ML* and thus avoid any conflict with p's values. However, without further consideration, we see this move as problematic. First, we cannot assume the physician to be the absolute knowledge-generating entity capable of epistemically overriding *med*+*ML*. In fact, medical ML cannot be taken as yet another medical instrument for decision-making (such as MRI or blood count analysis) since it effectively displaces physicians from their epistemic role. In our case, this means that *p*'s treatment is, at best, on standby, awaiting the physician's decision on a course of action. In more complex cases, this might not even be possible. Let us also notice that the introduction of moral values in the epistemic assessment of ML might require, as it does in our case, a new treatment recommendation. In such a case, the physician has one of two options: either disregard the ML altogether, effectively neutralizing its use, or "factor" the moral values into the system. Our argument is that the informativeness account fails to consider the latter case.

Second, and more importantly, the suggestion to reject the output of *med*+*ML* cannot eschew a case of value conflict. It presupposes that refusing to treat *p* with chemotherapy follows from the principle of non-maleficence. However, this decision also clashes with the principle of doing no harm insofar as, without treatment, *p*'s biomedical well-being

[10] By "most suitable," we mean a medical course of action that takes into account the biological metrics of a patient as well as their personal, moral, and other values to inform their decision-making.

[11] In this respect, we follow philosophers of science, moral philosophers, and sociologists of science who have long debated about epistemic and non-epistemic values and their crossovers (see, e.g., Douglas (2009) and Longino (2004)).

[12] If this scenario sounds unlikely to happen, consider, again, physicians who are compelled to act upon the recommendations produced by algorithmic Prescription Drug Monitoring Platforms (PDMPs) (see Sect. 1). We consider systems such as PDMPs in assuming the epistemic and moral dependence of the physician on the ML system.

will be neglected. For the reasons given above regarding physicians' epistemic displacement by ML, we cannot take for granted that the physician will have a further course of action clearly figured out once it becomes clear that *p* is against blood transfusion.[13] It can be challenging—if not entirely impossible—to find an alternative treatment compatible with *p*'s values, and using an ML system as an epistemically powerful entity could be of great support. For this purpose, we need an epistemological framework that allows relevant ethical information (in the case considered, pertaining to *p*'s values) to have a direct bearing on crucial epistemic functions (such as explanation).

Now, by construction, the informativeness account is not sensitive to how a new piece of ethically relevant information (e.g., *p*'s personal values) should be included in the evaluation of central epistemic features of the systems (i.e., crucial epistemic functions, such as explanations). Concretely, this means that in view of the informativeness account, the physician remains epistemically and morally justified in subjecting *p* to chemotherapy. This is the case because, within the theoretical framework of the informativeness account, how epistemic functions should be adapted in order to include relevant ethical considerations remains unconsidered. To admit the possibility that ethical properties have a regulatory influence on the epistemological functions of ML means accepting that the epistemology and ethics of ML must constantly be re-evaluated. However, as pointed out in Sect. 2, the informativeness account remains silent on the possibility that ethical considerations—in our case *p*'s values—can have a bearing on the epistemological assessment of *med*+*ML*.[14] For these reasons, the informativeness account does not fit situations similar to the one under scrutiny.

To render the output of *med*+*ML* based on *p*'s values actionable for the physician, the epistemic assessment of *med*+*ML* must be reconsidered, including relevant ethical information pertaining to *p*'s situation. To our mind, the fact that an ethical property of the situation (i.e., *p* being against blood transfusion) should lead to the reassessment of an epistemic function highlights how epistemological and ethical considerations of ML are closely intertwined and mutually regulatory instead of compartmentalized, as the informativeness account takes them to be.

Against this background, how can we evaluate the recommendation produced by *med*+*ML* for further medical action? On the one hand, *med*+*ML*'s recommendation is based on standard medical and biological theories, evidence, and the general body of knowledge on diverse types of cancer. In this respect, we can say that the physician is *epistemically justified* (Durán and Formanek 2018) in believing that the recommendation is pertinent since the epistemic state induced by *med*+*ML* supports such a recommendation (i.e., the output is based on an accurate analysis of the biological state of *p*). On the other hand, the physician is not *morally justified* in following through with the chemotherapy treatment because this conflicts with *p*'s values. The solution to this problem is to factor *p*'s values into the system to render a new treatment (e.g., the new best treatment given *p*'s values is surgery).

Let us also clarify the nature of the epistemic and moral normativity in place in the case under scrutiny. While, as already pointed out, the physician is justified to believe that chemotherapy is the best treatment biologically speaking, she is not justified to believe that it is the *overall* best treatment for *p*. This is because the moral claim entailed in *p*'s rejection of blood transfusion has a direct bearing on what the physician should believe is the most suitable course of action all things considered. Under a definition of health that exceeds the evaluation of biological parameters and also includes moral, social, and otherwise relevant considerations, the physician is epistemically justified to believe that chemotherapy is no longer the best treatment for *p*.[15]

If the above considerations are correct, then a more overarching view of the epistemology and ethics of ML emerges. Whereas Mittelstadt and colleagues rightly emphasize the informative value of the epistemology of ML on moral actions, we complement the missing parts of the framework by showing the merits of an epistemology regulated by the ethics of ML. We believe that cases similar to the one considered here are better analyzed through the lenses of a different approach, one that, as we argue, takes epistemic and moral features of medical ML as substantially regulatory—rather than informational—of each other. In a regulation-based framework, we submit that *p*'s personal values become a substantial part of the epistemology, regulating

[13] A related point concerns the extent to which physicians would consider personal values as relevant for diagnosis and treatment and thus as morally problematic. According to diverse approaches in medical ethics, the physical well-being of a person supersedes personal values (Richman 2004). Although we cannot elaborate on these considerations, they seem relevant to the issue at hand.

[14] The problem presented here is different from assuming that the epistemology of ML is empty of values (moral, cultural, economic, political, etc.). Mittelstadt et al. would admit that an explanation rendered by *med*+*ML* depends on the kind of question we want to answer, the information provided, etc. In summary, the epistemology of ML is not value-free. The crucial difference is that Mittelstadt et al. consider that once the epistemology of ML is settled, it is informative. Our contention is that there is a "loop back" from the ethics to the epistemology, a loopback that is unaccounted for by the informativeness account.

[15] This interpretation of health as more encompassing than biological health is aligned with the WHO definition of this concept: https://www.who.int/about/governance/constitution For further debates on different conceptions of health, see also Richman (2004). Unfortunately, we cannot expand on these issues in this paper.

the epistemological assessment of the system. The regulatory role of ethical features in the epistemology of ML comes to light in its considerable impact on the physician's beliefs. Even if, all things being equal, she would be justified in believing and acting upon the explanation provided by *med* + *ML*, this is no longer the case as soon as a relevant ethical property of the situation comes to the foreground. Only within a regulation-based framework do induced epistemic attitudes of the physician elicit a clear stand in either being (or not being) morally justified in proceeding with a given course of action.

Drawing on the previous discussion and on the example under scrutiny, we can now consider how, following the logic underlying the informativeness account, situations in which *p* is the victim of an epistemic injustice do not find treatment. We turn to this analysis in the next section.

4 Epistemic injustice

We argued in Sect. 3 that the informativeness approach does not account for information regarding the patient's (*p*'s) values that become relevant after the ML has outputted a treatment recommendation. How does this affect the practice of healthcare with ML, above and beyond the fact that the system's recommendations are unsatisfactory in cases such as ours? We submit that there is a wrong done to *p*, understood in terms of Miranda Fricker's account of *epistemic objectification*, which falls within her analysis of *epistemic injustice* (Fricker 2007).

In its broadest sense, epistemic injustice designates flawed practices in meaning-making and knowledge-creating processes, leading to marginalization, unfair distrust, silencing, and exclusion (among others) (Pohlhaus 2017). As such, epistemic injustice is a wrong done to epistemic subjects in their capacities as knowers, that is, as recipients and conveyors of knowledge. Issues of epistemic injustice have mainly been addressed in terms of a credibility deficit attributed to individuals belonging to vulnerable societal groups,[16] precisely due to their perceived social identity from the side of their interlocutor(s) (*testimonial injustice*) or to an inability to comprehend and make sense of their own social experience due to a lack of or access to shared hermeneutical resources (*hermeneutical injustice*) (Fricker 2007).

This multi-faceted phenomenon has been receiving increasing attention in the philosophical debate at the intersection between social epistemology and ethics in recent years (e.g., Byskov 2021; Carel et al. 2017; Chung 2021; Moes et al. 2020; Thomas et al. 2020; Wardrope 2015). Since ML systems are epistemically authoritative and increasingly involved in decision-making procedures that strongly impact patients' lives, it is of paramount importance to ensure that they do not undermine epistemic subjects in their capacities as knowers. As such, epistemic injustice in ML-mediated contexts requires particular attention (Symons and Alvarado 2022; Pozzi 2023a, b). Issues of epistemic injustice emerge, generally, if patients are excluded from influencing decision-making processes and if their lived experiences, testimony, and personal values (epistemic, moral, societal, etc.) are not acknowledged as legitimate sources of knowledge, among many other factors that still need to be explicitly addressed and investigated in depth.[17] The analysis of our example in light of the phenomenon of epistemic injustice should point to the importance of working toward the development and deployment of ML systems that do not represent an obstacle to the active participation of relevant stakeholders in shared decision-making. Operationalizing systems that do not impair the process of understanding and forming beliefs regarding our lived experiences is, in fact, essential to avoid genuinely epistemic forms of injustice that can otherwise emerge.

The following analysis allows us to shift the focus of the debate from a conception of epistemic injustice, which has been mostly considered in a human-centered fashion, to its application to cases in which epistemic subjects interact with ML systems. It is our aim to show that Fricker's concept of epistemic objectification can be successfully applied to our ML case to capture the moral wrong suffered by *p*.

Let us now turn to the reconstruction of Fricker's account of epistemic objectification so that we can, in a second step, show that it can capture at best the moral wrong inflicted on *p* in the case addressed in Sect. 3.

4.1 Epistemic objectification

According to Fricker, a subject is epistemically objectified in situations in which she is, due to prejudices from the side of the hearer(s), completely deprived of her active role as an informant and is, as such, reduced to a *mere* source of information.[18] Drawing on Craig's account of the State of Nature (Craig 1990), Fricker argues that the distinction between

[16] Vulnerable epistemic subjects can be considered such due to their gender and race but also because they find themselves in precarious health conditions.

[17] The analysis of all the aspects mentioned above would go well beyond the scope of this paper. In this contribution, our investigation is limited to pointing out how the example under scrutiny can be interpreted through the lens of Fricker's conception of epistemic objectification.

[18] The analysis excludes cases in which the hearer judges their interlocutor as epistemically untrustworthy and, for this reason, and not due to some forms of prejudices, she does not acknowledge them in their role of informants (Fricker 2007, p. 136).

'informant' and 'source of information' in the process of conveying knowledge is an epistemological aspect that entails relevant ethical meanings. Informants are to be considered epistemic agents who are able to convey information actively and share knowledge with their interlocutors (e.g., by communicating information). In the field of healthcare, a patient can be considered an active informant in that she can communicate relevant information regarding her physical and mental state to her physician, thus participating and playing a role in informing medical decisions.

Differently, sources of information are states of affairs from which an inquirer can deduce information. Therefore, as Fricker points out, individuals can be both informants, being able to actively express themselves and convey knowledge, and sources of information, in that the inquirer can derive information about their current state, for instance, through observational evidence of their behavior (Fricker 2007, 132). For a human being to be a source of information could be no reason for concern from an ethical point of view; this is the case, for example, in a situation in which a physician concludes that a patient suffers from a particular pathology due to the analysis of medical tests conducted on the patient in question. That is, the physician comes to a conclusion regarding the current state of the patient without the patient actively communicating it.[19] Against this background, it is undisputed that the dimensions of being an active informant and a source of information coexist in human beings as epistemic subjects.

By contrast, treating someone as a *mere* source of information implies an instrumentalization of the subject, depriving them completely of their role as active informants. One does not need to adopt Kantian principles of morality to acknowledge that the instrumentalization of subjects is universally wrong from a moral point of view. In the context of medicine and healthcare, being treated as a *mere* source of information would mean that the patient is expected to provide basic information regarding her current state but is deprived—due to, for example, prejudices that physicians or other healthcare professionals have related to their social identity—from the possibility of participating and contributing in a substantial way to the collective epistemic activity of sharing their lived experience. However, this is arguably key to making sense of their health situation and actively participating in shared medical decision-making processes (Carel and Kidd 2017).

Fricker takes the phenomenon of epistemic objectification as reconstructed as a particularly harmful form of silencing and the central wrong derived from epistemic injustice (Fricker 2007, 6). Indeed, the fact that a subject's active contributions are limited or impaired altogether represents a considerable restriction to their agential role as rational individuals and strongly constrains their participation in the production and exchange of knowledge. Thus, this can be considered the primary wrong that epistemic injustice understood in terms of epistemic objectification perpetrates on its victims since, in these cases, the knower is deprived of her active agential role and, as such, "wronged in a capacity essential to human value" (ibid., 44). A secondary kind of wrong can manifest in more practical—but not less detrimental—terms, also creating a clear disadvantage for the subjects involved. In the context of healthcare, for example, the risk of attributing to patients a deflated level of credibility on the basis of prejudices connected to their status as ill persons could lead to being misdiagnosed.[20]

A growing body of literature addresses the fact that ill persons can be considered a particularly vulnerable category inclined to suffer epistemic injustice in Fricker's sense (Carel and Kidd 2014). Kidd and Carel (2017) argue that judgments about the epistemic credibility of ill persons are often prejudicial, being produced and sustained by both negative stereotypes and the structural characteristics of healthcare practices (ibid., 175). In particular, they point out that ill persons are vulnerable to epistemic injustice through the supposed attribution of characteristics such as cognitive unreliability and emotional instability that deflate their testimony's credibility. However, there are also structural features of healthcare systems that can be regarded as the causes of hermeneutical forms of injustice (rather than the intentions of individuals). For example, considerable time limitations and the use of standardized protocols contribute to the marginal role assumed by personal needs and values (ibid., 176). Further, the difficulty of articulating particular aspects related to a patient's illness is an aspect that can be considered challenging from a hermeneutical point of view. Overall, being in a physically and mentally precarious condition puts the patient in a situation of vulnerability and dependence, which undermines her own epistemic confidence (ibid., 174).

Drawing on what has been said so far, we can conclude that in standard, that is, non-ML-mediated practices in healthcare, there are factors such as the ones previously mentioned that put the patient p into a position of epistemic vulnerability. We submit that the situation becomes even more pressing when an additional epistemically

[19] A more straightforward example from everyday life could be a case in which someone blushes, and from this behavioral feature, we derive that he or she is embarrassed.

[20] As a matter of fact, in a case in which "the style of interaction between clinician and patient is one that closes down communication, such that important information is potentially lost" (Havi Carel and Kidd 2014, 531), it is not too far-fetched to think of the possibility of misdiagnosing a patient as a legitimate practical concern deriving from an instance of testimonial injustice.

authoritative entity, such as *med + ML*, becomes involved in this relationship.

4.2 Informativeness and epistemic objectification in ML

We now consider how *med + ML*, without allowing for the possibility of integrating *p*'s values into its epistemology, brings about a case of epistemic objectification at *p*'s expense in the example under scrutiny. This analysis aims to further show the need to implement ML systems that allow ethical features (say, a patient's values) to regulate epistemologically relevant aspects (e.g., an explanation provided by the system). However, before turning to this analysis, some considerations are in order. Whereas Fricker sees epistemic objectification as the most direct expression of instances of testimonial injustice, we need to detach ourselves from her human-centric approach to make our case for epistemic objectification brought about by *med + ML* at *p*'s expense. As previously mentioned, the wrong that she aims to capture is caused by unjustly deflated credibility judgments that a subject receives from her interlocutor due to prejudices related to her social identity. Since, in our case, the physician does not play an active role in mediating between *med + ML* and *p*, prejudicial judgments that could be detrimental to *p*'s epistemic positions are out of place.[21] Even less plausible would be the assumption that *med + ML* holds prejudices that deflate *p*'s credibility. In fact, it goes without arguing that attributing these genuinely human traits to an ML system would be a category mistake. Thus, the epistemic objectification we aim to capture is one that emerges because *med + ML* cannot pick up on *p*'s values, and therefore, *p*'s agential contribution to the decision-making process cannot be successfully considered. This is, we claim, due to how *med + ML* operates, as elaborated in detail below.

To convincingly argue that *p* suffers a form of epistemic objectification brought about by *med + ML,* we need to account for the fact that *p*'s knowledge (e.g., in the form of her personal epistemic values) is wholly excluded from the decision-making process leading to the output. Relatedly, having shown this will allow us to argue that *p* is treated as a *mere* source of information and not as an active informant. It follows that *p* is epistemically objectified. We show that *p* is utterly excluded from the decision-making process by making explicit a *vicious circularity* in how *med + ML* produces its output, which is unsolvable following the informativeness account. We take the epistemic objectification of *p* as a direct consequence of the vicious circularity to which our discussion now turns.

As previously argued, by construction, the informativeness account puts forward an investigation of how the epistemology of ML informs the ethics of ML. We showed that ethical elements substantially affecting the epistemological counterpart are left unaddressed. Thus, the informativeness account adopts a unidirectionality that goes from the epistemology to the ethics of ML but not the other way around. The example in Sect. 3 illustrates this unidirectionality.

As indicated, the informativeness account considers, *ex hypothesi*, the epistemological assessment of *med + ML* to be "fixed" and therefore unmodifiable by new incoming information that may be relevant to *p*'s medical condition. From this perspective, *med + ML* induces in the physician the belief that the explanation is suitable, along with the moral justification for acting upon it. This informs, in turn, the physician's actions. It follows that at the moment in which *p* is confronted with the output brought about by *med + ML*'s suggestion of chemotherapy and, consequently, blood transfusion as the most suitable treatment, *p* will have to refuse the suggested therapy, restating that it goes against her personal values. At this point, the vicious circularity becomes obvious: since *med + ML* is unable to factor this relevant piece of information into central epistemological functions (such as in the explanation of the output), *p* can only be confronted anew with the same outcome produced by the medical ML system, an unsuitable treatment recommendation for her set of values. Our claim is that the reiteration of this hypothetical yet logically consistent scenario exerts distinctive negative influences on *p*'s epistemic confidence and, more importantly, strongly limits her agential role. Indeed, from the moment in which the output of *med + ML* is created, *p* is not in a position to actively influence the decision-making process and is, consequently, completely left out of it. This leads us to the second claim, that is, *med + ML* leads *p* to be treated as a *mere* source of information.

As previously pointed out, treating *p* as a source of information is generally unproblematic and thus also in cases in which a medical interaction is mediated by an ML such as *med + ML.* This is the case since *med + ML* can elaborate information regarding *p*'s physical state that she might not have directly provided but that has been acquired through different processes. Indeed, the system can effectively elaborate information regarding *p*'s physical condition from indirect sources, such as laboratory tests or any kind of medical examination she has undergone. The epistemically and ethically relevant problem in terms of epistemic objectification arises as soon as *p* is in possession of a relevant piece of information (i.e., the fact that *p* is against blood transfusion) that cannot, however, be accounted for by *med + ML.* That is, at the point where *p* should *actively* convey new relevant information, she is prevented from doing so due to the role played by *med + ML.* As a consequence, *p* cannot receive an

[21] Note that, as previously mentioned, we exclude the possibility of the physician intervening independently from the *med + ML.*

appropriate medical treatment compatible with her personal values.

Drawing on the discussion so far, it can be argued that *p* is treated as a *mere* source of information since her agential contribution is left unconsidered by the explanation provided by *med* + *ML*. This constitutes, as such, a case of epistemic objectification. Very crucially, the unidirectionality previously pointed out leads to a unidirectional exchange of knowledge: the end users of *med* + *ML* are merely recipients of knowledge but are not able to actively influence the knowledge-producing process itself. This outcome leads to undesirable consequences for *p*: either an endless circle in which no suitable alternative to the output produced by *med* + *ML* is found or a medical procedure that infringes on her personal—moral and epistemic—values.

Bottom line, the informativeness account does not provide the theoretical backdrop needed to tackle moral and epistemic issues in terms of patient objectification as we have been discussing them. These considerations reinforce the need for a flexible epistemology capable of incorporating new relevant information as it is acquired to overcome moral and epistemic concerns in connection with the epistemic objectification of the relevant stakeholders involved in medical decision-making processes.

5 Final remarks

This contribution aims to point out the limitations of an approach in the epistemology and ethics of ML that sees these two dimensions as compartmentalized. In particular, we analyzed considerations of the general relationship between the epistemology and ethics underlying the approach taken by Mittelstadt et al. (2016). We reconstructed their methodology in terms of an information-serving relationship between the epistemology and ethics of ML according to two dimensions that, to our mind, characterize their analysis (i.e., instrumentality and autonomy). In Sect. 3, we substantiated our claims by considering a case of explanatory medical ML that cannot be appropriately solved following the logic of the informativeness account. We analyzed the ethical consequences of this situation for the patient involved in the example considered in terms of Fricker's concept of epistemic objectification (see Sect. 4).

Our main criticism toward the informativeness account is that it is not designed to address cases that require the analysis of how an ethical property (such as patients' values as discussed in the case in Sect. 3) should lead to the re-evaluation of central epistemic functions in situations mediated by a medical ML. The informativeness account remains silent on the possibility that *p*'s values motivate the rejection of an otherwise well-constructed explanation (thus requiring a new one). Hence, morally problematic situations in which a physician is no longer justified to act upon the ML output produced do not find treatment within this theoretical framework.

Whereas, in line with the relevant literature, Mittelstadt and colleagues rightly emphasize the informative value of the epistemology of ML on moral actions, our aim was to make clear the need to complement their framework by showing the merits of an epistemology regulated by the ethics for ML. The regulatory role that ethical features have on epistemic functions has been made explicit in the example analyzed in Sect. 3: in that case, the physician is, in principle, not justified in acting upon the explanation provided by the ML system in the face of *p*'s values as a relevant ethical feature of the situation. In turn, this means that an otherwise sound explanation needs to be reformulated, including the consideration of *p*'s values. It is in this sense that we take that an ethical property *regulates* what counts as a morally acceptable explanation and what does not.

Admittedly, there is a need for more consideration of how a flexible epistemology can be formalized. However, with our work, we hope to have contributed to the debate, showing the importance of further pursuing this direction in future research to avoid epistemic and ethical issues such as those highlighted in this contribution.

Acknowledgements We are grateful to the participants of the CEPE/IACAP Joint Conference 2021 for their helpful feedback on a previous draft of this paper. We would also like to thank the Digital Philosophy Seminar group within the Ethics/Philosophy of Technology Section at TU Delft for fruitful discussions of the ideas we developed in this article.

This work was supported by the European Commission through the H2020-INFRAIA-2018-2020/H2020-INFRAIA-2019-1 European project "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (Grant Agreement 871042). The funders had no role in developing the research and writing the manuscript.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends intechnology, arts, science and society, commenting on issues of concernto the research community and wider society. Whilst the drivefor super-human intelligence promotes potential benefits to widersociety, it also raises deep concerns of existential risk, thereby-highlighting the need for an ongoing conversation between technology-and society. At the core of Curmudgeon concern is the question:What is it to be human in the age of the AI machine? -Editor.

Data availability A data availability statement is not applicable as there are no data associated with this article. The formulation already included 'We declare that the manuscript has no associated data' is thus pertinent.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Alpaydin E (2014) Introduction to machine learning. Massachusetts Institute of Technology.

Babushkina D, Votsis A (2022) Epistemo-ethical constraints on AI-human decision making for diagnostic purposes. Ethics Inf Technol 24:22

Beisbart C (2021) Opacity thought through: On the intransparency of computer simulations. Synthese 1–24.

Bjerring JC, Busch J (2021) Artificial intelligence and patient-centered decision-making. Philosophy Technol 34:349–371

Bryer E, Henry D (2018) Chemotherapy-induced anemia: Etiology, pathophysiology, and implications for contemporary practice. Int J Clin Trans Med 6:21–31

Byskov MB (2021) What makes epistemic injustice an "injustice"? J Soc Philos 52:114–131. https://doi.org/10.1111/josp.12348

Carel H, Kidd IJ (2014) Epistemic injustice in healthcare: a philosophial analysis. Med Health Care Philos 17:529–540. https://doi.org/10.1007/s11019-014-9560-2

Carel H, Kidd IJ (2017) Epistemic injustice in medicine and healthcare. In: Kidd IJ, Medina J, Pohlhaus G (eds) The Routledge Handbook of Epistemic Injustice. Routledge, pp 336–346

Carel H, Blease C, Geraghty K (2017) Epistemic injustice in healthcare encounters: evidence from chronic fatigue syndrome. J Med Ethics 43:549–557. https://doi.org/10.1136/medethics-2016-103691

Chung R (2021) Structural health vulnerability: Health inequalities, structural and epistemic injustice. Journal of Social Philosophy 1–16. https://doi.org/10.1111/josp.12393.

Craig E (1990) Knowledge and the state of nature: an essay in conceptual synthesis. Clarendon Press.

De Laat PB (2018) Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philosophy Technol 31:525–541. https://doi.org/10.1007/s13347-017-0293-z

Douglas H (2009) Science, policy, and the value-free ideal. University of Pittsburgh Press.

Durán JM (2021) Dissecting scientific explanation in AI (sXAI): a case for medicine and healthcare. Artif Intell 297:103498

Durán JM, Formanek N (2018) Grounds for trust: essential epistemic opacity and computational reliabilism. Mind Mach 28:645–666. https://doi.org/10.1007/s11023-018-9481-6

Durán JM, Jongsma KR (2021) Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 47:329–335

Esteva A et al (2019) A guide to deep learning in healthcare. Nat Med 25:24–29. https://doi.org/10.1038/s41591-018-0316-z

Fricker M (2007) Epistemic injustice: Power and the ethics of knowing. Oxford University Press.

Grote T, Berens P (2020) On the ethics of algorithmic decision-making in healthcare. J Med Ethics 46:205–211

Haines S et al (2022) Key considerations for the implementation of clinically focused prescription drug monitoring programs to avoid unintended consequences. International Journal of Drug Policy 101. https://doi.org/10.1016/j.drugpo.2021.103549.

Hatherley JJ (2020) Limits of trust in medical AI. J Med Ethics 46:478–481

Humphreys P (2009) The philosophical novelty of computer simulation methods. Synthese 169:615–626. https://doi.org/10.1007/s11229-008-9435-2

Kidd IJ, Carel H (2017) Epistemic Injustice and Illness. J Appl Philos 34:172–190. https://doi.org/10.1111/japp.12172

Longino HE (2004) How values can be good for science. Sci Values Objectivity 127–142.

Mittelstadt BD et al (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3:2

Moes F et al (2020) Questions regarding 'epistemic injustice' in knowledge intensive policymaking: two examples from Dutch health insurance policy. Social Science and Medicine 245.

Morley J et al (2020) The ethics of AI in health care: A mapping review. Soc Sci Med. https://doi.org/10.1016/j.socscimed.2020.113172

Oliva, J. (2022). Dosing Discrimination: Regulating PDMP risk scores (January 18, 2021). 110 California Law Review 47, Available at SSRN: https://ssrn.com/abstract=3768774 or https://doi.org/10.2139/ssrn.3768774

Pohlhaus G (2017) Varieties of epistemic injustice. In: The Routledge handbook of epistemic injustice, pp 13–26.

Pozzi G (2023a) Automated opioid risk scores: a case for machine learning-induced epistemic injustice in healthcare. Ethics Inf Technol 25:3

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

Richman KA (2004) Ethics and the metaphysics of medicine: Reflections on health and beneficence. MIT Press.

Russo F, Schliesser E, Wagemans J (2023) Connecting ethics and epistemology of AI. AI & Society 1–19.

Symons J, Alvarado R (2022) Epistemic injustice and data science technologies. Synthese 200:87. https://doi.org/10.1007/s11229-022-03631-z

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/

Thomas A et al (2020) What is "shared" in shared decision-making? Philosophical perspectives, epistemic justice, and implications for health professions education. J Eval Clin Practice 26:409–418. hy10.1111/jep.13370.

Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25:44–56

Tsamados A et al (2021) The ethics of algorithms: key problems and solutions. AI and Society 0123456789. https://doi.org/10.1007/s00146-021-01154-8.

Wardrope A (2015) Medicalization and epistemic injustice. Med Health Care Philos 18:341–352. https://doi.org/10.1007/s11019-014-9608-3

Zarsky T (2016) The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. Sci Technol Human Values 41:118–132