The American Journal of Bioethics

ISSN: 1526-5161 (Print) 1536-0075 (Online) Journal homepage: www.tandfonline.com/journals/uajb20

The Ethics and Epistemology of Clinician-AI Disagreement in Medicine: Beyond Opposition

Giorgia Pozzi, Martin Sand & Karin Jongsma

To cite this article: Giorgia Pozzi, Martin Sand & Karin Jongsma (27 Feb 2026): The Ethics and Epistemology of Clinician-AI Disagreement in Medicine: Beyond Opposition, The American Journal of Bioethics, DOI: 10.1080/15265161.2026.2632008

To link to this article: https://doi.org/10.1080/15265161.2026.2632008

THE AMERICAN JOURNAL OF BIOETHICS
https://doi.org/10.1080/15265161.2026.2632008

Taylor & Francis
Taylor & Francis Group

TARGET ARTICLE

OPEN ACCESS Check for updates

The Ethics and Epistemology of Clinician-AI Disagreement in Medicine: Beyond Opposition

Giorgia Pozzi[a], Martin Sand[a] and Karin Jongsma[b]

[a]Delft University of Technology; [b]University Medical Center Utrecht

ABSTRACT

The integration of AI systems in medical care magnifies questions related to how physicians should work with such systems to ensure the best patient outcomes. A particularly thorny issue is related to dealing with situations of possible disagreement between an AI system's recommendation and the course of medical action envisaged by a human clinician. The current academic debate has so far suggested three possible ways of dealing with such clinician-AI disagreements. First, by considering when clinicians are justified in deferring to the AI output (what we call the *deference approach*), second when the human user overrules the AI system's output in cases of disagreement (the *overruling approach*), and lastly when a second human opinion is deemed necessary to resolve disagreements (the *second opinion approach*). In this paper, we aim to spell out the shortcomings of these three approaches for dealing with clinician-AI disagreement and offer a more nuanced perspective on such disagreements. We argue that differentiation between types of disagreements, taking into account the role attributed to AI in medical practice, is essential before determining how clinician-AI disagreements should be dealt with. By drawing on a case that exemplifies how multifaceted medical decision-making is, we point out the normative implications of possible clinician-AI disagreements ensuing from it. We highlight the distinctive uncertainties inherent to medical decision-making, showing that disagreements in these contexts are not merely unavoidable but can even be epistemically valuable. Ultimately, by considering the epistemic positions of clinicians and AI systems, our analysis raises important questions for the epistemology of disagreement that need timely attention.

KEYWORDS

Medical decision-making; ethics and epistemology of AI; disagreement; human-AI collaboration

INTRODUCTION

Imagine being a patient who has undergone a CT scan, which is assessed both by an AI system and by your treating physician. The physician has judged the CT scan to be normal, yet the AI system has detected an abnormality. This may be understood as a human-AI disagreement,[1] and highlights unresolved epistemic and ethical questions about how we should respond to such situations.

In the ethical and legal debate, various suggestions for dealing with such situations have been proposed. For instance, in line with the widely circulated belief that AI "outperforms" human doctors, some have argued that clinicians should defer to the AI system's suggested course of action (Grote and Berens 2020). In contrast, others have suggested that humans should be able to overrule an AI system, for instance, when they cannot verify its output, a demand also found in ethical and legal guidelines (Declaration of Montreal 2018; AI Act, Regulation 2024/1689; Lekadir et al. 2025). Lastly, there is the "second opinion" approach, as advanced by Kempt and Nagel (2022). In cases of human-AI disagreement, these authors recommend "the rule of disagreement," according to which the

CONTACT Giorgia Pozzi g.pozzi@tudelft.nl Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands.

[1]In this paper, we conceive of disagreements as situations in which the AI system's output diverges from the clinician's preferred course of action. Most of the literature on the epistemology of disagreement focuses on disagreements occurring between epistemic peers. However, as acknowledged in the relevant literature (see, e.g., Christensen 2007), determining peerness in terms of epistemic equality between human agents is already difficult, and the challenge becomes even greater when we attempt to determine what epistemic peerness amounts to in human-AI interactions. This is the case because central concepts to this debate such as freedom from bias and intelligence (Christensen 2007) can hardly be applied to human-AI cases without risking a misclassification of AI systems. Thus, we are aware that the term "disagreement" might carry an anthropomorphizing connotation. However, when we speak of disagreement, we do not imply that AI systems are full-fledged agents, whether epistemic or moral. For further discussion of different interpretations of disagreement in the AI literature see Kempt et al. (2023). We thank an anonymous reviewer for encouraging us to clarify this point.

disagreement is resolved by turning to another human physician if the AI diagnosis does not align with that of the physician in charge.

Clearly, these approaches to handling human-AI disagreement vary in morally and practically significant ways. These approaches point in different directions when clinicians encounter a disagreement that has morally relevant consequences for patient care. While this body of literature has largely illuminated the intricate problems that medical AI raises, we believe these approaches have shortcomings, being partially based on wrongful assumptions and too schematic cases that neglect the complexity of medical practice and the particularities of these disagreements. In contrast to these positions, we argue that disagreements should not always be overcome, that they do not have to be resolved based on epistemic authority, and that we should consider the possibility that both the physician and the AI system might be wrong or both partially right (Babushkina and Votsis 2022; Van Leersum and Maathuis 2025).

With AI being introduced in routine medical care, disagreements between clinicians and AI systems' outputs are likely to emerge more frequently in AI-mediated medical practices. Therefore, it is important to find ways to navigate AI-clinician disagreements.[2] While we take a critical stance toward the main approaches currently available in the literature, it is not our aim to dismiss them entirely. In fact, there can be situations in which overruling the AI's output, deferring to it, or seeking a second opinion might be the appropriate course of action. However, contextual considerations are needed, pertaining to, among others, the nature of the disagreement at hand, which are often sidelined in the current debate. In this paper, we take a step back from prescriptive approaches on how to deal with disagreements and critically question what disagreements amount to, regarding which available evidence they emerge, and which normative considerations are relevant in moving forward. In doing so, our analysis should not be interpreted as being *against* existing approaches, but rather as offering ways that build upon and go *beyond* them.

In this paper, we consider the case of an AI system used to support clinicians in the diagnosis and treatment decision-making for suspected pulmonary embolism (PE) (see Section "Case: Pulmonary Embolism Treatment Decision in Patients with an Increased Risk of Bleeding"). We deem this case particularly fitting for our analysis because it illustrates the complexity of medical decision-making under uncertainty, which allows us to exemplify the different ways in which disagreements can be dealt with. Building on this case, we show that none of the three approaches for human-AI disagreement available in the literature - that we coin the *deference*, the *overruling*, and the *second opinion approach* - can satisfactorily account for how to deal with AI-clinician disagreement in the case considered (see Sections "Beyond Deference," "Beyond Overruling," and "Beyond Second Opinion," respectively). In Section "Beyond Resolving Disagreement: Kinds of Disagreement in Context and their Normative Implications," we provide a constructive way forward for navigating human-AI disagreement. Here, we also spell out relevant normative implications of this approach and shed light on the possible value of disagreement[3] in clinician-AI interactions, thus supporting a non-antagonistic approach.

CASE: PULMONARY EMBOLISM TREATMENT DECISION IN PATIENTS WITH AN INCREASED RISK OF BLEEDING

Let us consider the situation of a 75-year-old patient who presents to the emergency department with sudden-onset shortness of breath, mild chest pain, and tachycardia. The patient has a history of gastrointestinal bleeding and chronic kidney disease. The CT pulmonary angiogram shows a small, subsegmental pulmonary embolism.

Acute pulmonary embolism (PE), if untreated, is associated with a significant mortality rate (as high as 30%) (Belohlavek et al. 2013). Commonly, PE is clinically managed with pharmaceutical treatment (Zuo et al. 2021), which includes anticoagulation treatment ('blood thinners' like heparin) as a standard and thrombolytic treatment in more exceptional cases. Both for anticoagulation treatment and thrombolytic treatment, there is a risk of adverse effects such as minor or major bleeding, especially in patients with a high bleeding risk. Furthermore, studies suggest that

[2]In this paper, we are focusing on immediate patient health and well-being as the primary values to strive for in clinical settings when dealing with disagreements. However, we recognize that this does not exhaust other values relevant in clinical decision-making, such as healthcare costs and treatment demands in the long term. While working out the criteria for dealing with disagreement more explicitly will remain a task for future research, we see our current analysis as offering a relevant starting point for these considerations.

[3]For a general discussion on the value of disagreement in bioethics see Parker (2025).

subsegmental PE (small clots in small arteries) may not always require treatment, especially in patients with high bleeding risk (Baumgartner and Tritschler 2021). However, evidence is mixed, and clinical guidelines do not offer a definitive answer as to which treatment is better in specific circumstances, leaving room for physicians' discretion.

To date, several AI systems have been developed for the diagnostic process of PE (Allena et al. 2023; van Maanen et al. 2025). Diagnosing PE is challenging due to nonspecific symptoms, yet it can have possible life-threatening effects on the patient if the diagnosis is missed. For diagnosis, referral to a hospital is generally necessary. As far as we know, there is currently no AI system focusing on the treatment decision-making of PE once adequately diagnosed. Therefore, we draw on a hypothetical AI system for diagnosis and treatment decision-making for pulmonary embolism called PE-TER (Pulmonary Embolism TreatmEnt Recommender-system).[4] PE-TER can recommend a) whether pharmaceutical treatment is recommendable over no treatment with close monitoring, b) personalize this advice for patients with high bleeding risks and c) provide recommendations about which treatment alternative should be provided, including full dose of anti-coagulation (to prevent cloth growth and new cloths forming, higher bleeding risk), short-course anticoagulation (resolve existing cloths, lower bleeding risk or thrombolytic treatment; in cases of massive cloths, there is a dispute about the bleeding risk compared to anticoagulation). In our hypothetical scenario, PE-TER is a highly trained and validated system whose performance in providing potentially suitable treatment recommendations following a diagnosis of PE is comparable to that of human clinicians and has been used in routine patient care for over 5 years. Our hypothetical Dr. Parker, one of the experienced clinicians, has applied PE-TER to the case described above. PE-TER recommends full-dose anticoagulation treatment, while Dr. Parker suggests no treatment and close monitoring. It is unclear how this disagreement should be dealt with.

[4]Our case has been drafted with the help of clinical experts. It should be noted that even if the case involves a hypothetical AI system, such multi-purpose systems (e.g., used both for diagnostic and treatment recommendation purposes) are currently considered for inclusion in clinical care (see, e.g., Duwe et al. 2024; Mishra and Shridevi 2024). This shows that these systems are generally interesting for developers and clinicians and our analysis of possible emerging disagreements deserves timely scrutiny. Let us also point out that while the system's performance in providing diagnoses can be assessed in terms of accuracy, this notion is less applicable to treatment recommendations. We refer here to PE-TER's performance as a generally more suitable term.

RESOLVING DISAGREEMENT

In this section, we analyze three widely discussed and contrasting approaches for resolving human-AI disagreement available in the literature by applying the case provided in the previous section to these approaches. The first postulates that clinicians should defer to the AI output in cases of disagreement (hence we label this the *deference* approach), the second entails the need for the clinicians to override the AI output (hence, we dub this the *overruling* approach), and the third calls for redirecting the decision to a further human clinician in case of a disagreement with the AI (we call this the *second opinion approach*). In this section, we discuss these three approaches, outlining their shortcomings and assumptions to lay the foundations for a constructive way of dealing with human-AI disagreement, which we provide in detail in the following sections.

Beyond Deference

The deference approach holds that clinicians should defer to the AI system's suggested course of action in cases of disagreement. This position presupposes that clinicians and AI systems are epistemic authorities because of their ability to gather and process a considerable amount of clinically relevant information, thus providing, say, suitable treatment plans. The proponents of the deference approach argue that clinicians should adjust their judgment to align with the AI system's recommendation in cases of disagreement with it. For instance, Bjerring and Busch (2021) argue that "if a practitioner honors her epistemic obligation, she will align her medical verdict with that of the superior AI system, but in doing so, she will end up violating central tenets of patient-centered medicine" (352). So, even though these authors are worried that AI systems in medicine can result in paternalism, they hold that, from an epistemic perspective and all things being equal, clinicians should defer to the AI system because of its superior epistemic standing. We refer to this and positions advanced in the literature along similar lines as the *deference approach*.

Another sophisticated elaboration of this view has been provided by Grote and Berens (2020). These authors analyze possible cases of disagreement between a dermatologist's assessment of skin cancer classification and that of an AI system. The disagreement is problematized by considering a hypothetical situation in which the clinician arrives at a certain diagnosis while the AI produces an outcome that differs from the clinician's diagnosis and, crucially, both have a

similar degree of confidence. These cases are particularly challenging because it is not straightforward how much weight the clinician should attribute to the AI outcome and whether she should be less or more confident (or maybe even reconsider her diagnosis altogether). According to Grote and Berens, this kind of assessment is particularly difficult without additional information regarding how the algorithm produces its outcomes. Widely discussed epistemic limitations, such as these systems' epistemic opacity (Burrell 2016; Mittelstadt et al. 2016), notoriously prevent humans from accessing and evaluating the decision-making logic behind algorithmic outcomes. Grote and Berens argue that in cases of disagreement, "given that the algorithm is likely trained and validated on the opinions of several expert clinicians—deferring would seem like a reasonable choice." (ibid., p. 207)

While this approach has benefits, such as supporting faster decision-making, there are also shortcomings since it might lead, among others, to a possibly unjustified over-reliance on AI systems' outputs (Goddard et al. 2014). While Bjerring and Busch explicitly state that they take AI systems to be epistemically superior to humans, Grote and Berens seem to conceive of AI systems and human experts as comparatively epistemically authoritative. In both cases, the assumption is that AI systems have access to and can elaborate high amounts of expert knowledge and have, on occasion, demonstrated high accuracy (that might arguably be comparable to that of clinical experts), which speaks in favor of attributing to these systems a form of epistemic authority.

However, understanding human clinicians and AI systems as being positioned on a similar epistemic footing (even following the milder version of deference put forward by Grote and Berens) from an epistemic perspective has important normative implications, e.g., in terms of being answerable to patients by reconstructing the reasons behind a chosen course of medical action, a condition that underlies accountability (Coeckelbergh 2020). In fact, also in this case, it is not clear whether clinicians should revise their beliefs in cases of disagreement and align them with the AI's output since questions of epistemic authority are also tied to other normative considerations, such as those related to the existing asymmetry in the attribution of responsibility.

Let us now turn to the illustration of these shortcomings of the deference approach with the hypothetical PE-TER scenario outlined above, in which PE-TER recommends full anticoagulation dose treatment, and Dr. Parker suggests close monitoring. In the particular case under scrutiny, PE-TER can effectively support clinicians in the diagnostic process by recognizing patterns that a human clinician might overlook.[5] More specifically, PE-TER can effectively compare the data pertaining to the particular patient in question to the data of patients with similar characteristics and their outcomes. Under this heading, since the patient is stable and otherwise healthy, their situation is evaluated against the background of other comparatively stable and overall healthy patients for whom a full dose of anticoagulants turned out to produce the best outcome. Conversely, clinicians are particularly well-positioned to capture medically relevant information that patients can actively provide through their report of symptoms and lived experience of illness. The way patients communicate about their health status can be idiosyncratic and fraught with contextual specificities that can exceed what an AI system can process (e.g., on occasion, the experience of illness of the patient, which can be relevant for further treatment planning, might not be reducible to diagnostic categories available to the AI system's syntax and thus remain unconsidered). In the particular case we are analyzing, we further assume that the patient appears to the clinician in an overall stable state and expresses worries, due to former negative experiences about taking anticoagulants due to the increased risk of bleeding. Given that the patient is reactive and stable, Dr. Parker does not think that treatment with anticoagulants is absolutely necessary. Moreover, the worries explicitly expressed by the patient relating to the possible side effects of the treatment are additional reasons for the clinician to be even more cautious in moving forward with administering this medication to the patient. Ultimately, these considerations indicate that the epistemic bases of PE-TER and Dr. Parker differ considerably. They are prone to different types of errors and are efficient in capturing distinct relevant information from different sources (i.e., while PE-TER can extract and elaborate large amounts of data at considerable speed, Dr. Parker can understand unique contextual, idiosyncratic factors pertaining to an individual patient).

Ultimately and more generally, we recognize an important shortcoming of the deference approach in that it attributes to AI systems an expert-like status mostly based on claims regarding their epistemic qualities in terms of accuracy and ability to process high amounts of information in a limited time. However, we find the move from its performance to

[5]An example of how this works in dermatology can be found here: Winkler et al. (2023).

attributing to AI an expert status (that in turn warrants or even mandates deferring to its outcomes in cases of disagreement, following the deference approach) too hasty and often insufficiently justified. The reasons for this are twofold.

First, claims about AI systems' high levels of accuracy or even their likelihood to outperform medical professionals, which are frequently present in the literature, are often largely over-inflated and do not thoroughly reflect the current state regarding the real possibilities of AI systems in medical practice (Drogt et al. 2024). Research, thus, tends to overstate what AI systems can realistically achieve in terms of accuracy levels and might unjustifiably question the epistemic standing and authority of human clinicians, taking deferring to the AI's output as the most viable solution. However, as the case of PE under scrutiny indicates, clinical decision-making is more nuanced, and medical decisions are rarely as clear-cut as proponents of the deference approach seem to assume.

Second and relatedly, by putting considerable emphasis on AI systems' performance as one of the main reasons for clinicians to defer their judgment to AI's outputs, the deference approach seems to insufficiently account for the value of other crucial factors that play a decisive role in clinical decision-making. For instance, this account does not seem to duly recognize the value of forms of hard-to-quantify yet highly valuable knowledge, such as clinicians' tacit knowledge (or trained 'gut feeling'), which is often the result of years of experience and involvement in patient care, on top of the consideration of patients' input and narratives (Funer and Wiesing 2024; Durán and Jongsma 2021; Low 2020). In this respect, let us note that Bjerring and Busch (2021) admit that clinicians' tacit knowledge may, on occasion, override the AI's output. However, in these situations, the burden of proof is on clinicians to justify why the output has been overridden. In their own words, "we can imagine how an appeal to such medical know-how may help us explain how a practitioner can be *excused* from acting in accordance with the recommendations of an epistemically superior AI system" (p. 351) (our emphasis). As this passage shows, these authors still explicitly assume that clinicians' trained clinical eye is a subordinate epistemic resource and clinicians should, *prima facie*, defer to AI systems' outputs. It is precisely the assumption of AI systems' epistemic superiority that lies at the heart of the deference approach and, as we argue, possibly leads to the oversimplification of cases of disagreement in medical practice.

Furthermore, striking the right balance between embracing certain risks and adopting a more cautious attitude in decision-making requires considering a variety of clinically relevant factors and often requires more than the system's performance. Accuracy (of the diagnosis) is one value or datum to be considered in the treatment decision-making process, but certainly not the only one. Solely focusing on the performance of AI to justify deference in cases of disagreement depicts a too simplified and binary view of both clinical decisions and management of disagreement that does not do justice to the complexity and the variety of factors that flow into clinical judgment. While this seems like a natural product of the focus on diagnostic tools, it does not properly capture the intertwinement of accuracy considerations, risk assessments, and other values that should play a role in medical decision-making. We will return to this in more detail in Section "Beyond Resolving Disagreement: Kinds of Disagreement in Context and their Normative Implications."

Beyond Overruling

The overruling approach argues more or less the opposite, namely that humans should overrule AI systems in cases of disagreement. This approach aligns neatly with accounts that stress a human-in-the-loop requirement, according to which humans should have the final say over AI-supported decision processes. A prominent specification of this requirement can be found in efforts toward securing meaningful human control, which encompasses the idea that 'human operators and not computers should ultimately remain in control of, and thus morally responsible for, relevant decisions […]' (Santoni de Sio and Van den Hoven 2018, 1). This notion originates from debates about the use of automated weapon systems, but it is also frequently applied to the context of medicine and healthcare (Hille et al. 2023). Particularly with reference to often discussed forms of "monitoring control," Hille and colleagues explicitly refer to human control understood as overriding the decisions of an AI system, if needed (Hille et al. 2023, 6).[6]

[6]Let us clarify that not all formulations of human control over AI systems are incompatible with the more nuanced approach to disagreement that we present in Section "Beyond Resolving Disagreement: Kinds of Disagreement in Context and their Normative Implications." For example, meaningful human control understood as reason-responsiveness (Santoni de Sio and Van den Hoven 2018; Santoni de Sio 2024) takes that AI systems should be overridden if they do not mirror the relevant moral reasons of stakeholders affected by AI systems' decisions (e.g., are not compatible with their values).

The specificity and formulation of human-in-the-loop requirements vary in legal documents and broader guidelines, but share the assumption that humans have to be able to overrule AI systems (Weingart et al. 2003). Legal documents and guidelines explicitly state that 'human-in-the-loop mechanisms should be designed and implemented to perform specific quality checks (e.g., to flag biases, errors, or implausible explanations), and to overrule the AI predictions when necessary' (Lekadir et al. 2025). However, these phrasings are problematic when deciding how to settle a disagreement, because they are so vague that they leave it at the discretion of the human user to interpret when overruling is *necessary*. Thus, the provision does not provide users with actionable information, possibly risking overriding an otherwise reliable system due to contingent aspects pertaining to the human user themselves (e.g., attitudes of algorithmic aversion (Mahmud et al. 2022)). While the overruling account does not *require* humans to overrule the AI system in every case of disagreement, it leaves unspecified under which conditions overruling is appropriate, thus failing to provide orientation on how to act in specific situations of clinician-AI disagreement.

Moreover, and along similar lines, the American Medical Association (2024) recommends that "(c)linical decisions influenced by AI must be made with specified qualified human intervention points during the decision-making process. […] With few exceptions, there generally should be a human in the loop when it comes to medical decision-making capable of intervening or overriding the output of an AI model." Furthermore, the declaration of Montreal states in their principle 9.2 that "(i)n all areas where a decision that affects a person's life, quality of life, or reputation must be made, where time and circumstance permit, the final decision must be taken by a human being and that decision should be free and informed." Relatedly, the AI Act article 14.4d demands that natural persons to whom human oversight is assigned are enabled, as appropriate and proportionate "(d) to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system" (Regulation 2024/1689, 2024). Similar requirements are also voiced in the EU Guidelines for Trustworthy AI in terms of a "human in command approach" understood as the need "to ensure the ability to override a decision made by a system" (HLEG, 2019). Finally and in connection with protecting the principle of human autonomy, the WHO guidance on Ethics and Governance of Artificial Intelligence for Health states the need to ensure that "a clinician can override decisions made by AI systems and that machine autonomy can be restricted and made 'intrinsically reversible'" (WHO 2021).

While the overruling approach has benefits, including reducing the risk of AI errors, faster decision-making, eradicating responsibility gaps, and it may result in higher acceptance of AI use by patients (Lennartz et al. 2021), it has several shortcomings. Again, consider the pulmonary embolism case introduced in Section "Case: Pulmonary Embolism Treatment Decision in Patients with an Increased Risk of Bleeding". The clinician faces various decisional pathways, and the AI system might provide them with relevant information that they might overlook or may provide a treatment option that was not initially envisaged. For example, in the face of the output generated by PE-TER, Dr. Parker might be prompted to reconsider whether no treatment really is the best option for the patient and integrate the AI system's output to the extent that she recommends a short-course anticoagulation treatment. This would constitute a middle-ground solution between her initial plan and the AI system's suggestion. Without entirely following PE-TER, she attributes some value to its output.

However, the overruling account, as laid out in the legal and ethical guidelines, does not seem to allow for similar nuanced approaches in cases of disagreement, thus risking failing to leverage the added value of having an AI involved in decision-making. Overriding the AI system can directly undercut the intended benefits of these systems for medical practice to possibly improve accuracy and patient outcomes. In high-stakes settings such as medicine, the expected benefits of medical AI may not be realized, or worse, patient care faces severe risks, including inefficient care and direct harm. This is not a mere hypothetical: Several cases are already known from the literature where the human overruling of intelligent systems resulted in harm. For example, the Smiler rollercoaster incident at Alton Towers resulted in serious injury to five people, because a human overruled a computerized safety system (The Halliday

These can be well-justified cases of overriding and show that this account does not generally prescribe overriding the system as soon as there is *any kind* of disagreement. What we aim to problematize in this paper are rather clear-cut approaches endorsing a general overruling attitude and that: 1. either do not specify in which cases overruling might not be justified; 2. or simply implicitly assume that overriding the AI will generally lead to better outcomes. We provide some examples of the latter in the remainder of this section. Also note that human-in-the-loop approaches, where humans supervise decisions of AI systems, generally provide a binary decision situation to the human to follow or dismiss the system.

2015).[7] This example shows that cases in which humans unjustifiably overrule computer systems may undermine the prevention of harm and the promotion of well-being. Therefore, clarification is needed on whether overriding AI systems is desirable in cases of disagreement in medical practice. Deferring and overriding, when applied as "all-or-nothing" approaches, may overlook complexities and nuances of medical decision-making.

Beyond Second Opinion

Lastly, there is the "second opinion" approach, as advanced by Kempt and Nagel (2022). Kempt and Nagel argue for the uniqueness of the collaboration between doctors as epistemic agents in diagnostic processes. In their view, human-human collaborations usually mean that each of the agents involved has a chance to improve their knowledge and decision-making capacities by sharing the reasons for their particular beliefs and recommended courses of action. Thereby, both agents can also share responsibility for the decisions reached. Kempt and Nagel argue that neither reason- nor responsibility sharing on eye level is possible in settings where humans and AI collaborate. Therefore, in cases of human-AI disagreement, they recommend "the rule of disagreement": "If a diagnosis provided by an autonomous AI diagnostic system contradicts the initial diagnosis of the physician-in-charge, it shall count as disagreement requiring a second opinion of another physician" (227).

The rule of disagreement has been formulated to resolve disagreements about diagnosis, but can be extended to treatment decisions. The second opinion approach has benefits because it involves a wider array of expertise in cases with uncertainty, and it may result in higher patient acceptance, similar to cases in which patients know that the final decision is in the hands of a human professional (i.e., similar to overruling approaches as discussed in the previous section). It also seems to ensure answerability because the second clinician can settle the disagreement through an exchange of reasons with the physician in charge, thus justifying the final decision. Such answerability need not, but can ground accountability, when things have gone awry.

However, the second opinion approach does not seem to be the best way to go about disagreements on treatment recommendations, as in the case of PE-TER and Dr. Parker, for at least two reasons. First, seeking a further expert opinion does not seem like a viable approach for practical reasons that have to do with the timeframe in which a treatment for PE should be administered. As mentioned in Section "Case: Pulmonary Embolism Treatment Decision in Patients with an Increased Risk of Bleeding", the treatment of PE is only effective when promptly provided. Therefore, turning to a second (human) opinion might unnecessarily delay medical decision-making and eventual treatment provision, ultimately resulting in patient harm.

Second, including a further perspective in a decision-making process that is already fraught with uncertainty could complicate the situation even further by creating additional levels of disagreement that are hard to disentangle, particularly in time-sensitive situations. For example, if the second physician suggested yet another possible treatment (say, thrombolytic treatment), it raises the question of how this further uncertainty and disagreement should be dealt with. The first and second physicians will be able to exchange reasons for their perspectives, but this will not overcome the disagreement with the AI system. Furthermore, the second opinion approach seems to assume that the second physician will overcome the limits and biases of human decision-making and offer a fitting assessment of the situation that is, in turn, able to settle the initial disagreement. This seems to be a problematic assumption, particularly in situations in which AI systems demonstrably lead to suitable recommendations.

BEYOND RESOLVING DISAGREEMENT: KINDS OF DISAGREEMENT IN CONTEXT AND THEIR NORMATIVE IMPLICATIONS

The deference, overruling, and second opinion approach have several assumptions in common namely: 1) that human-AI disagreements should (always) be resolved, 2) that disagreement should be resolved by epistemic authority, 3) that cases where AI and clinicians agree are unproblematic, and 4) that humans and AI systems are in an antagonistic relation in a situation of disagreement.

[7] Let us note that we mention this example even if it is not connected to the use of AI in medical practice because it is paradigmatic of the risks connected to the overruling approach we aim to problematize. While there is a lack of documentation of similar incidents with the use of medical AI, it is important to be aware of these issues to prevent possible future problematic outcomes due to unjustified overruling of an AI system. While cases of possible medical malpractice arising when AI systems displace human clinicians abound in the literature (see, e.g., Bannon 2023), cases of patient harm occurring because an otherwise correct AI decision is unjustifiably overruled seem largely under-researched.

First, disagreement needs not always be seen as a problem that has to be overcome. Instead, it can serve as a productive condition that fosters, for example, epistemic humility and encourages careful deliberation (Parker 2025). In scientific inquiry, disagreement may prompt the refinement of methods, theories, and assumptions, contributing to the robustness of knowledge over time. Similarly, in political philosophy, sustained disagreement can reflect the plurality of values in a democratic society, where premature consensus may risk marginalizing minority voices (Hannon 2024). Rather than hastily resolving disagreements, actors might benefit from slightly adjusting their initial positions or adopting an incremental approach, allowing for ongoing evaluation of outcomes. This process acknowledges that the value of disagreement lies not only in eventual convergence but also in its capacity to enhance reflexivity, adaptability, and responsiveness in complex decision-making environments such as medicine.

Naturally, the epistemic opacity that characterizes many AI systems currently deployed in medicine can complicate this deliberative practice. In this regard, it is crucial to consider methods that can enrich the epistemic environment in which clinicians interact with AI systems, thereby making space for deliberation (Pozzi et al. 2025). To this end, often-discussed candidates are explanatory AI (XAI) methods that aim to reconstruct the system's decision-making process via, for example, the provision of post-hoc explanations. While sometimes helpful, these methods are not a panacea for addressing the black box problem. Additionally, there is literature casting doubts on whether the availability of explanations effectively increases a user's understanding of the system, which in turn would enable critical engagement with it (Babic et al. 2021). Moreover, there are also indications that explanations might even backfire as they might provide a false sense of justification (Bucinca et al 2021). Other methods to assess the reliability of an AI system are available in the literature, which are externalist to the algorithm insofar as they do not try to remedy their opacity. As Durán and Jongsma (2021) argue, one might consider reliability indicators as a means to assess whether a user is (epistemically) justified in believing the system's output. While the black-box nature of AI systems may still pose an obstacle to critical scrutiny, these and other methods can provide clinicians with crucial information needed to evaluate potential disagreements.

Second, the three approaches differ regarding to whom they attribute epistemic authority in cases of disagreement: while the overruling and second opinion approaches consider humans to be the epistemic authority in cases of disagreement (the latter because of humans' reason-giving ability, the former to sustain responsibility), the deference approach considers the AI system the epistemic authority (due to its ascribed accuracy). However, examining the problem through the lens of epistemic authority might be problematic altogether, irrespective of who (or what) the authority is assigned to. To ground this claim, consider the definition of epistemic authority advanced by Keren (2007)[8] in terms of the normative power of giving a special kind of reason for belief: "by expressing her belief that p, a person who has authority on p does not merely give us a reason to likewise believe that p. More than that, *she gives us a second-order, preemptive reason for disregarding other relevant evidence which we may have concerning p*" (ibid, p. 373, emphasis added). According to this definition, information coming from an epistemic authority is not additional evidence; it is rather what one decides to follow *instead of* looking for (additional) evidence. This view on epistemic authority seems to underlie an antagonistic approach, according to which disagreement is something to be resolved (by deferring to the AI system or human clinician, depending on which view one endorses). However, this proves problematic in both cases. Attributing epistemic authority to AI systems is problematic as the evidence they produce often needs to be contextualized and balanced against other available pieces of evidence, and, thus, should not entail this kind of normative power. Crucially, let us note that more often than not, AI systems provide input that needs to be included in a broader range of considerations pertaining to a patient's clinical situation. Cases in which the AI system's output can be considered a final, standalone solution for providing evidence (e.g., imaging applications to detect diabetic retinopathy) are the exception rather than the rule.[9] Second and conversely, attributing epistemic authority to clinicians (as the overruling and second opinion approaches entail), limits the possible epistemic contribution of AI systems to clinical deliberation and

[8]We consider this definition of epistemic authority because, contrary to others often referred to in the literature (e.g., Zagzebski 2012), it is not based on a virtue epistemological account of epistemic authority and can thus be applied to our considerations pertaining to human-AI interactions. For an analysis of epistemic authority in AI following Zagzebski's account see Ferrario et al. (2024).

[9]Let us also point out that attributing epistemic authority to the AI system might exacerbate a critical over-reliance on it, thus unjustifiably reducing clinicians' appropriate engagement with the case at hand. Research has expanded to show the risks of over-reliance in the form of automation bias (Goddard et al. 2014).

thus risks failing to harness the benefits of having an AI system involved in the decision-making process in the first place. These considerations thus seem to urge us to move beyond the epistemic authority paradigm altogether and embrace a more nuanced and less antagonistic approach.

Relatedly, let us take notice of the fact that, so far much of the literature on disagreement idealizes and simplifies decision-making. On the one hand, standard debates in the epistemology of disagreement mostly consider the question of whether a subject should revise their beliefs in view of a disagreement with an epistemic peer in situations in which there is a finite set of verifiable answers (Christensen 2007). On the other hand, literature on clinician-AI disagreement tends to oversimplification partly because of the focus on diagnostics rather than the more complex situation of treatment decision-making (the concern of oversimplification is also shared by, e.g., Kempt et al. 2023 even though the authors reach different conclusions than those we advance in our analysis).

However, we maintain that disagreements in medical cases require a more nuanced analysis. This is the case not only because illness is not binary (cancer is here a very obvious example, where not just occurrence, but stages and grades need to be distinguished, which both ought to inform treatment decisions (see Plutynski & Laplane, 2023)), but also because treatment is a complex collective action process that needs to align both with patient values and relevant moral and social values (Mukherjee 2015). This means that a disagreement between a clinician and AI system opens a plethora of possible responses both epistemically and practically. Epistemically, rather than merely overruling or sustaining one's judgment in a disagreement situation, after a) suspending and carefully reconsidering one's belief, which is in itself urged through the disagreement, the physician could b) reconcile the two opposing recommendations, e.g. adjust one's assessment of the grade or stage of a disease, for example, rather than completely rejecting the idea that the patient has the disease. Furthermore, c) additional evidence could be taken up, or d) a diagnostic process could be repeated to test the reliability of either of the two beliefs. In terms of the practical treatment decision, the possibilities are even vaster: Rather than suggesting a standard treatment, a) the treatment options too could be temporarily suspended, to establish a clearer diagnostic picture, b) drug dosages could be adjusted, c) the originally intended treatment type could be changed (again, in cancer treatment one could advise a change from chemotherapy to surgery or both, in psychology from behavioral to exposure therapy), d) if doubts arise about a possible different underlying cause, the patient could be referred to a different specialist. These non-exhaustive examples suggest that the disagreement opens these productive avenues, which straightforward overruling or deferring approaches, with the antagonistic assumptions that underlies them, neglect.

Third, it is assumed that situations in which physicians and AI agree are not problematic. Yet, the *reasons* for agreement between the AI system and a human doctor might be substantially different, and they could still both be wrong (Jongsma and Sand 2022). Specifically, the second opinion approach treats the AI purely as a confirmation machine, requiring explanation and justification only when the responsible physician disagrees with its output. Without prior knowledge about which of the two is right or wrong, such a differential response is unwarranted. Rosenbacke and colleagues (2025) indicate that cases in which a clinician and an AI agree but are both wrong amount to an often overlooked issue that they call "false confirmation." This scenario is, according to these authors, particularly problematic in medicine because it can wrongfully reinforce a clinician's belief that they are making the right decision. Moreover, cases of false confirmation could lead clinicians to accept the AI system's output without critically questioning its suitability, simply because it happens to be aligned with the preferred course of action (Rosenbacke et al. 2025).

Fourth, within the three approaches, there is an implicit assumption that disagreement indicates an antagonistic relationship between physicians and AI systems. However, this framing overlooks the possibility that disagreement can persist within cooperative and mutually respectful relationships. Good collaborators often encounter and even rely upon sustained disagreement as a means of testing ideas and improving outcomes. In such relationships, disagreement is not a sign of dysfunction but rather a marker of intellectual engagement and shared commitment to a rigorous process. Even though, as Krishnan (2020) points out, standard debates in the epistemology of disagreement do not consider those occurring between radically heterogeneous epistemic entities, such as those between humans and AI systems, we can argue that if they track different indicators, they are to be seen as independent sources. This means, in turn, that disagreements do not necessarily need to be overcome by uncovering a possible mistake on either side; rather, a disagreement might be indicative of different, but complementary, epistemic perspectives on the matter at hand. This consideration further supports the need to go beyond the assumption that disagreements

reflect an antagonistic relationship between AI and human clinicians that seems to underlie the three approaches previously analyzed.

Ultimately, the three approaches considered share an overly schematic and hypothetical analysis of cases in which disagreements emerge that do not do justice to the nuanced nature of medical practice. By highlighting their limitations, our previous discussion has paved the way to a more refined understanding of disagreement. Disagreement has various sources that need to be more sharply distinguished. Depending on the source of disagreement, different responses are warranted. First, disagreement might pertain to questions about what is used as evidence: Our envisioned AI tool, PET-ER, does not have access to relevant evidence that is available to Dr. Parker regarding the case at hand. PE-TER might base its decision primarily on an analysis of the CT scan, where blood clots are detected. While a patient record might also be available to PET-ER, Dr. Parker might - as mentioned before - gain an immediate visual impression of the (physical) state of the patient and, therefore, an indication of how much the patient is suffering and whether a treatment decision contributes to lowering the suffering of the patient (Anjum et al. 2020). It seems crucial that the more complex PET-ER is in the types of evidentiary sources that it can process, the more opaque it becomes to Dr. Parker which factors have been given more or less weight in generating its recommendation (see, e.g., London 2019).

Moreover, it is clear that there can be disagreements about evidence without disagreements about treatment: the same treatment could be administered based on two different sources of evidence, such as the physical and verbal expressions of pain of the patient and lab data about blood toxin levels. Similarly, disagreements about treatment options may arise despite agreements about evidence. PET-ER might base a suitable treatment recommendation purely on the CT scan, while Dr. Parker, in agreement with those results but aware of the idiosyncratic data about the physical state of the patient that is unavailable to PET-ER at this point, rejects its treatment decision. Moreover, it is important to note that disagreement about evidentiary sources can be productive: Dr. Parker might, for instance, decide against full-coarse anticoagulation treatment in our previous example, but synthesize PE-TER's recommendation and opt for a short-course anticoagulation treatment. In this case, she *did not overrule* PE-TER's recommendation. She does value and incorporate the recommendation into her decision-making; otherwise, she would have refrained from administering a treatment altogether. Hence, her decision is not against the AI. Since the disagreement is a product of differential access to evidence (i.e., physical perception of the patient or merely a CT scan), Dr. Parker's decision is one that is better informed through the consideration and contextualization of PE-TER's output (this is thus in stark contrast with approaches that rely on considerations of epistemic authority instead of incorporating the AI's output against the background of other relevant evidence). Ultimately, disagreement in medicine can indicate reasonable uncertainty and make us behave more cautiously. Moreover, it may foster critical thinking, enhance the rigor of evidence, and contribute to innovation (in terms of practical ways forward, whether that concerns the goal of curing, stabilizing, or relieving pain, among others). Recognizing such uncertainty can thus be valuable, as it may also lead to more informed and better overarching opinions and decisions.

Furthermore, the latter situation might also be framed as a disagreement about the relative weight of evidence and which practical conclusions to draw from it. Diagnostic and treatment decisions are fundamentally evaluative judgments for which risks and uncertainties have to be weighed against a backdrop of medical knowledge, expert knowledge, and intuitions (Durán and Jongsma 2021). Assume that PET-ER can process information about the fragility of the patient related to previous health issues (e.g., history of bleeding complications). What weight will it attribute to those previous health issues when suggesting a treatment? Can it evaluate their severity - can that severity be quantified without taking into account what the patient would gain from the treatment - hence can it be quantified without making a complex reference to the benefits? Risk assessments are often comparative in this regard: all treatments come with risks and, therefore, should be assessed for their proportionality and against a certain standard. If there is no objective baseline for the acceptability of certain risks, then disagreements are inevitable in medicine. The baselines with which Dr. Parker compares the risks of the patient might simply be different than the ones PET-ER uses.

Lastly, these considerations tie into questions of value. Disagreements about what is valuable can lead to considering new values that have not been previously taken into account. While it initially may seem that disagreement concerns only epistemic values (e.g., accuracy or reliability), the interaction between AI and doctors may give rise to a broader set of considerations and values. Many patients enter a healthcare setting to be cured or relieved from pain, but this is

not the only thing that matters to them. They may want to understand their conditions, assess risks of different treatment options in the light of their broader views and values, and be seen and heard by their health care provider. In the particular case under consideration, PE-TER might recommend a full-dose anticoagulation treatment to the patient, which can be interpreted as a risk-averse decision aimed at preventing any chance of acute PE. However, Dr. Parker might be inclined to recommend close monitoring and no medication to avoid the risk of excessive bleeding, i.e., thereby also acting in a risk-averse fashion, but for different reasons. PE-TER and Dr. Parker might, thus, suggest different courses of action based on the same principle of risk aversion.[10]

Furthermore, if there is a disagreement about the treatment recommendation (e.g., administer blood thinners or not), on the basis, for instance, of different evaluations of the risks of the treatment for the patient, but agreement about the evidentiary sources (CT scan showing embolism), then a new duty arises for Dr. Parker, that includes the value of monitoring and transparency. She will have to inform the patient about the diagnosis of the pulmonary embolism (PE-TER and Dr. Parker are in agreement about this), and potentially base her judgment partially on the AI output. This value or duty would not emerge if, in the case of agreement, there is, as the previous account holds, no problem at hand, or in the case of disagreement, she is attributed the epistemic authority to overrule the AI output and, thereby, declare its input nil. All in all, we have argued that in all these cases, disagreement is often viewed as an obstacle to a swift medical decision-making process, but actually has the potential to enhance medical decision-making, both in its process and its outcome.

FINAL REMARKS

As AI systems are increasingly applied in clinical care, situations where human-AI disagreements emerge will become more frequent. This paper outlines three proposed ways of overcoming human-AI disagreements, namely the *deference*, the *overruling*, and the *second-opinion* approach. We argued that the outlined approaches fall short of acknowledging the benefits and value that disagreement may have for clinical reasoning and clinical decision-making. We maintained that overcoming human-AI disagreement is not always the best way to handle disagreement, as disagreement can indicate reasonable uncertainty about evidence, risk assessment, and relevant values that would otherwise go unnoticed. Sustaining disagreement should make us behave more cautiously, may foster critical thinking, enhance the rigor of evidence, and contribute to innovation. Recognizing such uncertainty can thus be valuable, as it may also lead to reaching better overarching views and medical decisions while harnessing the benefits of involving AI systems in medical practice. We conclude that it is not desirable to have a standard protocol to handle all occasions of all types of disagreement. The sources of disagreement are too varied, and the possibility that both parties are right, but for different reasons, or that they are both wrong, suggests that it would be ill-advised to develop a go-to heuristic pathway to resolve human-AI disagreement in medicine. It must be accepted that AI systems will contribute new data to the clinical diagnostic and decision-making process. Depending on context, patient condition, and physician knowledge and confidence, this information must be taken into account.

[10]Durán and Jongsma (2021) advance similar considerations regarding different possible interpretations of the evidence provided by an AI system following the principle of safety.

ACKNOWLEDGMENTS

We thank Geert-Jan Geersink for helping us draft the hypothetical case and Alberto Giubilini for his thoughtful comments on an earlier draft of this manuscript.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

FUNDING

KJ received funding from NWO [VI.Vidi.221F.010].

ORCID

Giorgia Pozzi http://orcid.org/0000-0001-8928-5513
Martin Sand http://orcid.org/0000-0001-8167-4581
Karin Jongsma http://orcid.org/0000-0001-8135-6786

REFERENCES

Allena N, Khanal S. 2023. The algorithmic lung detective: artificial intelligence in the diagnosis of pulmonary embolism. Cureus. 15(12):e51006. https://doi.org/10.7759/cureus.51006

American Medical Association (AMA). 2024. Augmented intelligence development, deployment, and use in health care. https://www.ama-assn.org/system/files/ama-ai-principles.pdf

Anjum RL, Copeland S, Rocca E. 2020. Medical scientists and philosophers worldwide appeal to EBM to expand the notion of 'evidence'. BMJ Evid Based Med. 25(1):6–8. https://doi.org/10.1136/bmjebm-2018-111092

Babic B, Gerke S, Evgeniou T, Cohen IG. 2021. Beware explanations from AI in health care. Science. 373(6552):284–286. https://doi.org/10.1126/science.abg1834

Babushkina D, Votsis A. 2022. Epistemo-ethical constraints on AI-human decision making for diagnostic purposes. Ethics Inf Technol. 24(2):22. https://doi.org/10.1007/s10676-022-09629-y

Bannon L. 2023 Jun 15. When AI overrules the nurses caring for you. The Wall Street Journal. https://www.wsj.com/articles/ai-medical-diagnosis-nurses-f881b0fe

Baumgartner C, Tritschler T. 2021. Clinical significance of subsegmental pulmonary embolism: an ongoing controversy. Res Pract Thromb Haemost. 5(1):14–16. https://doi.org/10.1002/rth2.12464

Bělohlávek J, Dytrych V, Linhart A. 2013. Pulmonary embolism, part I: epidemiology, risk factors and risk stratification, pathophysiology, clinical presentation, diagnosis and nonthrombotic pulmonary embolism. Exp Clin Cardiol. 18(2):129–138.

Bjerring J, Busch J. 2021. Artificial intelligence and patient-centered decision-making. Philos Technol. 34(2):349–371. https://doi.org/10.1007/s13347-019-00391-6

Buçinca Z, Malaya MB, Gajos KZ. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc ACM Hum-Comput Interact. 5(CSCW1):1–21. https://doi.org/10.1145/3449287

Burrell J. 2016. How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. 3(1):1–12. https://doi.org/10.1177/2053951715622512

Christensen, D. (2007). Epistemology of disagreement: the good news. Philos Rev. 116(2):187–217. https://doi.org/10.1215/00318108-2006-035

Coeckelbergh M. 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics. 26(4):2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Drogt J et al. 2024. Ethical guidance for reporting and evaluating claims of AI outperforming human doctors. NPJ Digit Med. 7(1):271. https://doi.org/10.1038/s41746-024-01255-w

Durán J, Jongsma K. 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics. 47:medethics-2020-106820. https://doi.org/10.1136/medethics-2020-106820

Duwe G et al. 2024. Challenges and perspectives in use of artificial intelligence to support treatment recommendations in clinical oncology. Cancer Med. 202413(12):e7398. https://doi.org/10.1002/cam4.7398

High-Level Expert Group on Artificial Intelligence. 2019. Ethics guidelines for trustworthy AI. Brussels: European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hille EM, Hummel P, Braun M. 2023. Meaningful human control over AI for health? A review. J Med Ethics. https://doi.org/10.1136/jme-2023-109095

Ferrario A, Facchini A, Termine A. 2024. Experts or authorities? The strange case of the presumed epistemic superiority of artificial intelligence systems. Minds Mach. 34(3):30. https://doi.org/10.1007/s11023-024-09681-1

Funer F, Wiesing U. 2024. Physician's autonomy in the face of AI support: walking the ethical tightrope. Front Med (Lausanne). 11:1324963. https://doi.org/10.3389/fmed.2024.1324963

Goddard K, Roudsari A, Wyatt J. 2014. Automation bias: empirical results assessing influencing factors. Int J Med Inform. 83(5):368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Grote T, Berens P. 2020. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 46(3):205–211. https://doi.org/10.1136/medethics-2019-105586

Halliday J. 2015, November 24. Alton Towers Smiler ride crash caused by human error, theme park says. The Guardian. https://www.theguardian.com/uk-news/2015/nov/24/alton-towers-rollercoaster-crash-caused-by-human-error-theme-park-says

Hannon M. 2024. Disagreement and contemporary political philosophy. In: Baghramian M, Carter JA, Cosker-Rowland R, editors. The Routledge handbook of philosophy of disagreement. Routledge. p 385–397.

Jongsma K, Sand M. 2022. Agree to disagree: the symmetry of burden of proof in human–AI collaboration. J Med Ethics. 48(4):230–231. https://doi.org/10.1136/medethics-2022-108242

Kempt H, Heilinger JC, Nagel SK. 2023. "I'm afraid I can't let you do that, Doctor": meaningful disagreements with AI in medical contexts. AI & Soc. 38(4):1407–1414. https://doi.org/10.1007/s00146-022-01418-x

Kempt H, Nagel S. 2022. Responsibility, second opinions and peer- disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts. J Med Ethics. 48(4):222–229. https://doi.org/10.1136/medethics-2021-107440

Keren, A. (2007). Epistemic authority, testimony and the transmission of knowledge. Episteme. 4(3):368–381. https://doi.org/10.3366/e1742360007000147

Krishnan M. 2020. Against interpretability: a critical examination of the interpretability problem in machine learning. Philos Technol. 33(3):487–502. https://doi.org/10.1007/s13347-019-00372-9

Lekadir K et al. 2025. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 388:e081554. https://doi.org/10.1136/bmj-2024-081554

Lennartz S et al. 2021. Use and control of artificial intelligence in patients across the medical workflow: single-center questionnaire study of patient perspectives. J Med Internet Res. 23(2):e24221. https://doi.org/10.2196/24221

London A. 2019. Artificial Intelligence and Black-Box Medical Decisions: accuracy versus Explainability. Hastings Cent Rep. 49(1):15–21. https://doi.org/10.1002/hast.973

Low M. 2020. Above and beyond statistical evidence. why stories matter for clinical decisions and shared decision making. In Anjum RL, Copeland S, Rocca E, editors. Rethinking causality, complexity and evidence for the unique patient. Springer. p 127–136.

Mahmud H et al. 2022. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Technol Forecast Soc Change. 175:121390. https://doi.org/10.1016/j.techfore.2021.121390

Mittelstadt BD et al. 2016. The ethics of algorithms: mapping the debate. Big Data Soc. 3(2):1–12. https://doi.org/10.1177/2053951716679679

Mishra R, Shridevi S. 2024. Knowledge graph driven medicine recommendation system using graph neural networks on longitudinal medical records. Sci Rep. 14(1): 25449. https://doi.org/10.1038/s41598-024-75784-5

Montréal Declaration for a Responsible Development of Artificial Intelligence. 2018. https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl_IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf

Mukherjee S. 2015. The laws of medicine - field notes from an uncertain science. Simon & Schuster.

Parker MJ. 2025. Bioethics and the value of disagreement. J Med Ethics. 52(1):7–13. https://doi.org/10.1136/jme-2024-110174

Pozzi G, Buijsman S, van den Hoven J. 2025. Harmful epistemic dependence on medical machine learning and its moral implications. J Med Ethics. 52(1):42–49. https://doi.org/10.1136/jme-2024-110552

Plutynski A, Laplane L. 2023. Cancer. In: Zalta EN, Nodelman U, editors. The Stanford encyclopedia of philosophy (Winter 2023 Edition). https://plato-stanford-edu.tudelft.idm.oclc.org/archives/win2023/entries/cancer/

Regulation (EU). 2024/1689. European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Rosenbacke R et al. 2025. AI and XAI second opinion: the danger of false confirmation in human–AI collaboration. J Med Ethics. 51(6):396–399. https://doi.org/10.1136/jme-2024-110074

Santoni de Sio F. 2024. Human freedom in the age of AI. Routledge.

Santoni de Sio F, Van den Hoven J. 2018. Meaningful human control over autonomous systems: a philosophical account. Front Robotics AI. 5:323836.

van Leersum CM, Maathuis C. 2025. Human centred explainable AI decision-making in healthcare. J Responsib Technol. 21:100108. https://doi.org/10.1016/j.jrt.2025.100108

van Maanen R et al. 2025. YEARS clinical decision rule for diagnosing pulmonary embolism: a prospective diagnostic cohort follow-up study in primary care. BMJ Open. 15(2):e091543. https://doi.org/10.1136/bmjopen-2024-091543

Weingart SN et al. 2003. Physicians' decisions to override computerized drug alerts in primary care. Arch Intern Med. 163(21):2625–2631. https://doi.org/10.1001/archinte.163.21.2625

Winkler JK et al. 2023. Assessment of diagnostic performance of dermatologists cooperating with a convolutional neural network in a prospective clinical study: human with machine. JAMA Dermatol. 159(6):621–627. https://doi.org/10.1001/jamadermatol.2023.0905

World Health Organization (WHO). 2021. Ethics and governance of artificial intelligence for health: WHO guidance. https://iris.who.int/bitstream/handle/10665/341996/9789240029200-eng.pdf?sequence=1

Zagzebski L. 2012. Epistemic authority: a theory of trust, authority, and autonomy in belief. Oxford University Press.

Zuo Z et al. 2021. Thrombolytic therapy for pulmonary embolism. Cochrane Database Syst Rev. 4(4):CD004437. https://doi.org/10.1002/14651858.CD004437.pub6