Technology in Society 81 (2025) 102791

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

An interdisciplinary perspective on AI-supported decision making in medicine

Jonas Ammeling [a], Marc Aubreville [b], Alexis Fritz [c], Angelika Kießig [d], Sebastian Krügel [a], Matthias Uhl [e,*]

[a] Technische Hochschule Ingolstadt, Esplanade 10, D-85049, Ingolstadt, Germany
[b] Hochschule Flensburg, Kanzleistr. 91-93, 24943, Flensburg, Germany
[c] Albert-Ludwigs-Universität Freiburg, Friedrichstr. 39, 79098, Freiburg, Germany
[d] Katholische Universität Eichstätt-Ingolstadt, Ostenstr. 28a, D85072, Eichstätt, Germany
[e] Universität Hohenheim, Schloss Hohenheim 1, 70599, Stuttgart, Germany

ARTICLE INFO

Keywords:
Machine learning
Decision-support system
Human-machine interaction
Medical image diagnosis
Ethically aligned design
Value-based design

ABSTRACT

Artificial intelligence (AI)-supported medical diagnosis offers the potential to utilize the collaborative intelligence of context-sensitive humans and narrowly focused machines for patients' benefit. The employment of machine-learning-based decision-support systems (MLDSS) in medicine, however, raises important multidisciplinary challenges that cannot be addressed in isolation. We discuss three disciplinary perspectives on the topic and their interplay. Ethical issues arise at the level of changing responsibility structures in healthcare. Behavioral issues relate to the actual impact that the system has on physicians. Technical issues arise with respect to the training of a machine learning (ML) model that gives accurate advice. We argue that the interaction between physicians and MLDSS including the concrete design of the interface in which this interaction occurs can only be considered at the intersection of all three disciplines.

1. Introduction

Algorithms permeate our society. One important field of application is delivering advice in medical decision-making. In this domain, the use of the collaborative intelligence of context-sensitive humans and narrowly focused AI-powered machines promises substantial benefits for patients [1]. It could also help to reduce exploding costs in the healthcare sector as physicians may be able to save valuable time resources by utilizing decision support systems [2]. Inter alia, algorithms can be used to support physicians when making a diagnosis based on the content of medical images.

The utilization of algorithms in medical decision-making, however, raises many questions and causes challenges at the intersection of several scientific disciplines. From the growing literature on AI-generated advice giving, we identified three distinct scientific perspectives that will have to be synthesized to achieve progress in human-centered MLDSS. The overarching guiding perspective is the ethical one. The task of this perspective is to discuss the appropriateness of traditional philosophical concepts like autonomy and responsibility for the realm of MLDSS. An essential question to be addressed here is whether the emergence of MLDSS in the context of medical decision-making causes responsibility gaps and whether the concept is adequate to capture the challenges of hybrid decision-making. To this purpose, we will argue, it may be helpful to consider the epistemic condition and the control condition separately. The epistemic condition refers to the question of how the generation of medical knowledge is transformed if the interaction between humans is increasingly replaced by hybrid interaction. The control condition tackles the problem of many hands in medical decision-making and problematizes how the attribution of responsibility may be maintained.

To avoid that well-intended normative ideas unfold undesired consequences, the ethical perspective has to be complemented by the behavioral perspective. This perspective focuses on the empirical phenomena that arise from the interaction of humans and MLDSS. One key factor when it comes to using algorithms in the medical domain is physicians' trust in these systems [3]. The extent to which physicians rely on automated decision support will depend largely on this. It is owed to the fact that the mediation of human behavior by technology shapes this behavior through feeding back into preferences and beliefs. Numerous studies in psychology and behavioral economics have

* Corresponding author.
E-mail address: matthias.uhl@uni-hohenheim.de (M. Uhl).

https://doi.org/10.1016/j.techsoc.2024.102791
Received 10 November 2023; Received in revised form 27 October 2024; Accepted 4 December 2024
Available online 5 December 2024

investigated the factors that influence trust between human advisors and advisees and established a phenomenon coined egocentric discounting. This describes an advisees' tendency to give the advice inadequately little weight by primarily relying on their own beliefs. There is also growing research that studies human advisees' trust in AI-based advisors that provides mixed evidence with some studies documenting less trust in algorithmic than human advisors ("algorithm aversion") and others documenting the opposite ("algorithm appreciation"). Gaining a more systematic understanding of the drivers of trust in algorithms seems crucial to design a successful interaction between physicians and MLDSS.

Finally, the technological perspective deals with the specific materialization of the interaction between human decision-makers and MLDSS. There exist, of course, challenges that are of technological nature. These relate to training the algorithmic model and to fostering its accuracy in terms of a trade-off between specificity and sensitivity, i.e., optimizing the tension between avoiding false positives and false negatives alike. Given the widening space of technological opportunities, the scope of this perspective in isolation is only restricted by technical feasibility. It is by the empirically informed normative concepts derived from the first two perspectives that further restrictions are imposed on technological implementation. Not everything that is feasible should be done and other things that may be desirable may not (yet) be feasible. It is the task of the technological perspective to shift the limits of what is feasible. Technological progress in the realm of AI raises new ethical problems of opacity in algorithmic advice generation, that are now being addressed by technological advances in Explainable AI (XAI).

A lot of research on hybrid interactions between humans and AI-based advisors is already done in the behavioral sciences and the technological sciences. In our opinion, there is however a lacking exchange of these fields with technology ethics. Ethics is often detached from the empirical disciplines and engages in normative reflection on an often rather abstract level. Sometimes it remains unclear how these reflections are to be considered in a concrete technological interaction design. We therefore propose an interdisciplinary approach in which ethicists, behavioral scientists and computer scientists work jointly and in close exchange on the challenge of human-centered design by bringing in their own respective expertise. In this context, it should be emphasized that the relationship of the three perspectives is not reducible to a top-down relationship from the ethical via the behavioral to the technical perspective. The behavioral and technical implications of a specific design of the interaction between users and MLDSS may give rise to ethically relevant phenomena that will need to be subjects of reflection on the ethical level. This is particularly true of emergent ethical phenomena that were hard to anticipate in a world prior to the technological development in question.

It is the aim of this article to elaborate on the multi-faceted problem of using MLDSS in medical decision-making by discussing three different disciplinary perspectives and their dependencies. Our intention is to show that the societal potential of utilizing ML models, in particular deep models, in medical decision-making can only be raised by addressing this technology holistically and in an interdisciplinary manner.

To this aim, our article proceeds as follows. We will first discuss each perspective in isolation and summarize important problems viewed from each discipline separately. Afterwards, we will conclude by arguing for the need of synthesizing the three perspectives and by calling for an interdisciplinary research agenda on MLDSS in medicine.

2. Three disciplinary perspectives on AI-supported decision making in medicine

2.1. Ethical perspective

The responsibility structure of classical medical diagnostic processes is – at least legally – clearly regulated, insofar as the attending physicians are responsible for the respective medical diagnosis as well as the subsequent indication and therapy plans [4]. In the discussion about the integration of MLDSS into these classical medical diagnostic processes, it is emphasized, in addition to the enumeration of weighty opportunities, that the effective use of algorithms can produce responsibility diffusion or responsibility gaps with regard to the diagnosis [5]. This in turn can have a particularly negative impact on patients and their relatives [6].

2.1.1. Responsibility gaps in medical decision-making?

In principle, responsibility gaps in the field of technology ethics in their original sense describe a situation in which a complex as well as opaque autonomous system decides or acts independently, making both human control and insight into its decision almost impossible [7,8]. However, since moral responsibility presupposes both "control" and "epistemic" cognition (Aristotle cited in Ref. [9,10]), a gap in the attribution of responsibility arises in such scenarios [7] – because the autonomous system cannot be morally responsible either, due to the fact that it is not a moral actor [11,12]. In contrast, with regard to the use of non-autonomously functioning MLDSS in the medical diagnostic process, it can be stated that this standard understanding of responsibility gaps falls short. This is because the use of the algorithm within medical diagnostics merely constitutes an element of the physician's final diagnosis and does not describe the autonomous final decision [13]. The result of an MLDSS is therefore considered one criterion among others, which the physician has to include and evaluate in the diagnosis. The physician remains in control; she is the final decision-maker. Thus, the diagnostic process remains the task of the medical professional and consequently the entire moral responsibility for the diagnosis can clearly be attributed to her, or so it seems.

For even if in the medical context it is therefore less a question of "standard" responsibility gaps – there is no "loss of control" in the medical field caused by autonomous AI taking final decisions – the question of responsibility in the medical diagnostic process has not yet been fully clarified. In order to gain a more differentiated understanding of the attribution and structure of responsibility within a diagnostic process that utilizes MLDSS, the diagnostic situation should be reflected on with the help of the conditions of moral responsibility already described above: the "epistemic" and the "control" condition. In the context of MLDSS, the epistemic condition refers to the question of how knowledge is produced if a medical specialist "cooperates" with an MLDSS that feeds information into the specialist's decision-making process. The control condition builds on this condition, as the synthesized knowledge produced will co-determine the decision that is taken by the human, and raises the specific questions of who is actually in control in hybrid constellations of physicians and MLDSS and how substantial human control can be maintained in these contexts of many hands.

2.1.2. The epistemic condition

Let us start with some thoughts on the epistemic condition: A physician is considered a "domain-relative 'epistemic authorit[y]'" [14] – an expert in her respective medical field – who, with regard to the diagnostic process, has the task of reflectively applying her medical knowledge. This knowledge is composed of a "*variety of methods and heuristics* (e.g., 'consensus conferences', 'evidence-based medicine', 'translational medicine', and 'narrative medicine')" [14] to the patient's specific individual disease situation (see also [15]). However, this "interpretative capacity that examines, evaluates, and selects the existing medical knowledge" [14] is considerably undermined by the opaque nature of the algorithm, insofar as it is only possible to a limited extent for the medical professional to interpret the result of the MLDSS and thus become aware of faulty outputs [11].

Nevertheless, this does not imply that the focus should be on the MLDSS as a "technical disruptive factor" when attributing responsibility – conventional medicine is by no means free of knowledge gaps either

[16]. Instead, the focus should be on the contextual transformations and differences that arise during a diagnostic process that utilizes MLDSS. The handling of "second opinions" becomes crucial in this context. Even though an evaluation of a MLDSS is just one piece of information that a physician considers when making a diagnosis, it still must be taken into account. In a conventional diagnostic process involving two physicians, there can be an exchange of arguments in the event of opposing views. However, this is not possible in the case of a ML-supported diagnosis [17]. This limitation is not only due to the fact that the evaluation of a MLDSS is based on statistical correlations and differs from a physician's judgment but also because a MLDSS cannot engage in a dialogue to provide arguments, reasons or explanations – even if the ML-supported result can be interpreted using XAI methods. Philosophically, questions arise of what constitutes an explanation. One may insistthat the physician is in fact not given discursive reasons by the "partner." While methods using explanatory cycles (e.g., Ref. [18]) try to approximate such discourses, they will fall short of respective human capabilities in the eyes of many. Ultimately, what constitutes a convincing advice or meaningful explanation still depends on the physician's – or the patient's – judgment [14,17].

In this context, it is particularly important to focus on the patient (and, if relevant, her relatives). From the physician's duty to inform, the principle of "informed consent" or the model of a deliberative physician-patient relationship with "shared agency", it becomes clear that the epistemic aspect and the moral responsibility of physicians are inherent in a relational dimension from the very beginning – namely towards the patients [19,20]. It is the individual situation of the patient that requires a diagnosis, and this diagnosis is to be justified towards her. The patient herself is also part of the diagnostic process, as she helps to develop it from her personal history and, in turn, has to incorporate it "meaningfully" into her self-understanding [14,21]. So, patient communication requires a certain amount of information and knowledge, which are significantly restricted by the opacity of the algorithm, in order to be able to "speak and answer" meaningfully [11,14,21].

In this regard, both the interpretive task of the physician in the diagnostic process and the communicative aspect are integral to the core responsibilities of medical practice. Therefore, the question arises as to whether the physician can still be held morally responsible if the use of a MLDSS leads to structural and integral changes in the diagnostic process. In particular, interpretation and communication are impaired epistemically by, for example, the opaque nature of the algorithm, which is not entirely comprehensible and interpretable by its users [22]. Furthermore, the MLDSS lacks a reasoning structure that would allow the physician to meaningfully integrate the additional information from the MLDSS into the diagnosis, especially in the case of conflicting opinions. So, there is still the risk of a gap in responsibility or responsibility diffusion – at the epistemic level. It is important to examine here, on the one hand, the requirements that actors in the medical context – particularly physicians and patients – have to meet. On the other hand, the design of the algorithm as well as human-machine interaction hold considerable importance. In this regard, it appears valuable to consider the aspect of "forward-looking responsibility" ([23]) to proactively define the respective responsibilities of individual actors in advance and thereby prevent the diffusion of responsibility [13,24,25].

2.1.3. The control condition

Secondly, it should be noted that the condition of control is also in question due to the transformed diagnostic process. Even though the physician represents the final link in the diagnostic process and is responsible for integrating the MLDSS into the diagnosis, biases or incorrect decisions relating to the development process of the MLDSS, e. g. "during the data collection and acquisition process" [5], can be inherent to the algorithm and consequently falsely affect the diagnosis. A medical professional who has acted in accordance with her duty of care and has relied on the MLDSS – e. g., due to her experience or on the recommendation of a regulatory authority – is not responsible for the data fed into the algorithm or its development. Instead, with the incorporation of an algorithm into the diagnostic process, additional actors come into focus. This is referred to as the "problem of many hands and many things" [19]. Those who have been involved in development and distribution of MLDSS now might play a role in the question of moral responsibility [5]. The use of MLDSS in medicine thus leads to a disruptive change in the existing structure of responsibility, insofar as there is now a risk of incorrectly attributing responsibility, failing to fulfill responsibilities, or "passing the buck" to other actors. The attribution of responsibility becomes diffuse and, in turn, carries the jeopardy of a responsibility gap [5]. It is therefore clear that "responsibility as a multidimensional and relational concept" [5] needs to be understood and differentiated within the context of medical decision-making. To meaningfully identify responsibilities, the concept of "meaningful human control" appears promising. It attempts to identify all relevant actors within the medical diagnostic context and analyzes their duties, roles, moral reasons, and responsibilities, thus re-establishing attributions of responsibility [8,26].

In summary, the employment of MLDSS in medical decision-making contexts raises a plethora of challenging ethical questions. It seems useful to focus on the epistemic and control condition in AI-supported decision-making because moral responsibility presupposes control and epistemic cognition. A focus of the epistemic condition is on impairments of a physician's opportunities to interpret information and communicate it to a patient due to an MLDSS's opaqueness. While methods of XAI work intensively to address this challenge, one might argue that the explanatory discourse of humans can only be imperfectly approximated in hybrid constellations. The control condition focuses on "meaningful human control" in light of the "problem of many hands and things." It thus deals with problems of responsibility diffusion between physicians, developers and other actors in a technologically permeated medical decision-making context.

2.2. Behavioral science perspective

Utilizing MLDSS in medical diagnostic processes constitutes a complex decision-making situation where physicians are the decision-makers and the technical system is the advisor. The optimal way to facilitate the interaction between human medical experts and MLDSS is non-trivial and varies depending on the specific application. In the following, we will first look at some determinants of trust in human advice identified in the literature and then consider hybrid advice-giving situations.

2.2.1. Determinants of trust in human advice

The determinants of advisees' trust in advice from other humans can be broadly categorized into factors that relate to characteristics of (1.) the advisee him- or herself, (2.) the advisor and (3.) the decision-making context.

Advisees are keen to preserve a favorable (self-)image and are consequently less likely to seek advice if they fear that this advice-seeking makes them appear incompetent [27]. Moreover, there is a strong negative correlation between people's level of confidence in their decision-making and their inclination to seek advice in the first place [28] or to give weight to it once they receive it [28–30]. Consequently, if advisees' confidence is fostered by more experience through longer tenure or by perceptions of high power within their organization, they tend to seek less advice [31]. A systematic investigation of advice seeking reveals that subjective confidence predicts decision-makers' propensity to seek and use advice and that, conversely, their advice-seeking behavior is a reliable proxy for subjective confidence [32]. The same seems to be true for the advisor, as perceived self-competence also tends to increase people's desire to influence others by providing advice [33].

Further evidence on determinants of a decision-makers inclination to seek advice concerns the characteristics of the advisor. Advisees prefer

those advisors to whom they ascribe expertise and those who resemble themselves more closely [34]. For example, advisors within an organization are often preferred over external advisors. It was found that in a task in which decision-makers were advised by three peers, a single peer who confirmed a decision-maker's own initial judgment clearly decreased the influence of the other more distant peers [35]. Furthermore, advisees are more likely to seek advice from advisors whom they assume have positive feelings toward them [36]. In a recent meta-analysis of 129 independent datasets on advice-taking, the only unique predictor of the weight of advice was information about the advisor that influenced decision-makers' perception of advice quality [37].

A third determinant relates to the characteristics of the context in which the advice is being made, it has been demonstrated that decision-makers tend to seek more advice if the situation involves more uncertainty [38–40]. They are also more likely to solicit additional advice in situations in which the advice that they received was more dissimilar to their initial opinion [41]. The cited studies often rely on stylized estimation tasks with a convenience sample of undergraduate students in which these are shown pictures of glasses filled with coins and have to guess their number [30], estimate tuition fees for different colleges [29] or caloric content of various foods per serving [41]. Studies with professionals mostly focus on business contexts, studying advice-seeking on strategic issues or cost-cutting measures by CEOs [34,39], entrepreneurial teams in software ventures [38] or alumni of top-ranked business schools [27].

2.2.2. The phenomenon of egocentric discounting

Generally, it has been established that advisees exhibit "egocentric discounting" in quantitative estimation tasks where advice is present. This bias implies that decision-makers place more weight on their own estimates than on their advisor's [42,43] and has by now been consolidated in numerous studies of estimation tasks which have shown that decision-makers adjust, on average, about one third of the distance from their initial estimate to the advisor's [28,44–46]. In a more recent review, the authors conclude that the social information waste caused by egocentric discounting cannot be explained away with a single-cause account which stresses the importance of studying it from the perspectives of multiple research traditions [47]. One proposed explanation resonates with the first determinant of trust mentioned above. Advisees privy their own thoughts but not those of others which means that they know the reasons for which they hold their own initial opinions, while the reasons for which their advisors hold theirs remain dubious. According to several authors, it is this asymmetry that may lead to the decision-maker's egocentric discounting of other people's advice ([41]; [48]; [45,49]). In line with this explanation is the fact that decision-makers who had not previously generated an initial estimate themselves, i.e., who had not formed their own rationale, tended to make quantitative estimates that were closer to those provided by their advisors [50,51]. This also suggests that egocentric discounting is smaller if the advisor is perceived to be more similar to oneself as suggested by the second determinant introduced above.

The third determinant, being the context of decision-making, may be of special interest, because it seems most amendable by intentional design. In this respect, it has been found that egocentric discounting seems somewhat less pronounced in task environments where choices are binary instead of continuous, such as in a card game task, in which participants could "stay" with a card of known value or "switch" to a card of unknown value [52]. Participants were sensitive to the advice of whether to stay or switch and use the advice strategically depending on the information that they knew the advisor had about the value of the card unknown to them. Here, the authors conclude that the participants engage successfully in mental-state reasoning by "putting two heads together" and integrating their own knowledge and the advisor's complementary knowledge to improve their decision. For instance, they ignored the advice if they knew that the advisor could not see any cards and followed it if the advisor could see both cards. This may suggest that the level of uncertainty inherent in decision-making is another important driver of a decision-maker's reliance in ML-generated advice. Having more information, for instance, about the data on which the ML-generated advice is based could increase advisees' inclination to adopt it.

2.2.3. Implications for trust in machine advice

The literature on advice-based decision making where the advisor is a human, taken as a whole, thus indicates that decision makers tend to underweight rather than overweight advice. It is important to investigate whether this general finding carries over to constellations where the advisor is a machine instead of a human. The literature on these hybrid constellations is growing but still scarcer than the one referring to interhuman constellations. Which difference might one expect prima facie? An important factor that is likely to determine how ML-generated advice is received and utilized is the emotional dimension of the advice. This will likely depend on the affective attitude that the decision-maker has toward the MLDSS as opposed to a human expert. Given that the advisee's perception of the advisor has been found to be an important determinant of trust it seems important to understand how they perceive MLDSS. This may of course be influenced directly by the human-AI interface or more indirectly by other characteristics of the interaction. For instance, the way the advice is communicated to the decision-maker might matter, as it is likely to induce feelings of dominance over, subordination to or partnership with the MLDSS in the advisee. Another factor that is likely to play a role is the dissimilarity between the ML-generated advice and the perspective of the decision-maker. It has been found that pooling individual information to achieve collective intelligence is particularly successful if it capitalizes on individual heterogeneity [53].

Given the explanation of egocentric discounting cited above that decision-makers have difficulties to understand an advisor's thoughts, however, one might expect that a machine's "rationale" for giving a specific advice is perceived as even more opaque and thus dubious. In this sense, the MLDSS's advice may be perceived as more dissimilar from the decision-maker's perspective than the advice of a fellow human advisor. This may also be due to the fact that humans have a harder time emphasizing with artificial advisors and that building a theory of mind of entities whose inner workings seem alien to them comes less natural. Similarly, the observations that decision-makers are more likely to accept advice from people who resemble them or whom they ascribe more positive feelings towards them might suggest that humans are more likely to discount the advice of unfamiliar artificial agents. Overall, these considerations may suggest an even stronger discounting of the advice of artificial as compared to human advisors.

Contrary to this, however, decision-makers may also be aware that artificial agents are tasked to attain a certain and clearly defined objective, while people's motives often may be more ambiguous. Relatedly, it might be the case that artificial agents that are designed to be "sympathetic" to decision-makers are perceived as being less contemptuous or dominant than human experts and appear thus more likable. These opposed intuitions seem to suggest a higher reliance on artificial agents' as opposed to human agents' advice. The tendency to trust or distrust in AI-generated compared to human advice is likely to depend substantially on the experience that individual decision-makers have made with MLDSS and the degree to which they are used to working with these entities. Chong et al. (2022) have shown that decision-makers' individual experience with AI-assisted decision-making has a profound influence on their reliance on the advice through the effect that the collaboration has on their self-confidence. From a meta-analysis on trust in AI, Kaplan et al. [54] conclude that the interaction between the many different antecedents of trust in AI has not been empirically investigated. Promoting this kind of research is important, particularly in ethically charged contexts like medicine.

2.2.4. Algorithm aversion vs. algorithm appreciation

Several research findings point to the phenomenon of algorithm aversion, implyingthat humans underuse or mistrust artificial agents' compared to human agents' advice. These findings show that humans do not distrust artificial agents in principle and from the outset but depending on their expectations about their own and the algorithm's performance, experience with the artificial advisor and the context of the task. Informational mechanisms that are linked to the perceived competence of the advisor seem to play a stronger role for human-robot interactions than for interhuman interactions which are highly co-determined by social normative mechanisms [55].

The perception of the MLDSS as a superior intelligence seems important as expectations are easily disappointed. If decision-makers see a forecasting algorithm err, their algorithm aversion increases, even when they see it outperform a human forecaster [56]. Concerning the decision-making context, it has been found that algorithm aversion is more pronounced for tasks that seem subjective in nature [57]. With increased difficulty in intellective tasks with clearly correct answers, subjects in three experiments relied more on algorithmic advice with increasing difficulty [58]. It is suggested that algorithm aversion is driven by biased evaluations of human capabilities and tends to be stronger in domains where people's identity is threatened [59]. In line with this, trust in algorithms can be brought about by the opportunity to (even slightly) modify an imperfect forecasting algorithm, i.e., by the decision-makers' feeling of being more in control [60]. Finally, Dietvorst and Bharti [61] found that trust in algorithms will depend on the uncertainty inherent in the decision domain and that people have diminishing sensitivity to forecasting error which lets them rely on riskier human judgment. They suspect that people may be unwilling to use even the best possible algorithms in inherently uncertain domains like medical decision-making.

Algorithm appreciation describes the opposite phenomenon that decision-makers adhere more to advice from algorithms than from persons. In a sequence of six experiments, people showed algorithm appreciation when making numeric estimates about visual stimuli and forecasts about song popularity and romantic attraction [62]. In contrast to results cited above, it has also been reported that an increased trust in algorithmic advisors does not erode when decision-makers are informed of the algorithm's prediction errors [63] or when they knowingly interact with dubious algorithms ([64]). Research on automation bias focuses on the problems that are caused by users' failure to recognize new errors that are introduced by automated advice [65,66].

Hou and Jung [67] have attempted to reconcile the contradicting phenomena of algorithm appreciation and algorithm aversion in AI-supported decision-making. Their research focuses on the importance of how humans and algorithms are framed in the different studies. They argue that different framings produce the inconsistent results of algorithm aversion and algorithm appreciation observed in previous studies. According to the authors, framing refers to the relative description of the human and of the artificial advisor: what kind of people and what kind of algorithm are we comparing? The importance of framing emphasizes the need to investigate to which degree advice by artificial agents in the medical domain is sought and weighted and how this compares to advice by humans. Among several personal factors like a physician's level of experience and general confidence, this will probably also depend on features that are amendable to the design of the interaction, i.e., factors that concern the perception of the advisor or the decision situation. These factors could, based on previous results, consist in the information that the algorithm provides regarding the criteria on which it based its advice or in emphasizing the uncertainty of the situation by communicating its own imperfect confidence in the accuracy of the advice. Even the sequence of the interaction can have an impact on decision-making [68].

To sum up, the behavioral perspective focuses so far mainly on an advisees' trust in the advice as determined by participants' behavioral tendency to follow it in their choices. Especially in behavioral experiments, these choices are often consequential as they imply lower monetary payoffs if participants perform worse. Interhuman advice giving has been studied extensively and determinants that relate to characteristics of the advisee him- or herself, the perception of the advisor and the decision-making context were identified. Hybrid interactions of human and AI are increasingly studied but behavioral evidence is still much scarcer. There is mixed empirical evidence on algorithm aversion and algorithm aversion with few attempts to reconcile these phenomena. Understanding determinants of physicians' trust in MLDSS will therefore be indispensable.

2.3. Technical perspective

Machine learning algorithms, especially those using complex (deep) networks, are being perceived as "black boxes" [69], and their use in safety critical environments such as autonomous driving or in the medical context is met with criticism. This is likely rooted in the inability of an observer to understand such complex systems and not in a lack of transparency regarding the model and its individual components and values. To overcome these limitations, van Lent et al. [70] introduced the concept of XAI already in the early days of AI, which should provide the user with an "easily understood chain of reasoning". XAI has been a very dynamic field of research centered around developing technologies to make machine learning models more interpretable. Methods in XAI typically aim to either explain the decision for a single data instance (local explanations) or to provide a global explanation of the model's decision-making process across all data instances. In the context of the medical diagnostic process, local explanations are of highest relevance, as they help clinicians understand the specific reasons behind a model's prediction for an individual case.

2.3.1. Explainability of deep machine-learning models

One approach to achieving explainability is through the employment of inherently interpretable models, such as linear regression or compact decision trees. This, however, has been criticized for resulting in restrictions in machine learning model design, which can lead to lower performance of those models [71].

Initially, researchers have focused on post-hoc, model-agnostic interpretability methods. These methods preserve the structure and performance of machine learning models by analyzing them through their input-output relationships. One example of this is the communication of prototypical input space representatives of a decision (e.g., exemplary images) together with 'criticisms' (non-prototypical examples of the same class) [72]. Contrastive methods offer another layer of insight by explaining why a particular decision was made in contrast to other potential decisions [73]. For example, these methods aim to clarify why a medical model diagnosed one disease instead of another similar disease, enhancing our understanding of the model's decision-making process.

Another similar method is the use of counterfactuals, which provide examples that demonstrate how slight modifications in the input could change the model's decision [74]. For instance, a chest x-ray image with a certain model diagnosis could be contrasted to how the same patient's image could look like for another diagnosis [75]. All of these methods, which essentially show and contrast exemplary behaviors of the model, follow the idea of familiarizing the user with the model and aligning expectations with the actual behavior.

More recently, the design of deep model architectures that are explainable by design has received growing interest. The core idea is to incorporate the pattern matching capabilities of modern deep neural networks (and hence to achieve highly performant models), with inherently explainable methods. For instance, the PIP-Net approach by Nauta et al. [76], 2024) uses convolutional feature extraction in the early layers of the network, providing high non-linear pattern recognition capacity, followed by spatial aggregation and linear combination, which are inherently explainable methods. This allows for a "scoring

card" interpretation, facilitating insight into how predictions are made. Unfortunately, this structure also comes with a trade-off: while increasing explainability, the model's ability to capture and recombine more complex spatial patterns is constrained, potentially limiting its overall performance. The very property of being inherently explainable (in the last part of the model) limits the model to discover and map more complex spatial relationships in the input image. Furthermore, while the first part of the model, which performs a more complex pattern matching against intuitive prototypes, is in fact a powerful pattern recognizer, it has some of the same black box properties that have been criticized and may be vulnerable to challenges such as adversarial attacks and unpredictable performance on out-of-distribution cases.

2.3.2. Limits of explainability

The remainder of the section focuses on the discussion of the limits of explainability that has been emphasized by some authors. The idea that explanation is limited is that the methods discussed aid in understanding the model in a local context of the input space, but fall short of achieving comprehensive or holistic interpretability as defined by Miller et al. (2019), i.e., that a human can understand the cause of a decision, which is crucial for applications like medical MLDSS, where understanding the entire decision-making process, in particular also for edge cases, is important. This limitation is especially relevant for increasingly large models (e.g., large language models and other recent foundation models) and becomes particularly evident in the case of images and text as input, which is the most prevalent scenario in medical MLDSS. Exemplary images alone communicate only distinct modes of the data, thereby falling short of providing a comprehensive understanding of the model's decision-making process.

Furthermore, these examples are typically drawn from a certain data distribution, which may not be representative of the target distribution in actual use cases [77]. For example, models trained on data from one demographic might perform unpredictably when applied to a different demographic, due to covariant data distribution shifts [78]. These shifts can result in unexpected and potentially harmful model predictions. Even further, they might be contradictory to the expectations of the user that were established during familiarization, e.g., by using counterfactuals, prototypes and criticisms.

This highlights a critical aspect in assessing the reliability of large models: the effectiveness of these models is inherently linked to the representativeness of the data used during training and user familiarization. Typically, training data lacks uniform representation across different conditions, such as different demographic groups or clinical scenarios, which can lead to inherent biases in model behavior. Similarly, these representational imbalances can introduce perceptual biases in the explainability of models. While explainability methods can certainly help to detect biases in models and/or datasets, they are also dependent on the data distribution. Explainability patterns, which are derived from training data, might not accurately reflect the actual use case scenarios, leading to misleading or incomplete explanations.

2.3.3. The consequences of non-explainability

Consequently, achieving explainability in model decisions is a complex and nuanced topic of research. The clinical utility of explainability methods is not yet fully understood, and it is still unclear whether the diagnostic process will be positively influenced by it ([79]; [80]). Explainable methods might also be persuasive and increase cognitive biases [81]. This may raise an important question: while explainability is very much desirable – are we perhaps focusing too strongly on interpretability at the expense of other crucial aspects? While the call for transparency or traceability of the reasoning of the algorithm is easy to comprehend, it could also be considered a case of an anthropomorphic fallacy [82]: Potentially based upon a perceived parallelism of a human medical expert and a medical decision support system, we assume the algorithm to *reason* and *decide,* ignoring that the very nature of today's (large) machine learning models does not include reason or a conscious decision.

Machine learning systems work by optimizing what can be described as a gigantic equation that incorporates lots of variables, typically with the target of maximizing the correlation to the desired output and reducing the number and magnitude of errors such a system makes. This is very different to human reasoning, which is ideally based on the logical combination of verified assumptions, but also subject to an often-underestimated number of known fallacies and biases [83–85]. The restriction of AI methods to models that are logically understandable, is also known to restrict the complexity and hence the capability, versatility, or efficiency of said models [71]. If the introduction of transparency into machine learning model design comes at the cost of reduced performance, this directly leads to a moral dilemma.

2.3.4. Not interpretable but safe and effective

In the realm of medical research, however, one may argue that withholding a particular treatment or medication solely due to an incomplete comprehension of its mechanism of action can be ethically problematic, particularly when the treatment's efficacy has been comprehensively substantiated, and its potential side effects and interactions have been fully identified. This seemsparticularly true for the function of well-known and widely used drugs, such as the painkiller paracetamol [86] or the anesthetic ketamine [87], which have been used for decades besides a poorly understood mechanism of action.

In laboratory testing of diagnostic samples (such as blood and other bodily liquids), it has well been accepted that the nature of a test result is purely statistical and, in practice, subject to many factors that can influence the test outcome that are rooted in the preparation of the samples [88] and hence hard to trace in retrospect. Each test method is known to be subject to a defined set of statistical properties (e.g., recall and specificity for the detection of a disease, standard error of a measurement) and to be used only given a defined set of circumstances (such as mode of application or even patient population) that the method has been validated for [89].

This could serve as a blueprint for ML in medical decision making: Given properly and independently validated ML models, where the boundary conditions of the use (i.e., the expected data distribution) are described in sufficient detail [90], the potential risks to safety are mitigated and the benefit through efficiency and performance outweighs the risks.

In summary, while explainability seems desirable for machine learning models and a lot of progress in the field of explainable AI in recent years allows for better explainability even of deep learning models, increased explainability may come at the cost of reduced accuracy. At the same time, the lack of independent validation for many AI models poses its own risks, potentially leading to generalization errors that reduce accuracy in clinical settings. To ensure reliable performance, robust validation of MLDSS using sufficiently large and, most importantly, representative datasets is essential and the precondition to their application being safe and effective. Ultimately, the question of whether explainability should take precedence over accuracy is not just a technical one. It is a question where the patient as the primary stakeholder should be involved as well.

3. The need for interdisciplinary exchange

The utilization of MLDSS in medical decision-making raises important scientific questions with a high relevance for society. In a way, the scientific questions that MLDSS pose at the ethical and behavioral level are even more fundamental than the technical ones. On the one hand, ethical issues arise regarding changes in the responsibility structures in healthcare that may be caused by the introduction of decision-support systems. Can we still reliably assign individual responsibility in case of a medical error or is the responsibility of physician, algorithm developer, regulator and the hospital management implementing the hybrid interaction diffused to such a degree that individual responsibility

dissipates? In case of the latter, may it be ethically appropriate to withhold a decision-support system from a patient even though its performance has been comprehensively validated or does performance trump transparency and clear attribution of responsibility? When does one outweigh the other?

On the other hand, issues in behavioral sciences arise at the level of the actual effects that the interaction between physicians and algorithms has on the diagnostic process and thus ultimately on the patient. An extensive literature on advice-based decision-making reveals many factors at play in human-human interactions. To what extent do these results transfer to human-AI interactions? How can physicians' trust in decision-support systems and reliance on the systems' advice be increased or decreased through specific design of the interaction, the human-machine interface or the way the system "justifies" its advice? How do we resolve the potential tension between explainability and accuracy of machine learning models and, in particular, who decides which trade-offs are acceptable in this regard?

3.1. Toward a normatively and empirically informed technical implementation

These ethical and behavioral questions ought to be addressed before any adequate comprehensive technical solution can be even conceptualized, let alone implemented. Conceptual issues regarding responsibility as posed by ethics will be essential to judge whether MLDSS should give specific advice at all or whether they should only help to structure the decision situation by, for instance, highlighting segments of a given medical image that the physician should turn her attention to. Moreover, ethicists and behavioral scientists need to join the debate in computer science about the importance and usefulness of explainability in machine learning models, especially when there are trade-offs between explainability and model performance. All of this would have direct implications for the design of the human-algorithm interaction through the information that the system's interface communicates to the diagnostician. Similarly, empirical phenomena regarding any influences that the advice has on the diagnostician will be crucial for a comprehensive assessment of the technology. Assume that it is found that a certain kind of representation of the system's advice makes it systematically more likely that the diagnostician will follow the advice. From an ethical view, it is important to decide whether intentionally exploiting this effect to foster adherence to the system's advice is legitimate or whether it undermines the physician's autonomy in an unacceptable way. Only if the respective manipulation is considered morally legitimate would the identified behavioral implications be used to actively shape the human-system interaction accordingly through the concrete technical design.

It therefore becomes clear that it is by no means sufficient to discuss each perspective's challenges in isolation but indispensable to take an interdisciplinary perspective on the challenges raised. The examples stated above illustrate how all the perspectives are interrelated and that scientists working in each field have to complement each other to enable human-centeredness in AI-supported medical decision-making. It takes behavioral scientists to explore the factors influencing the interaction, computer scientists and XAI researchers to improve the performance of machine-learning algorithms with due consideration of possible influencing factors, and it takes ethicists to work out any red lines in the application of these systems in medical practice. It is clear that different methodologies and terminology can lead to misunderstanding and skepticism between disciplines. But if the goal is ethically aligned, human-centered AI, then this endeavor necessarily transcends the expertise of any single discipline.

Although there is already important work on the impact of AI-based systems on users being done in XAI research (for an overview, see for instance, Ref. [91–93]), the focus of this research is usually not on ethically motivated questions. Focusing on the ethical implications of human-AI interaction will inevitably lead to the definition of different dependent variables and different experimental manipulations than the ones that behavioral scientists and computer scientists will typically address. Ethicists alone, however, will hardly be able to investigate the complex interplay between human behavior and technical artifacts without profound expertise from behavioral and computer science. In the following, we will outline two illustrations of how the three perspectives discussed in this article interlock with each other and how a fruitful interdisciplinary exchange could look like.

3.2. Example 1: The Impact of Epistemic vs. Ethical Advice

Most tasks that were cited in the behavioral perspective of this article were epistemic tasks in which advisees received advice on a knowledge problem. As discussed, there are important behaviorally oriented studies on the effect of XAI that refer to the question of whether certain means of explainability make advisees trust more in the advice. These studies systematically vary explanatory features of the MLDSS to study their influence on human performance and satisfaction of the explanation. There are, however, ethically highly relevant features of the decision situation that remain typically out of scope in this research. One crucial aspect relates to the diffusion of responsibility between the user and an MLDSS.

Several ethicists assume that responsibility can categorically not be attributed to AI-based systems [94]. Empirical research, however, suggests that people factually tend to ascribe responsibility to MLDSS in the context of medical image diagnosis and that this is especially the case if they attribute consciousness to the system [95]. The psychological reality of people shifting responsibility to MLDSS may, however, have implications for a physician's willingness to take risk (see also [17]). Behaviorally, the desire to shift responsibility to an MLDSS will be reflected in one's willingness to rely on the system's advice. To study the empirical relevance of this ethically motivated question a behavioral research design that is embedded in a concretely designed context of human-AI interaction is needed. This is an example of a challenge that can only be addressed by the three perspectives working together and thereby tackling the problem holistically.

The ethically informed concept of responsibility diffusion is first to be operationalized in a concrete behavioral design. This can be done, for instance, by varying the perception of a given task as either primarily epistemic or primarily ethical. One possibility would be to frame the decision-making scenario by making its epistemic or ethical nature more salient. Another means for doing so is the use of either self-regarding or other-regarding incentives for the correct classification of images in laboratory studies. Self-regarding incentives mean that an advisee receives a financial reward for his or her classification accuracy. Other-regarding incentives mean that another (passive) participant receives a financial reward for an advisee's accuracy. This intends to capture consequential spillovers of one's decision on other people as is the case when physicians make decisions affecting patients. It is to be expected that tasks with self-regarding incentives are relatively more likely to be primarily interpreted as epistemic tasks by the advisee because performance failures "only" lead to one's own financial loss. Vice versa, it is to be expected that tasks with other-regarding incentives are relatively more likely to be primarily interpreted as ethical tasks by the advisee because failures lead to another person's financial harm. In this experimental setup, the hypothesis could then be that advisees' reliance in MLDSS is stronger in treatments with other-regarding incentives than with self-regarding incentives because in the former decision-makers have a more intensive desire to shift responsibility to and hide behind the MLDSS.

It is important to note that the cooperation between advisee and MLDSS does not happen in an abstract setting but is to be studied in a concrete context of decision-making. The manipulation of a task as being perceived as primarily epistemic or primarily ethical is therefore likely to interact with the mode of XAI used to justify the advice. Studying the emerging interaction effects is crucial for an empirically informed

human-centered design approach. In this context, it seems particularly worthwhile to systematically study how the use of methods from XAI in advice-giving influences a participant's perception of owning one's decision after receiving advice or not. One may hypothesize that receiving an explanation for a given advice makes it more likely that an advisee feels that the advice has truly persuaded him or her and that it is therefore more considered to be integrated in the advisee's decision. In case of a wrong diagnosis, an advisee may then feel more responsible than in a case where the advice was absorbed in a more unreflected way and where he or she feels rather manipulated.

3.3. Example 2: The Perception of Manipulative vs. Persuasive AI

Manipulation is a concept that is at the core of philosophical ethics [96]. It seems valuable to operationalize this abstract discussion to make it fruitful for a holistic empirical investigation of what a manipulation-free human-AI interaction could look like. Operationalization will have to be done jointly by behaviorists and computer scientists who will need ethicists' guidance to navigate through the complex ethical debate on manipulation and boil it down to its key ideas. Ethicists, on the other hand, will need their empirically oriented colleagues to help them translate their conceptual notions in specific research design choices. The concepts of manipulative vs. persuasive AI [97,98] may therefore serve as another useful illustration of interdisciplinary collaboration. From an ethical perspective, it might be considered important that physicians are not subconsciously nudged into following the advice of an MLDSS because this would undermine their autonomy as decision-makers. If nudging occurs, they are only nominally kept in the loop of decision-making, but their role is factually reduced to the role of vicarious agents of the technical system (see, for instance, Ref. [99]). Generating faithful explanations that are provided to substantiate an MLDSS's advice to convince the advisee to follow it seems ethically more desirable than generating explanations that are behaviorally most successful in inducing follower behavior. The next paragraphs may help to illustrate how a holistic and interdisciplinary collaboration between ethicists, behaviorists and computer scientists could look like in this context.

Ethicists disagree to a substantial degree about the identification question of what makes advice manipulative. For instance, some philosophers argue that manipulation should be characterized by the intentions of the manipulator. The Trickery Account of manipulation argues that it is characterized by a deliberate attempt to induce faulty beliefs in a person to make him or her act in the way desired by the manipulator [100]. Klenk [101], however, argues for an Indifference Account where manipulation is not necessarily caused by deceptive intentions but by a manipulators' indifference to the means with which he or she achieves her goal to influence the decision-maker. This latter approach seems better suited for the concept of "manipulative AI" as the MLDSS will by definition be lacking intentions to deceive and thus be indifferent to the means it uses. It should be noted that this discussion is not only of academic interest as policy makers are increasingly keen of banning AI systems that manipulate people's decisions or exploit their vulnerabilities (EU AI Act, Article 5).

It seems important to explicitly consider the intuitions of the potentially manipulated decision-makers and those affected by their choices to understand whether an explanation is interpreted as faithful or manipulative. It is therefore crucial to address this question in behaviorally designed experiments. To do so, it seems necessary to explicitly contrast participants' perception of an advice when receiving it from a human that provides certain explanations with an advice that is provided by MLDSS that provide certain explanations. This comparison may help to disentangle between the above cited Trickery and Indifference Account and their relevance for the actual perception of human-AI interaction and their behavioral implications. It is unlikely that behaviorists and XAI researchers will be able to design adequate experimental setups to capture this distinction without closely collaborating with ethicists.

Analogously, will XAI researchers naturally focus on different kinds of questions that are more closely related to the questions that are at the core of their domain than ethicists or social scientists. Nauta et al. [102], for instance, focus on the alignment of explainable image classifiers with medical classification standards and the possibility of human decision-makers to manually correct the system's reasoning. It would be fascinating to expand their research by studying whether the possibility to correct reasoning increases an advisees' perception of owning a decision. Kulesza et al. [18], for instance, focus on Explanatory Debugging Cycles of Explanation and find that users using these cycles understood the learning system better and expressed a preference for them. Their research design, however, does not test for ethically relevant characteristics like a potential divergence in people's keenness for explanatory cycles in ethical tasks as opposed to epistemic ones. Furthermore, the explicit elicitation of the impact of the explanation on subjective feelings of being manipulated and whether this perception differs between the recipients of the advice and those who observe the influence from the outside is not considered. Establishing such a gap, however, may be very relevant for a policy debate that often relies on the identification criteria of outside observers like politicians or lawyers. It is unlikely that such a question would be addressed without a profound knowledge of the philosophical manipulation debate that behaviorists and XAI specialists are unlikely to have.

Similarly, previous research that sought to address real-world users instead of outsider observers has relied on qualitative interviews with UX specialists to create user-centered AI [103]. This is doubtlessly important. It seems, however, important to complement this research by quantitative studies that systematically investigate the intuitions of affected users, as designers might wrongly project their own intuitions on others. Here, we take up a call from Miller [104] who already argued that most work in XAI tends to take up researchers' intuitions of what constitutes an explanation instead of relying on expertise from other disciplines how people define and evaluate explanations. This underlines that the implementation and implication of an ethical concept like "persuasion instead of manipulation" has to be tested using behavioral methods with real users in the context of concrete technological solutions. This may necessitate manipulations that compare perceptions of human advice versus AI-based advice, experiencing and observing influence or performing for oneself or others.

These are manipulations from social science research that are typically and understandably out of the scope of more technically oriented XAI researchers.

The use of XAI has profound ethical implications that need to be studied holistically. Specific manipulations and experimental manipulations that happen at the behavioral level within concrete interaction designs will naturally not arise from behavioral or technical research questions that are often based on the intention to increase performance. This is natural because different perspectives tend to focus on different questions. The diversity of these different foci leads to a more holistic consideration of AI-based decision-making in medicine. This requires the working together of experts from all three disciplines on joint research designs.

4. Conclusion

In a seminal paper, Rahwan et al. [105] draw attention to the need for cross-disciplinary scientific collaboration to examine how intelligent machines affect humans and vice versa. Medicine is one area where, due to numerous advances in automatic processing and interpretation of data, AI-based decision support systems are already making their way into everyday practice. The use of AI-based systems will continue to increase significantly in the near future in this area. Given its societal relevance, medicine therefore represents a use case where the need for cross-disciplinary research is particularly urgent.

In this paper, we have focused on three perspectives that seem to us

highly relevant for achieving human-centered medical AI. We acknowledge, however, that the selection of these perspectives was likely coined by a professional bias of the authors who happen to be ethicists, behavioral scientists, and computer scientists. One might rightly miss the perspectives of scholars from, say, law, sociology, the political sciences and many other fields. One may hope that the successful collaboration of different fields will make it ever more likely that further perspectives will be integrated over time.

CRediT authorship contribution statement

Jonas Ammeling: Writing – review & editing, Writing – original draft, Conceptualization. **Marc Aubreville:** Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization. **Alexis Fritz:** Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization. **Angelika Kießig:** Writing – review & editing, Writing – original draft, Conceptualization. **Sebastian Krügel:** Writing – review & editing, Writing – original draft, Conceptualization. **Matthias Uhl:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization.

Conflict of interest statement

The authors declare that there is no conflict of interest.

Acknowledgments

The authors are grateful to the Bavarian Research Institute for Digital Transformation (bidt) for funding our research.

Data availability

No data was used for the research described in the article.

References

[1] Ami B. Bhatt, Jennifer Bae, Collaborative intelligence to catalyze the digital transformation of healthcare, NPJ Digital Medicine 6 (2023) 177.
[2] G. Campanella, et al., Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat. Med. 25 (2019) 1301–1309.
[3] W. Jorritsma, F. Cnossen, P.M. van Ooijen, Improving the radiologist–CAD interaction: designing for appropriate trust, Clin. Radiol. 70 (2) (2015) 115–122.
[4] N. Donner-Banzhoff, Die Ärztliche Diagnose. Erfahrung – Evidenz – Ritual, Bern, 2022.
[5] H. Bleher, M. Braun, Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems, AI and Ethics 2 (2022) 747–761.
[6] M. Braun, et al., Primer on an ethics of AI-based decision support systems in the clinic, J. Med. Ethics 47 (12) (2020) e3.
[7] A. Matthias, The responsibility gap: ascribing responsibility for the actions of learning automata, Ethics Inf. Technol. 6 (3) (2004) 175–183.
[8] F. Santoni de Sio, G. Mecacci, Four responsibility gaps with artificial intelligence: why they matter and how to address them, Philosophy & Technology 34 (4) (2021) 1057–1084.
[9] R. Crisp (Ed.), Aristotle: Nicomachean Ethics, Cambridge University Press, 2014.
[10] M. Coeckelbergh, Narrative Responsibility and Artificial Intelligence. How AI Challenges Human Responsibility and Sense-Making, AI & Society, 2021.
[11] M. Verdicchio, A. Perin, When doctors and AI interact: on human responsibility for artificial risks, Philosophy & Technology 35 (1) (2022).
[12] A. Fritz, W. Brandt, H. Gimpel, S. Bayer, Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI), De Ethica. A Journal of Philosophical, Theological and Applied Ethics 6 (1) (2020) 3–22.
[13] T. Grote, Machine learning in healthcare and the methodological priority of epistemology over ethics, Inquiry 2 (2024) 1–30.
[14] F. Funer, The deception of certainty: how non-interpretable machine learning outcomes challenge the epistemic authority of physicians. A deliberative-relational approach, Med. Healthc. Philos. 25 (2022) 167–178.
[15] M. Solomon, Making Medical Knowledge, 2015. Oxford.
[16] A.J. London, Artificial intelligence and black-box medical decisions: accuracy versus explainability, Hastings Cent. Rep. 49 (1) (2019) 15–21.
[17] H. Kempt, S.K. Nagel, Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts, J. Med. Ethics 48 (4) (2022) 222–229.
[18] T. Kulesza, M. Burnett, W.K. Wong, S. Stumpf, Principles of explanatory debugging to personalize interactive machine learning, in: Proceedings of the 20th International Conference on Intelligent User Interfaces, 2015, March, pp. 126–137.
[19] M. Coeckelbergh, Artificial intelligence, responsibility attribution, and a relational justification of explainability, Sci. Eng. Ethics 26 (2020) (2020) 2051–2068, 4.
[20] T. Gundersen, K. Bærøe, The future ethics of artificial intelligence in medicine: making sense of collaborative models, Sci. Eng. Ethics 28 (2) (2022).
[21] J.C. Bjerring, J. Busch, Artificial intelligence and patient-centered decision-making, Philosophy & Technology 34 (2) (2021) 349–371.
[22] K. Baum, et al., From responsibility to reason-giving explainable artificial intelligence, Philosophy & Technology 35 (1) (2022).
[23] I. Poel van de, M. Sand, Varieties of responsibility: two problems of responsible innovation, Synthese 198 (2021) 4769–4787.
[24] M. Sand, J.M. Durán, K.R. Jongsma, Responsibility beyond design: physicians' requirements for ethical medical AI, Bioethics 36 (2) (2021) 1–8.
[25] S. Nyholm, Responsibility gaps, value alignment, and meaningful human control over artificial intelligence, in: A. Placani, S. Broadhead (Eds.), *Risk And Responsibility In Context,* Routledge Studies in Ethics and Moral Theory, 2023, pp. 191–213. New York.
[26] L. Cavalcante Siebert, et al., Meaningful human control: actionable properties for AI system development, AI Ethics 3 (1) (2023) 241–255.
[27] I. Cojuharenco, N. Karelaia, When leaders ask questions: can humility premiums buffer the effects of competence penalties? Organ. Behav. Hum. Decis. Process. 156 (2020) 113–134.
[28] J.B. Soll, R.P. Larrick, Strategies for revising judgment: how (and how well) people use others' opinions, J. Exp. Psychol. Learn. Mem. Cognit. 35 (3) (2009) 780–805.
[29] K.E. See, E.W. Morrison, N.B. Rothman, J.B. Soll, The detrimental effects of power on confidence, advice taking, and accuracy, Organ. Behav. Hum. Decis. Process. 116 (2) (2011) 272–285.
[30] X. Wang, X. Du, Why does advice discounting occur? The combined roles of confidence and trust, Front. Psychol. 9 (2018). Article 2381.
[31] A. Vestal, R. Guidice, The determinants and performance consequences of CEO strategic advice seeking, J. Gen. Manag. 44 (4) (2019) 232–242.
[32] N. Pescetelli, A.K. Hauperich, N. Yeung, Confidence, advice seeking and changes of mind in decision making, Cognition 215 (2021) 104810.
[33] U. Hertz, E. Tyropoulou, C. Traberg, B. Bahrami, Self-competence increases the willingness to pay for social influence, Sci. Rep. 10 (1) (2020) 17813.
[34] M.L. McDonald, P. Khanna, J.D. Westphal, Getting them to think outside the circle: corporate governance, CEOs' external advice networks, and firm performance, Acad. Manag. J. 51 (3) (2008) 453–475.
[35] L. Molleman, A.N. Tump, A. Gradassi, S. Herzog, B. Jayles, R.H. Kurvers, W. van den Bos, Strategies for integrating disparate social information, Proceedings of the Royal Society B 287 (1939) (2020) 20202413.
[36] J.D. Hur, R.L. Ruttan, C.T. Shea, The unexpected power of positivity: predictions versus decisions about advisor selection, J. Exp. Psychol. Gen. 149 (10) (2020) 1969–1986.
[37] P.E. Bailey, T. Leon, N.C. Ebner, A.A. Moustafa, G. Weidemann, A meta-analysis of the weight of advice in decision-making, Curr. Psychol. 42 (28) (2023) 24516–24541.
[38] A.S. Alexiev, J.J.P. Jansen, F.A.J. Van den Bosch, H.W. Volberda, Industry differences in strategic decision making of Dutch top management teams, in: K. J. McCarthy, M. Fiolet, W. Dolfsma (Eds.), Nature of the New Firm: beyond the Boundaries of Organizations and Institutions, 2011, pp. 58–75.
[39] M.L. Heyden, S. Van Doorn, M. Reimer, F.A. Van Den Bosch, H.W. Volberda, Perceived environmental dynamism, relative competitive performance, and top management team heterogeneity: examining correlates of upper echelons' advice-seeking, Organ. Stud. 34 (9) (2013) 1327–1356.
[40] B. Vissa, A.S. Chacar, Leveraging ties: the contingent value of entrepreneurial teams' external advice networks on Indian software venture performance, Strat. Manag. J. 30 (11) (2009) 1179–1191.
[41] M. Hütter, F. Ache, Seeking advice: a sampling approach to advice taking, Judgment and Decision Making 11 (4) (2016) 401–415.
[42] S. Bonaccio, R.S. Dalal, Advice taking and decision-making: an integrative literature review, and implications for the organizational sciences, Organ. Behav. Hum. Decis. Process. 101 (2) (2006) 127–151.
[43] I. Yaniv, Receiving other people's advice: influence and benefit, Organ. Behav. Hum. Decis. Process. 93 (1) (2004) 1–13.
[44] P. Ecken, R. Pibernik, Hit or miss: what leads experts to take advice for long-term judgments? Manag. Sci. 62 (7) (2016) 2002–2021.
[45] J.A. Minson, V. Liberman, L. Ross, Two to tango: effects of collaboration and disagreement on dyadic judgment, Pers. Soc. Psychol. Bull. 37 (10) (2011) 1325–1338.
[46] J.B. Soll, A.E. Mannes, Judgmental aggregation strategies depend on whether the self is involved, Int. J. Forecast. 27 (1) (2011) 81–102.
[47] O. Morin, P.O. Jacquet, K. Vaesen, A. Acerbi, Social information use and social information waste, Philosophical Transactions of the Royal Society B 376 (1828) (2021) 20200052.
[48] M. Milyavsky, A.W. Kruglanski, M. Chernikova, N. Schori-Eyal, Evidence for arrogance: on the relative importance of expertise, outcome, and manner, PLoS One 12 (7) (2017). Article e0180420.
[49] I. Yaniv, S. Choshen-Hillel, M. Milyavsky, Spurious consensus and opinion revision: why might people be more confident in their less accurate judgments? J. Exp. Psychol. Learn. Mem. Cognit. 35 (2) (2009) 558–563.

[50] D.J. Koehler, T.A. Beauregard, Illusion of confirmation from exposure to another's hypothesis, J. Behav. Decis. Making 19 (1) (2006) 61–78.
[51] I. Yaniv, S. Choshen-Hillel, Exploiting the wisdom of others to make better decisions: suspending judgment reduces egocentrism and increases accuracy, J. Behav. Decis. Making 25 (5) (2012) 427–434.
[52] N. Vélez, H. Gweon, Integrating incomplete knowledge with imperfect advice, Topics in Cognitive Science 11 (2) (2019) 299–315.
[53] T. Kameda, W. Toyokawa, R.S. Tindale, Information aggregation and collective intelligence beyond the wisdom of crowds, Nature Reviews Psychology 1 (6) (2022) 345–357.
[54] A.D. Kaplan, T.T. Kessler, J.C. Brill, P.A. Hancock, Trust in artificial intelligence: meta-analytic findings, Hum. Factors 65 (2) (2023) 337–359.
[55] J. Zonca, A. Folsø, A. Sciutti, Social influence under uncertainty in interaction with peers, robots and computers, International Journal of Social Robotics 15 (2) (2023) 249–268.
[56] B.J. Dietvorst, J.P. Simmons, C. Massey, Algorithm aversion: people erroneously avoid algorithms after seeing them err, J. Exp. Psychol. Gen. 144 (1) (2015) 114.
[57] N. Castelo, M.W. Bos, D.R. Lehmann, Task-dependent algorithm aversion, J. Market. Res. 56 (5) (2019) 809–825.
[58] E. Bogert, A. Schecter, R.T. Watson, Humans rely more on algorithms than social influence as a task becomes more difficult, Sci. Rep. 11 (1) (2021) 8028.
[59] C.K. Morewedge, Preference for human, not algorithm aversion, Trends Cognit. Sci. 26 (10) (2022) 824–826.
[60] B.J. Dietvorst, J.P. Simmons, C. Massey, Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them, Manag. Sci. 64 (3) (2018) 1155–1170.
[61] B.J. Dietvorst, S. Bharti, People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error, Psychol. Sci. 31 (10) (2020) 1302–1314.
[62] J.M. Logg, J.A. Minson, D.A. Moore, Algorithm appreciation: people prefer algorithmic to human judgment, Organ. Behav. Hum. Decis. Process. 151 (2019) 90–103.
[63] S. You, C.L. Yang, X. Li, Algorithmic versus human advice: does presenting prediction performance matter for algorithm appreciation? J. Manag. Inf. Syst. 39 (2) (2022) 336–365.
[64] S. Krügel, A. Ostermaier, M. Uhl, Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions, Philosophy & Technology 35 (1) (2022) 17.
[65] K. Goddard, A. Roudsari, J.C. Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators, J. Am. Med. Inf. Assoc. 19 (1) (2012) 121–127.
[66] K. Goddard, A. Roudsari, J.C. Wyatt, Automation bias: empirical results assessing influencing factors, Int. J. Med. Inf. 83 (5) (2014) 368–375.
[67] Y.T.Y. Hou, M.F. Jung, Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making, Proceedings of the ACM on Human-Computer Interaction 5 (CSCW2) (2021) 1–25.
[68] S. Krügel, A. Ostermaier, M. Uhl, Algorithms as partners in crime: a lesson in ethics by design, Comput. Hum. Behav. 138 (2023) 107483.
[69] D. Castelvecchi, Can we open the black box of AI? Nature News 538 (7623) (2016) 20.
[70] M. Van Lent, W. Fisher, M. Mancuso, An explainable artificial intelligence system for small-unit tactical behavior, in: Proceedings of the National Conference on Artificial Intelligence, 2004, pp. 900–907.
[71] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
[72] B. Kim, R. Khanna, O.O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, Adv. Neural Inf. Process. Syst. 29 (2016).
[73] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.
[74] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the GDPR, Harv. JL & Tech. 31 (2017) 841.
[75] Y. Gu, J. Yang, N. Usuyama, C. Li, S. Zhang, M.P. Lungren, H. Poon, Biomedjourney: counterfactual biomedical image generation by instruction-learning from multimodal patient journeys, arXiv preprint arXiv:2310.10765 (2023).
[76] M. Nauta, J.H. Hegeman, J. Geerdink, J. Schlötterer, M.V. Keulen, C. Seifert, Interpreting and correcting medical image classification with pip-net, in: European Conference on Artificial Intelligence, Springer Nature Switzerland, Cham, 2023, September, pp. 198–215.
[77] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, C.C. Loy, Domain generalization: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2022) 4396–4415.
[78] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (2019) 1–9.
[79] M. Nagendran, P. Festor, M. Komorowski, A.C. Gordon, A.A. Faisal, Quantifying the impact of AI recommendations with explanations on prescription decision making, NPJ Digital Medicine 6 (1) (2023) 206.
[80] J. Jiang, S. Kahai, M. Yang, Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty, Int. J. Hum. Comput. Stud. 165 (2022) 102839.
[81] M. Vered, T. Livni, P.D.L. Howe, T. Miller, L. Sonenberg, The effects of explanations on automation bias, Artif. Intell. 322 (2023) 103952.
[82] L.R. Caporael, Anthropomorphism and mechanomorphism: two faces of the human machine, Comput. Hum. Behav. 2 (3) (1986) 215–234.
[83] R. Pohl (Ed.), Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory, Psychology Press, 2004.
[84] H. Albisser Schleger, N.R. Oehninger, S. Reiter-Theil, Avoiding bias in medical ethical decision-making. Lessons to be learnt from psychology research, Med. Healthc. Philos. 14 (2011) 155–162.
[85] I.E. Dror, Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias, Anal. Chem. 92 (12) (2020) 7998–8004.
[86] S.P. Clissold, Paracetamol and phenacetin, Drugs 32 (1986) 46–59.
[87] O.K. Sial, E.M. Parise, L.F. Parise, T. Gnecco, C.A. Bolaños-Guzmán, Ketamine: the final frontier or another depressing end? Behav. Brain Res. 383 (2020) 112508.
[88] M. Plebani, Quality indicators to detect pre-analytical errors in laboratory testing, Clin. Biochem. Rev. 33 (3) (2012) 85.
[89] R.M. Walton, Validation of laboratory tests and methods, Seminars Avian Exot. Pet Med. 10 (2) (2001) 59–65.
[90] H. Chen, C. Gomez, C.M. Huang, M. Unberath, Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review, NPJ Digital Medicine 5 (1) (2022) 156.
[91] A.F. Markus, J.A. Kors, P.R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies, J. Biomed. Inf. 113 (2021) 103655.
[92] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Inf. Fusion 76 (2021) 89–106.
[93] M. Nauta, J. Schlötterer, M. Van Keulen, C. Seifert, Pip-net: patch-based intuitive prototypes for interpretable image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2744–2753.
[94] E. Di Nucci, The Control Paradox: from AI to Populism, Rowman & Littlefield, 2020.
[95] S. Krügel, J. Ammeling, M. Aubreville, A. Fritz, A. Kießig, M. Uhl, Perceived Responsibility in AI-Supported Medicine, AI & SOCIETY, 2024, pp. 1–11.
[96] R. Noggle, The ethics of manipulation, in: E.N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, 2018.
[97] M. Ienca, On artificial intelligence and manipulation, Topoi 42 (3) (2023) 833–842.
[98] M. Dragoni, I. Donadello, C. Eccher, Explainable AI meets persuasiveness: translating reasoning results into behavioral change advice, Artif. Intell. Med. 105 (2020) 101840.
[99] S. Krügel, A. Ostermaier, M. Uhl, ChatGPT's inconsistent moral advice influences users' judgment, Sci. Rep. 13 (1) (2023) 4569.
[100] V. Kasten, Manipulation and teaching, J. Philos. Educ. 14 (1) (1980) 53–62.
[101] M. Klenk, (Online) manipulation: sometimes hidden, always careless, Rev. Soc. Econ. 80 (1) (2022) 85–105.
[102] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, C. Seifert, From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI, ACM Comput. Surv. 55 (13s) (2023) 1–42.
[103] Q.V. Liao, D. Gruen, S. Miller, Questioning the AI: informing design practices for explainable AI user experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, April, pp. 1–15.
[104] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[105] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.F. Bonnefon, C. Breazeal, M. Wellman, Machine behaviour, Nature 568 (7753) (2019) 477–486.