The Role of Empathy in Critical Reasoning and the Limitations of Medical AI Systems

Please cite the final version, forthcoming in the *Journal of Medicine and Philosophy*.

Abstract: *The recent developments of medical AI systems (MAIS) open up questions as to whether and to what extent MAIS can be modeled to include empathetic understanding, as well as what impact MAIS' lack of empathetic understanding would have on its ability to perform the necessary critical analyses for reaching a diagnosis and recommending medical treatment. In this paper, we argue that current medical AI systems' ability to empathize with patients is severely limited due to its lack of first-person experiences with human interests and that efforts to correct for this deficit – by having MAIS interpret patients' medical and non-medical interests – will encounter significant obstacles. Finally, we demonstrate how MAIS' lack of empathy is likely to hinder its performance in crucial aspects of the processes through which useful medical diagnoses are reached and through which appropriate treatment recommendations for patients are determined.*

Keywords: Medical AI Systems; Ethics of AI; Bioethics; Empathy; Kant

I. INTRODUCTION

Empathy is the capacity through which we represent the viewpoint of others and understand what having that viewpoint feels like. As such, empathy is often regarded as necessary for a "humanely grounded" approach to medicine because of its important role in clinical practice.[1] However, recent developments of medical AI systems (MAIS) have opened up questions as to whether and to what extent MAIS can be modeled to include empathetic understanding, as well as what the implication of MAIS being more or less able to perform critical reflection on the basis of empathy are. As the *Oxford Handbook of the Ethics of AI* states, "AI-driven diagnosis is certainly one of the most promising fields of application for AI in patient care" (Blasimme and Vayena, 2020, 708).

Some MAIS have already received clearance for marketing from the US Food and Drugs Administration (FDA) and have been deployed in the clinic. This includes algorithms used to detect wrist fractures in X-ray scans[2] and a machine learning software that detects diabetic retinopathy (DR) by automatically interpreting images from the back of the patient's eye.[3] Many more algorithms have appeared in the literature, including algorithms that can compute cardiovascular risk factors based on retinal images,[4] algorithms that can predict the decline of glomerular filtration rate in patients with polycystic kidney disease,[5] deep learning models that can process images from endoscopy and ultrasounds to detect abnormal structures such as colonic polyps,[6] and a machine learning algorithm used to identify cancer patients at high risk of thirty-day mortality before they start chemotherapy (both palliative and curative).[7] For most of these studies, "the performance of the algorithms was tested against the benchmark of certified specialists' assessments revealing equal or superior outcomes for AI systems as compared to human physicians" (Blasimme and Vayena, 2020, 708).[8]

Several of these MAIS have proven to be useful tools that are capable of providing guidance to physicians in their efforts to reach diagnoses and make treatment recommendations and there have been indications that medical experts working with MAIS are able to make more reliable diagnoses than either physicians without MAIS or the predictions of MAIS on its own.[9] However, many view the upside of these technologies as being associated with their ability to save valuable time and resources by allowing physicians to rely on the programs' suggested diagnoses or recommended treatments without having to independently analyze the relevant medical data themselves.[10] Furthermore, recent advances in large language models (LLMs) – with much attention going to the increased capabilities of programs such as ChatGPT – have led some to promote the development of generalist medical artificial intelligence (GMAI) that would provide diagnoses or treatment recommendations directly to patients using apps on their smartphones.[11] As we will be arguing, MAIS' inability to form an empathetic understanding of patients' interests is a significant and oft overlooked limitation that makes problematic both patients' and physicians' reliance on MAIS' suggested diagnoses and/or recommendations for treatment.[12]

The plan for the paper goes as follows. First, we draw from Immanuel Kant's notion of what today is considered 'empathy' – though which Kant labels as 'sympathy' – to highlight the epistemic and ethical functions of this capacity (Section II). Second, we argue that the possibility of current MAIS empathizing with patients is significantly limited by their inability to accurately interpret patients' medical and non-medical interests (Section III). Third, we argue that AI programs' lack of empathy will hinder its performance in two crucial stages of the process of reaching useful medical diagnoses and making appropriate treatment recommendations (Section IV). The first one is the ability to make judgments about the risks to patient interests when determining the relevant diagnoses or medical treatments that should be considered viable possibilities – a part of the diagnostic process which we refer to as 'abduction.' The second is the ability to make judgments about the risks to patients' interests when deciding which of the diagnoses identified through abduction should be given priority in further investigations and which of the treatments identified through abduction should be given preference as the first to try – a part of the diagnostic process which we refer to as 'prioritization'.

Finally, we point out how AI programs' lack of empathy should inform decisions about their proper use as a medical technology. On this matter, we reach three conclusions: 1) MAIS' lack of empathy means that they should not be used as a more general diagnostic tool for reaching diagnoses or recommending treatment for multiple conditions; 2) Since MAIS' lack of empathetic understanding leads these AI programs to be insensitive towards aspects of patients' situations that may be highly significant for making judgments about what will hinder/further those patients' interests, the use of these programs should involve high levels of supervision from physicians; and 3) The less opaque programmers are able to make the algorithms by which MAIS reach outputs, the more useful these outputs will be for physicians when reaching a diagnosis or making a treatment recommendation.

II. KANT ON THE ROLE OF SYMPATHY/EMPATHY IN CRITICAL REASONING

To get a proper handle on empathy's significance in the context of medical AI, we have to reckon with why and how empathy matters, both epistemically and morally. This is not a novel subject, as it has been explored both in the contemporary philosophical literature about emotions and in the history of philosophy. One philosopher who offers an account of empathy that is able to cash out both its moral relevance and its epistemic significance is Immanuel Kant. On Kant's view, the virtuous moral agent relies on the feeling of sympathy as an epistemic aid to act from duty. That Kant attributes this special role to the feeling of sympathy can be seen from his remarks in the *Metaphysics of Morals* that

> While it is not in itself a duty to share the sufferings (as well as the joys) of others, it is a duty to sympathize actively in their fate; and to this end it is therefore an indirect duty to cultivate the compassionate natural (aesthetic) feelings in us, and to make use of them as so many means to sympathy based on moral principles and the feeling appropriate to them. – It is therefore a duty not to avoid the places where the poor who lack the most basic necessities are to be found but rather to seek them out, and not to shun sickrooms or debtors' prisons and so forth in order to avoid sharing painful feelings one may not be able to resist; for this is still one of the impulses that nature has implanted in us to do what the representation of duty alone might not accomplish. (MS 6:457)

Here, Kant writes that, while it is not a duty to be in the same emotional state (e.g., of grief) of somebody else, what is a duty is to sympathize actively in their fate. So, we have a duty – though indirect, but we don't have to worry about this characterization here – to cultivate the sympathetic feelings in us, and to make use of them. Kant then gives us some examples of how we should cultivate such feelings, and he give us some indication of how we should make use of them. He writes that we should not avoid the places where the poor who lack the most basic necessities are to be found but rather to seek them out; and we should not shun sickrooms or debtors’ prisons. What is the use we should make of the sympathetic feelings that might arise from these experiences? The answer is that such feelings are supposed to do “what the representation of duty alone might not accomplish”.

As many commentators have pointed out, the passage is at first sight very puzzling. Acting from duty, Kant tells us throughout his work (e.g., in *Groundwork* I), cannot rely on the motivational import of any sensible feeling; rather, it must be done out of the representation of duty alone. So, what does Kant mean when he claims that sympathy does what the representation of duty alone cannot do? The answer is that, as scholars like Nancy Sherman and Anne-Margaret Baxley have pointed out, sympathy works as an epistemic aid, as opposed to a motivational one, that helps the virtuous moral agent to act from duty. How does it do that? Emotions such as sympathy serve as ‘modes of attention’ that help us discern what is morally relevant as such. Sherman writes:

> The point is that the thought of duty alone is insufficient to provide information about which objects and circumstances require our moral attention. Given a practical interest in the moral law and its spheres of justice and virtue, we still require further information about when and where and how to deploy our practical interest. And such information is often provided through the emotions. Capacities for grief prime us to notice human mourning and loss; capacities for compassion, to notice that others suffer in ways that often seem undeserving. (Sherman, 1997, 146)

Sympathy in particular, by drawing us to occasions of distress or need, gives us epistemic insight that informs our practical judgments about what we ought to do.[13]

An important point is that, on Kant’s view, sympathy not only tells us *that* others suffer, but also tells us *what* their suffering feels like. As Baxley writes, “sympathy brings to our attention certain facts about the lives of other people in ways that lead us to give due weight to the reasons these facts contain" (Baxley, 2010, 164). Kant makes this clear when he writes that "to put ourselves in the other's place [..] is sympathy" (VAnth 25:476). He also writes that "one calls sympathetic feeling humane [*menschlich*]" (VAnth 25:608) and being humane is the "capacity and will to communicate with one another [*sich einander ... mitzutheilen*] with regard to one's feelings*"* (MS 6:456). So, being sympathetic amounts to communicating with the other, putting oneself in the other's place, and 'tuning in' to the other's point of view by means of feeling. This amounts to gaining a first-personal understanding of what it feels like to have their point of view, and is what, according to Kant, *being humane* means.

Though often labeling as "empathy" what Kant calls "sympathy"[14], contemporary philosophers have focused on the same features of this feeling and attributed to it a similar role. For instance, L. A. Paul describes empathy as imaginatively representing someone's situation such that one simulates their point of view or what it is like to be them. Doing this, Paul argues, allows one to "*understand* their point of view, in a way that *explains* their point of view" (Paul, 2021, 3). According to Olivia Bailey, empathy is the unique source of a particular form of understanding, namely humane understanding. To humanely understand another's emotion "means imaginatively taking up the other's first-personal perspective" (Bailey, 2022, 57) such that "we do not merely imagine *that* we are feeling some emotion. Rather, we actually experience an emotion" (Bailey, 2022, 53).[15]

Some might call into question the possibility of imaginatively reconstructing *exactly* how the other feels. But it is important to notice that achieving humane understanding through empathy does not require one to feel *exactly* as the other feels. As several scholars have pointed out, in order to empathize with others, it is sufficient that, as Bailey puts it, "the emotional experience of the one who empathizes closely resembles the emotional experience of the target of empathy" (Bailey 2022, 52). That is to say, one can still gain these significant moral insights by taking up a point of view that *closely resembles* the other's and this does not require an *exact replication* of the other's emotional experience. Joel Smith argues that what matters for empathy is that that the emotional experience of the one who empathizes *affectively matches* the emotional experience of the target of empathy. On Smith's account, this affective matching will depend on the extent according to which the one who empathizes knows, or is aware of, how the target of empathy feels. "Even if A is not, and has never been, in exactly the same psychological state as B, she may nevertheless be, or have been, in a state that affectively matches it at some level of determinacy. It is reasonable to suppose that such levels of determinacy will correspond to the extent to which A knows, or is aware of, how B feels" (Smith, 2017, 715).

Another objection that has been raised against the notion that empathy can provide us with important moral information is the suggestion that empathy is often subject to bias (Prinz, 2011a, 2011b; Hoffman, 2000; Hoffman, 2011). Prinz, for instance, claims that "we are grotesquely partial to the near and dear" and thus that in empathizing with others, we run the risk of "profound moral error" (Prinz, 2011b, 224). Hoffman writes that we are more prone to empathizing with those who look similar, are most familiar, and are proximally closer: "Although people tend to respond empathically to almost anyone in distress, they are vulnerable to bias in favor of victims who are family members, members of their primary group, close friends, and people who are similar to themselves; and to bias in favor of victims who are present in the immediate situation" (Hoffman, 2000, 13–14).

For Kantian empathy, however, this objection appears to miss the mark. To see this, recall that Kant tells us that it is a duty to empathize precisely with those against whom we tend to be negatively biased – from Kant's perspective, these would be, for example, the poor, the sick, the convicted. In claiming that we have a duty to actively gain understanding of the perspective of those who are not dear and near, Kant seems to hold that empathizing must be conducted such that we are aware of our tendency of being biased and that we try to correct for it. Doing so requires making a conscious effort to include those perspectives we are tempted to exclude. After all, in-group bias, selectivity, arbitrariness, proximity effect are examples of empathy for just one set of perspectives to the exclusion of the others. That these phenomena occur is reason to hold, like Kant does, that empathy's cultivation should be conducted in certain ways rather than others; it is not reason to think that that empathy is not necessary for our moral lives.

These considerations shed light on an important feature of Kantian empathy: the kind of empathy that Kant recommends cannot be a non-conscious and automatic coming to feel as another feels simply because another feels that way. Kant expressively discourages this kind of empathy, which he describes as a mere "receptivity, given by nature itself, to the feeling of joy and sadness in common with others" and calls it "unfree" (MS 6:456–7). In contrast, the empathy that Kant encourages requires some kind of cognitive effort on the part of the empathizer, including imaginative, perspective-taking and bias-corrective elements, and can thus be regarded as higher-order empathy.[16] Only this higher-order empathy, according to Kant, "is based on practical reason" and is therefore "free" (MS 6:456–7).[17]

A final objection that we want to consider pertains to whether we would be better off cultivating emotions other than empathy. For example, in his work in psychology Paul Bloom distinguishes between rational compassion and empathy, and argues that we should cultivate only the former. Bloom holds that "in contrast to empathy, compassion does not mean sharing the suffering of the other. Rather, it is characterized by feelings of warmth, concern and care for the other, as well as a strong motivation to improve the other’s well-being” (Bloom, 2016, 138). Bloom then claims that compassion is less likely than empathy to motivate morally problematic behavior. For instance, while empathizing with another can be so painful that those experiencing it may deliberately think of something else or leave the situation in order to alleviate their own distress (Prinz, 2011a; Prinz, 2011b; Hoffman, 2011), those with compassion experience “a warm positive feeling with a strong prosocial motivation” (Bloom, 2016, 139).

The problem with this line of reasoning is that compassion may very well fail to provide proper moral and epistemic guidance without the imaginative, effortful, and phenomenological prospective-taking that empathizing provides. That is, if one were to be compassionate without being empathetic, one would not be able to properly understand what the perspective of the other would be in order to promote their well-being. This means that compassion, when considered on its own, is flawed in an important way: it is possible to have compassion for somebody while misunderstanding what is in their best interest, what would constitute their well-being, and what would generally help them. Here, we agree with Shoemaker when he argues that empathy is morally significant because it allows us to grasp second-personal reasons embedded in the emotional perspectives of others, and this is always a moral improvement: “Our coming to recognize the emotional perspectives of others is *always* a moral improvement, for those perspectives are the source of a central category of moral reasons, namely, the second-personal reasons expressed in the authoritative demands we make on one another” (Shoemaker, 2014, 513). Shoemaker then goes on to argue that, even if one doesn't buy that these reasons are coherent or a central part of morality, favoring instead something like purely agent-neutral consequentialist reasons, “any moral reasons having to do with welfare, happiness, or utility generally are going to have to draw from facts about the psychological well-being, preferences, or utility functions of others, and these must make reference to agents' actual perspectives" (Shoemaker, 2014, 513).

To sum up: Kantian sympathy/empathy involves taking up the point of view of the other, and imagining what it feels like to have that point of view. As such, empathy plays the epistemic role of revealing important aspects of the other's point of view that are only salient from a first-person perspective. This includes things such as the person's conception of their own happiness and psychological well-being, the values they identify as promoting these, and their preferences regarding how to further those values. In the following sections, we argue that MAIS do not appear to be capable of empathizing in a way that would enable them to properly understand this type of information and that such information is crucial for understanding the inductive risks associated with reaching a diagnosis or providing recommendations for medical treatment.

III. ARE MAIS CAPABLE OF HAVING EMPATHETIC UNDERSTANDINGS OF PATIENTS?

At this point, one might ask: are medical AI systems capable of feeling emotions in general, and empathy in particular? Since Rosalind Picard's 1997 *Affective Computing*, a case can be made for AI 'having' emotions. Picard's point was that AI are capable of detecting and recognizing human emotional states and providing convincing feedback, thus creating a satisfactory emotionally intelligent interactive framework. Moreover, Picard argued that "it is possible to give computers a modicum of self-awareness: a name for where they are, what they are doing, what state they are in affectively, what their sensors are reading, the name of the problem they are trying to solve, and so forth" (Picard, 1997, 75). The upshot of this reasoning is that "we can argue that machines might have some awareness, and some feelings. However, [..] we can expect the internal subjective feelings to differ between human and machine" (Picard, 1997, 75). This means that, according to Picard, AI can show emotional activity and could be said to ‘have’ emotions, though in a limited sense that allows for AI to have emotional internal conscious states but rules out that those could be of the same kind than the emotional internal conscious states of human beings. The reason for ruling this out, according to Picard, is simply because machines and humans have radically different physiologies – different ‘brains’ and ‘bodies’.[18]

However, even proponents of the view that AI ‘have’ emotions have often been skeptical about AI’s capacity for feeling empathy. The reason for this is that, as we have seen, empathy is that emotion through which we accurately represent the viewpoint of others and understand what having that point of view feels like. Presumably, the capacity for empathy is contingent upon having the capacity for experiencing the same kind of internal conscious states. As we have seen, however, even if we were to grant that AI can have some emotional internal conscious states, these states would be of a different kind than the emotional internal conscious states of human beings. Thus, it is concluded that AI cannot feel empathy, at least towards humans.[19]

Some may wonder whether recent advances in LLMs might allow MAIS to obtain a form of humane understanding pertaining to the relevant information about patients’ values and preferences by prompting certain inputs from patients. Although most current MAIS have not been designed to regularly receive information about patients’ interests, one may be inclined to believe that this is something for which future models would be able to correct. Against this, we argue that, even if MAIS were to have access to patients’ medical and non-medical interests, it seems highly unlikely that they could become capable of accurately interpreting this type of information – and thus gain proper humane understanding – precisely because of their lack of first-person experiences.

To see why this would be the case, we will first consider the kinds of strategies that physicians use to obtain a humane understanding of their patients' interests and show why MAIS are unable to replicate these strategies. We will then consider the only recent attempt – conducted by Meier et al., 2022 – to use recent improvements in AI capabilities to design algorithms that would be capable of making reliable suggestions for ethical decisions in medical contexts. As we will argue, the efforts of Meier et al. reveal how the latest AI do not appear to be able to interpret information about human interests and values that would enable them to reach accurate normative conclusions in new scenarios. As such, MAIS' current failures to interpret human interests and values are attributable to two things: 1) MAIS are unable to achieve an empathetic understanding of human perspectives; and 2) Programmers have not developed any viable alternatives of how MAIS could learn the information that is gained when one achieves a humane understanding of others. Although the Meier et al. algorithm's task of suggesting ethical decisions to physicians is significantly different from the tasks of reaching a diagnosis or making a treatment recommendation, we will argue in the following section that MAIS' inability to adequately interpret human interests and values would also be a significant hindrance to successfully performing these latter tasks.

To explain MAIS' struggles to adequately account for human interests and values in decision-making, it is worth considering the strategies that physicians themselves are trained to use for improving their understanding of patients' interests and values. For instance, Samuel Gorovitz writes that "there is nothing like being a patient to open the eyes of a physician to the perils and indignities associated with receiving medical care," but "rarely does the experience provide them with technical data or other factual knowledge about medical care that was not previously known to them. Rather, what they gain is a sense of what it feels like to be a patient –

to undergo the fears, confusions, angers, and hopes that even they face when cast in the patient's role" (Gorovitz, 1988, 435).

The difficulties for MAIS that Gorovitz's discussion highlights are: 1) MAIS do not have available to them the strategy of gaining their own similar first-person experiences to improve their empathetic understanding of patients' interests; and 2) It is unclear how one could convey this type of information without in some way appealing to first-person experiences of holding such interests or values. Although physicians do not always need to gain first-person experiences that are very similar to a patient's in order to obtain an empathetic understanding of that patient's interests and/or values, there does appear to at least be a need for physicians to draw upon some commonalities with their own first-person experiences that would allow them to imaginatively reconstruct those patients' perspectives. This means that programmers would be facing the seemingly intractable problem of needing to convey this information in a way that would be intelligible for a MAIS that lacks any kind of embodied first-person experiences. In this regard, it is not clear at all from what basis MAIS could begin to accurately interpret these aspects of the first-person accounts of patients' experiences. And, as we will see, efforts to do precisely that have not come close to making up for the information lost from developing this type of empathetic understanding of others.

For instance, let's consider the recent effort of Meier et al. to have AI usefully interpret information about human values and interests by designing a 'fuzzy cognitive mapping' (FCM) algorithm – which is a type of algorithm that incorporates deep neural networks and the symbolic methods of LLMs[20] – which they hoped would be capable of providing physicians with recommendations for some of the ethical decisions that they are likely to face in medical clinics. An important reason for focusing on the work of Meier et al. is that, as they note, "as of today, no machine-intelligence systems exist that are designed specifically for the making of sophisticated moral decisions" and "to the best of our knowledge, the creation of an algorithmic advisory system for clinical ethics has only been attempted once, and it was not developed beyond the early prototype stage" (Meier et al., 2022, 4). In describing how their algorithm learned to make such decisions, Meier et al. write that "the algorithm gradually learned which constellations of input parameters are supposed to be associated with which ethical outcome" (Meier et al., 2022, 12). In this way, Meier et al. used a set of 20 parameters that were meant to capture connections between human interests (e.g., the patient wanting to make the choice for themselves) and human values (e.g., respecting patient autonomy) to interpret the context in which the ethical decision is being made.

However, as Barwise and Pickering point out, the relevant "ethical decision-making cues… are nuanced, not readily accessible to a machine, and the outcomes are highly contextual" (Barwise and Pickering, 2022, 47). Amongst these nuanced cues are the features of a context, such as the patient's interests and values, that physicians will not adequately grasp unless they develop an empathetic understanding of them. In this regard, it is unsurprising that the algorithm of Meier et al. was unable to use the inputs about patients' relevant interests and values to reach appropriate conclusions in new contexts. As Barwise and Pickering conclude, although "the authors make some attempts to address the constraint," the solution of Meier et al. "is a highly artificial construct that fails to demonstrate that the proposed design concept is feasible outside of their unique training setting" (Barwise and Pickering, 2022, 47). Here the problem is that the AI is unable to interpret these interests or values in a way that would allow them to properly identify the decision-making contexts in which each should be given decisive weight.

Altogether, the processes that human doctors use to improve upon their own empathetic understandings of patients seem to be out of reach for MAIS and there does not appear to be any prospect of MAIS learning to compensate for their lack of empathy in a way that would allow them to accurately interpret inputs about patients' interests or values. While we are not claiming to have proven that empathy is the only possible way that one could gain the information associated with developing a humane understanding of the other person, there have not been any viable proposals of alternative ways for AI to do so and it is difficult to see where one could even start given MAIS' lack of first-person, embodied experiences. That is to say, unlike the physician, MAIS do not have the option of drawing upon similarities in their own first-person experiences to reach the kind of humane understanding that would allow them to adequately interpret information about patients' interests or values and this appears to be essential to any recognized strategies that people have for gaining these insights. Without MAIS having at least some of its own experiences of the fears, confusions, angers, and hopes that are part of what it means to have human interests or values, it is unclear on what basis these programs could start to form an empathetic understanding of their patients' interests or values. And, as we will now argue, an adequate understanding of patients' interests/values is crucial for reaching diagnoses and making treatment recommendations on which physicians can rely.

IV. HOW LACKING EMPATHY LIMITS MAIS' ABILITY TO REACH USEFUL DIAGNOSES AND MAKE TREATMENT RECOMMENDATIONS

We turn now to the importance of empathy in reaching medical diagnoses and making treatment recommendations. In the context of medical practice, we will focus on the ability to reach a 'useful diagnosis'; i.e., a diagnosis that provides guidance for patient treatments that would increase the likelihood of a positive real outcome. As Elfiky et al. note, "to be useful, predictive models must improve decision making in the real world. Thus, rigorous evaluation of predictions' influence on outcomes is the criterion standard test but one that is often neglected in the literature, which focuses primarily on measuring predictive accuracy rather than real outcomes" (Elfiky et al., 2018, 10). While we are willing to grant that empathy may not be needed for achieving the narrower predictive accuracy mentioned in the quote above, we wish to suggest that empathy plays a crucial role in making decisions about when reaching a diagnosis or recommending treatment is appropriate given the associated inductive risks. Moreover, we claim that the opacity of many current MAIS will make it even more difficult for doctors or patients to rely upon MAIS' diagnoses and/or treatment recommendations.

There are two important stages in the process of making useful medical diagnoses that medical AI programs' lack of empathy would appear to hinder. The first we will refer to as 'abduction' or 'abductive reasoning,' which is the ability to make judgments about which hypotheses have the potential to explain a particular patient's condition. The second we will refer to as 'prioritization,' which is the ability to make decisions about which of the options identified by the process of abduction are worth further investigating and the order in which these should be investigated. In both cases, we are claiming that medical AI programs' inability to form an empathetic understanding of human interests will prevent them from being able to make appropriate calculations of the inductive risks (i.e., the risks associated with making a particular inductive leap from the given evidence to a conclusion or belief) associated with: a) Missing the diagnosis of another relevant medical condition; and b) Mistakenly recommending treatment for the diagnosis it favors.

Given the considerable harms that can result from these failures to appropriately understand the inductive risks in question, we will be reaching three conclusions about the limitations of current AI in medical diagnoses. First, unless MAIS become capable of empathetic understanding, their use in diagnosis should be limited to very specific and well-defined tasks. Second, a high degree of ‘supervision’ of MAIS – i.e., these AI programs’ outputs being reviewed by physicians in the diagnosis and treatment of a particular patient – will be necessary for their safe use. Third, the more opaque the MAIS, the less useful its outputs will be to physicians when reaching diagnoses or making treatment recommendations. Ultimately, though MAIS may well prove to be useful tools for physicians, we argue that their limitations speak against the view that MAIS’ diagnoses would allow physicians to forego analysis of the relevant medical data and against the notion that MAIS would have many safe applications for direct patient use.

A. How MAIS’ Lack of Empathy Impacts its Performance of Abduction

To better understand the way that MAIS’ lack of empathy will hinder its ability to contribute to useful medical diagnoses, we should first consider the role ‘abduction’ plays in doctors’ processes of reaching a diagnosis and recommending treatment.[21] Doctors must be able to evaluate the first-person account patients provide regarding what has led them to seek medical assistance and must identify which features of the patient’s experiences are likely signs of ‘abnormal’ body functions (i.e., symptoms of a medical condition) and which of these fall within the ‘normal’ range of the patient’s bodily functions.[22] This is the first place in which the process of abductive reasoning in medical diagnosis requires an understanding of human interests and experiences. Doctors must identify what parts of the patient’s experiences indicate potential threats to their interests that could be caused by medical conditions.[23] Without such evaluations, the doctor or MAIS would be unable to account for the risks associated with both misdiagnosis and failing to diagnose that will inform their decisions about whether reaching a diagnosis or making a treatment recommendation is appropriate at this time given the information that is available to them.

Once certain aspects of the patient's bodily functions are identified as 'symptoms' that have the potential to problematically interfere with the patient's interests, doctors must then make abductive judgments about what medical conditions are relevant to explaining those symptoms.[24] These types of judgments do not simply involve identifying the most obvious or most likely causes of these symptoms, but also involve an understanding of both the extent to which a given diagnosis is underdetermined by the evidence, the possible areas of deficiency in the evidence that have been obtained,[25] and the severity of the risks to the patient's interests. For the first of these, doctors will want to determine what conditions should be seriously considered according to the likelihood of the patient having each particular disease and the severity of the harms that would result from mistakenly ruling out this option.[26] While MAIS may be able to reach accurate judgments about the statistical probability of the patient having a particular disease, the MAIS' inability to empathize with patients makes them unlikely to be able to reach proper judgments about the severity of the harms associated with missing other possible diagnoses. Similarly, while MAIS might be able to predict outcomes of treatment for patients in a high percentage of cases, it will struggle to make evaluations of which of the treatments for different diagnoses will pose unacceptable risks on patients given the treatments available for the other possible, though less likely, diagnoses under consideration. Again, doctors' ability to empathize with patients will play a crucial role in determining whether the evidence they have is sufficient for reaching a diagnosis in relation to the potential risks to the patient's interests. This in turn will inform doctors' judgments about whether more information must be obtained before reaching a diagnosis. As Balogh, Miller, and Ball iterate, the decision to "begin treatment based on a working diagnosis is informed by: (1) the degree of certainty about the diagnosis; (2) the harms and benefits of treatment; and (3) the harms and benefits of further information-gathering activities, including the impact of delaying treatment" (Balogh, Miller, and Ball 2015, 49). In other words, reaching a diagnosis that can actually be 'useful' – in the sense that it could be used to guide decisions about whether to treat and which treatments to recommend – or choosing to refrain from doing so until further information is acquired, requires an ability to first account for these harms and benefits to the patient's interests.

The role of empathy in these different parts of the 'abductive' process for making medical diagnoses appears to be capable of explaining some of the ways in which the use of MAIS is currently limited to particular kinds of tasks. For instance, let's consider the IDx-DR program. First, this program is "only designed to detect diabetic retinopathy. IDx-DR is not intended to detect concomitant diseases. Patients should not rely on IDx-DR for detection of any other disease" (US FDA, 2018a, 1). Moreover, "IDx-DR is only for use in people already diagnosed with diabetes mellitus" (US FDA, 2018a, 2). As these descriptions indicate, the IDx-DR program's uses are limited to the detection of a single condition and, even then, it is not recommended for patients who have not already been diagnosed with diabetes mellitus. Far from being a feature of the early stages of the technology, the need to limit MAIS to testing for only a single disease that medical experts believe the patient is already predisposed to develop is precisely the type of significant limitations that we believe MAIS is unlikely to overcome because of their inability to make accurate judgments about the inductive risks involved in the process of abductive reasoning. Without a sense of the risks to the patient's interests that would result from misdiagnosis or a failure to diagnose, MAIS will be more likely to reach medical conclusions that would be inappropriate given the limitations in the information on which they are relying to reach a diagnosis.

Second, despite these limitations in the cases in which it has been approved to be used, IDx-DR still requires significant supervision from medical experts both in reviewing the diagnosis it provides and before using this diagnosis to recommend potential treatments to the patient. As the FDA notes, IDx-DR has patients' images "analyzed to determine whether further examination is needed by an eye care provider. Physicians should review IDx-DR results and advise patients of recommended referrals to an eye care provider for evaluation and potential treatment" (US FDA, 2018a, 2). Not only are physicians required to review the outputs of IDx-DR, but an entirely independent evaluation by eye care providers is further needed to determine potential treatment recommendations. Again, our argument is that the need for this kind of supervision is not just an initial precaution to be discarded after the technology has proven to be reliable, but a necessity following from MAIS' inability to develop empathetic understandings of patients' interests. To be clear, we acknowledge that IDx-DR has proven useful in guiding physicians' decisions about whether to perform additional tests for diabetic retinopathy and reach their own diagnosis given a review of the medical data available to them. What IDx-DR cannot – and, we are claiming, there is not a foreseeable way in which MAIS would be able to – do is provide a diagnosis on which the patient or the physician should rely upon for making subsequent treatment decisions.

B. How MAIS' Lack of Empathy Impacts its Performance of Prioritization

Next, we should consider the process of 'prioritization' that follows the identification of the possible relevant medical conditions and the treatments for those conditions. The process of prioritization in diagnoses involves medical experts making judgments about which explanations (out of those that have been identified as relevant possibilities) should be the first to be investigated or treated. As Balogh, Miller, and Ball note, "of major importance in the diagnostic process is the element of time" and how this requires physicians to make judgments about when "some diagnoses may be more important to establish immediately than others" (Balogh, Miller, and Ball, 2015, 51). This is because, in some cases, "the benefit of treating the disease promptly can greatly exceed the potential harm from unnecessary treatment," while in other cases "the potential harm from rapidly and unnecessarily treating a diagnosed condition can lead to a more conservative approach in the diagnostic process" (Balogh, Miller, and Ball, 2015, 51). Consequently, doctors may order tests or even begin treating quickly-evolving conditions that are possible, but unlikely, explanations of particular symptoms when the potential harm of failing to treat those serious conditions early is a significant risk to the patient's interests. In making these kinds of judgments, an understanding of patients' values and how patients are likely to prioritize these – as well as an understanding of when it is reasonable to ask about how the patient prioritizes these[27] – is crucial for deciding when reaching a diagnosis is likely to increase the chances of a positive outcome for the patient.

These evaluations regarding which treatments have a substantial risk to patients and how this should influence the order in which the possible diagnoses are to be further investigated will depend on having an accurate understanding of patients' interests. Moreover, part of this process will be for doctors to identify what kinds of tradeoffs in patients' interests will be relevant to which treatments they recommend and to ask patients for more information about how they prioritize the interests in question in order to make a proper recommendation. Balogh, Miller, and Ball point out that “the goal of patient engagement in diagnosis is to improve patient care and outcomes by enabling patients and their families to contribute valuable input” including their “varying needs, interests, and preferences,” which require their “roles in diagnosis” to be “individually tailored” (Balogh, Miller, and Ball, 2015, 153). As discussed above, MAIS’ lack of empathy makes it unlikely that they will either: a) be able to recognize when there are significant human interests at stake in the treatments of the different possible diagnoses such that they will perceive the need to ask patients to provide further information about their prioritization of these interests; or b) be able to interpret patients’ reports about the differences in the value they assign to each of the relevant interests even if these were provided to them as inputs.

To get a better sense of this, let’s consider the discussion of Elfiky et al. about machine learning programs that are designed to help doctors make decisions about whether to start a given cancer patient on chemotherapy. In this case, MAIS has access to significant amounts of information about the particular patient’s medical history (including quantitative information from recent medical tests, the patient’s diagnosis in having a particular kind of cancer, etc.) and will compare this to previous cancer patients’ medical histories in order to make a prediction about whether the patient will die within two weeks of starting chemotherapy. The stakes in making these kinds of medical judgments are clearly very high. In making these kinds of decisions, doctors and their patients must balance the significant harms associated with chemotherapy – including the possibility of the chemotherapy itself causing the death of the patient – with the harms of leaving the cancer untreated or treating the cancer in another (sometimes significantly less likely to be successful) way. If the MAIS could predict with certainty that the patient will die from chemotherapy in the next two weeks, then this would seem to make the decision straightforward: the cost is clearly too high to be worth starting the chemotherapy treatment. Of course, this is not a realistic goal for any MAIS to achieve and the inductive risks that become relevant when making judgments involving any uncertainty – even in cases of *near* certainty – will be problematic. Due to the enormous stakes involved, even a small chance of chemotherapy successfully treating the patient's cancer could still be incredibly relevant to patients' decisions. If the choice is between having a high chance of dying from chemotherapy in the next two weeks and a small chance of having one's lifespan extended significantly because the treatment is successful vs. having a high chance of dying in six months with incredibly miniscule chances of living beyond that without chemotherapy, then the MAIS recommending against a patient undergoing chemotherapy may well not be properly accounting for the inductive risks of being mistaken in reaching this conclusion.

This example illuminates some of the major concerns about physicians relying on MAIS' recommendations to make these types of treatment decisions. First, if the MAIS' calculation of the likelihood of death in the next two weeks is opaque to doctors, then it will be unclear to them to what extent this AI program is considering different possible explanations of the patients' symptoms and the risks associated with the different forms of available treatment. Second, given this opacity, it will not be possible for doctors to add their empathetic understanding of patients' interests to determine whether the strength of the evidence that the MAIS is considering is sufficient to outweigh the inductive risks. That is to say, not only would the MAIS' output be uninformed by empathetic considerations of the risks to the patient that are at stake, but doctors would be unable to supplement the MAIS' output with their own empathetic understandings because doctors would not have access to the evidence that such 'blackbox' AI programs are considering.

As one can see, the difficulties that MAIS will have in the ‘prioritization’ of diagnoses to further investigate will be substantial and these arise directly from the way that MAIS’ lack of empathy diminishes their abilities to accurately calculate the risks to patients’ interests. While it does seem that MAIS could be useful tools for doctors to quickly obtain a more precise understanding of the statistical likelihood of the patient having a certain condition, they will still need to conduct an independent analysis of the evidence available to them to be able to weigh this against the relevant inductive risks before making a judgment about which diagnoses should be investigated at this time and when – as well as which – medical treatments should be recommended to their patients. These are not simply decisions about what is the most likely cause of the patient’s symptoms, but judgments about what other conditions *could* be the cause given the information available to them and which of these conditions is most likely to be harmful to the patient’s interests when it is not diagnosed quickly or when the recommended treatments for a particular condition are more likely to threaten certain human interests.

Given the struggles that MAIS are likely to encounter within the ‘abduction’ and ‘prioritization’ parts of reaching a useful diagnosis, due to their lack of empathy, there are severe limitations on the extent to which these programs are capable of providing diagnoses or treatment recommendations on which physicians can rely. First, physicians must make judgments to help patients decide when testing for certain medical conditions is appropriate and when the ‘symptoms’ they are experiencing are not causes for concern (i.e., when such ‘symptoms’ fall within the normal range of the body’s functioning and are not being caused by an underlying disease). Second, physicians must make judgments about what possible causes of the patients’ symptoms are reasonable to consider and the potential harm to patients in failing to take any of these possibilities seriously before reaching a diagnosis. Third, physicians must make judgments about the safety of available treatments for the different possible diagnoses that remain viable explanations of the patients' symptoms and know when these treatments are likely to threaten patients' interests. Fourth, physicians must consider the tradeoffs in human interests that inform their decisions about which of the possible conditions they should recommend investigating further and which of these conditions they will recommend investigating before the others. Finally, physicians must make judgments about the tradeoffs in human interests that inform their decisions about what treatments would need to be started immediately and which can wait without serious risks to the patient's interests. And, as we have argued, MAIS' lack of empathetic understanding makes them unlikely to be able to sufficiently improve their performance on any of these tasks.

C. Recommendations for Future Uses of MAIS Given its Limited Capacity for Empathy

Finally, we wish to briefly outline the implications that MAIS' lack of empathy will have for future developments of the technology and how this should inform decisions about their proper use in processes of diagnosing patients or recommending medical treatments. The first conclusion to draw is that, unless these AI programs become capable of empathetic understanding, they should only be used for very specific and well-defined diagnostic tasks. That is to say, MAIS should be limited to more narrow tasks of diagnosing a single condition and should be primarily used to help medical experts to confirm or further investigate a specific diagnosis that they have already identified via the 'abduction' process. To the extent that a MAIS' process of reaching a diagnosis is opaque to medical experts, its use should be limited to situations where alternative diagnoses for patients experiencing particular symptoms are highly unlikely and where the risks of overlooking those diagnoses are sufficiently low. By keeping MAIS' tasks limited to helping doctors confirm a specific diagnosis they already suspect is the one that should be investigated at this time, one is able to limit the situations in which MAIS is being asked to make judgments between diagnoses and in which the inductive risks of each would need to be weighed.

A second – and closely related – conclusion to draw is that MAIS should be used to compliment, rather than replace, the diagnoses reached by physicians and the treatment recommendations they make. This would recommend against the use of MAIS that are not highly supervised by physicians; e.g., apps designed primarily for patients' direct use. The less physicians contribute to limiting the range of the likely diagnoses and/or treatments that are being considered for the patient, the riskier patients' reliance on these AI programs' outputs will be. Since MAIS' lack of empathetic understanding leads them to be insensitive to aspects of patients' conditions that may pose significant risks to their interests, physicians should only be using the recommendations of MAIS as one data point among several when reaching a diagnosis or a decision about treatment. To be clear, most of the MAIS that are currently being used have appeared to acknowledge this need to limit the risks of false positives by keeping the MAIS' recommendations to things such as "See your doctor for further examination and treatment" and the risks of false negatives by providing outputs that "strongly encourage" patients to "test again at an appropriate time in the future" (US FDA, 2018a, 2). And, to reiterate, we wish to emphasize the fact that such limitations should not be treated as precautionary – i.e., ones expected to be lifted once the technology has been proven reliable – but as necessary limitations for ensuring that MAIS' lack of empathetic understanding does not unnecessarily increase the risks to patients' interests.

The final conclusion we wish to emphasize is that MAIS will be more useful as tools to help doctors reach diagnoses and make treatment recommendations when the ways in which they reach such recommendations are made less opaque. The more doctors understand about the processes that lead MAIS to provide certain outputs, the better those doctors will be able to make judgments about the inductive risks associated with MAIS' recommendations. This is not an uncontroversial recommendation and some have argued against the notion that the 'black box' aspect of MAIS is problematic for its use in medicine. For instance, Cadario, Longoni, and Morewedge argue that the concerns about 'black box' algorithms are misplaced because "human decision-making is often as much of a black box as decisions made by algorithms" (Cadario, Longoni, and Morewedge, 2021, 1636).[28] While it may be true that it will be difficult for patients to identify when their doctors have developed an empathetic understanding of their interests and have used this to calculate the inductive risks, this does nothing to eliminate the concerns that MAIS appear to be incapable of doing so. Moreover, since physicians cannot rely on MAIS to have a sufficient empathetic understanding of the patient's interests, they cannot assume that MAIS have not overlooked the way that some of this information could have an impact on their patients' interests. As a result, efforts to make MAIS less opaque will prove crucial for making their outputs useful to doctors when reaching diagnoses or making treatment recommendations. The more physicians are aware of what features MAIS are likely to not be including in their analyses, the more they will be able to supplement these programs' outputs with judgments about the inductive risks involved. Until this happens, the potential harm of MAIS overlooking the risks to patients' interests associated with each of the relevant possible diagnoses and/or treatments will simply be too high to justify physicians' reliance on their outputs when reaching diagnoses or making treatment recommendations.

V. CONCLUSION

We have argued that the possibility of current MAIS empathizing with patients is significantly limited by the fact that, unlike doctors, MAIS do not have the option of gaining first-person experiences of having human interests. Moreover, even if MAIS were to receive descriptions of patients' experiences as inputs, it is difficult to see by what means they could adequately understand information about patients' interests from such first-person accounts. We have also argued that MAIS are likely to have difficulties – due to this lack of empathy – in their performance of 'abduction' and 'prioritization' when performing the tasks of reaching a diagnosis or making a treatment recommendation. In both cases, we are claiming that MAIS' inability to adequately understand human interests due to their lack of empathy will prevent them from being able to make appropriate calculations of the relevant risks to patient interests that arise when making the inductive leap from the given evidence to conclusions about patients' medical conditions and the treatments those patients should receive. In particular, we have focused on the risks to patients' interests associated with missing other possible diagnoses and the risks of overlooking other possible treatments.

Given the extent of these limitations arising from MAIS being unable to empathetically understand patients' interests, we do not recommend the development of MAIS for more general diagnostic tasks because of the way that the inductive risks of missing diagnoses will significantly rise with each increase in the potential conditions that the MAIS is required to choose between. Furthermore, we argue that the outputs of MAIS should be closely supervised by physicians, which speaks against the development of this technology for providing diagnoses on which physicians can rely and against its direct use by patients (e.g., via apps on smartphones). Until MAIS is able to develop the empathetic understanding needed to properly account for the inductive risks of reaching a diagnosis or making a recommendation for treatment, physicians must reach their own considered judgements in order to ensure that the outputs of MAIS do not pose unjustifiable risks to patients' interests.

REFERENCES

Bailey, O. 2022. Empathy and the value of humane understanding. *Philosophy and Phenomenological Research* 104(1): 50-65.

Balogh, E., B. Miller, and J. Ball. 2015. *Improving Diagnosis in Health Care*. Washington: The National Academies Press.

Baron, M. 1995. *Kantian Ethics Almost without Apology*. Ithaca: Cornell University Press.

Barragán-Montero et al. 2021. Artificial intelligence and machine learning for medical imaging: A technology review. *Physica Media* 83: 242-56.

Barwise, A. and B. Pickering. 2022. The AI needed for ethical decision making does not exist. *The American Journal of Bioethics* 22(7): 46-9.

Baxley, A. M. 2010. *Kant's Theory of Virtue: The Value of Autocracy*. Cambridge: Cambridge University Press.

Blasimme, A., and E. Vayena. 2020. The ethics of AI in biomedical research, patient care and public health. In *Oxford Handbook of Ethics of Artificial Intelligence*, Ed. M. Dubber et al., 703-18, Oxford: Oxford University Press.

Bloom, P. 2016. *Against Empathy: The Case for Rational Compassion.* New York: HarperCollins.

Briganti, G. and O. Le Moine. 2020. Artificial intelligence in medicine: Today and tomorrow. *Frontiers in Medicine* 7(27): 1-6.

Cadario, R., C. Longoni, and C. Morewedge. 2021. Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour* 5: 1636-42.

Coplan, A., and P. Goldie, eds. 2011. *Empathy: Philosophical and Psychological Perspectives.* Oxford: Oxford University Press.

Daniels, N. 1996. *Justice and Justification: Reflective Equilibrium in Theory and Practice*. Cambridge: Cambridge University Press.

Elfiky, A. A., M. J. Pany, R.B. Parikh, and Z. Obermeyer. 2018. Development and application of a machine learning approach to assess short-term mortality risk among patients with cancer starting chemotherapy. *JAMA Network Open* 1(3): e180926.

Fernandez-Esperrach et al. 2016. Exploring the clinical potential of an automatic colonic polyp detection method based on the creation of energy maps. *Endoscopy* 48: 837-42.

Goldie, P. 2011. Anti-empathy. In *Empathy: Philosophical and Psychological Perspectives*, eds. A. Coplan and P. Goldie, 302–17. Oxford: Oxford University Press.

Goldman, A. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Goldman, A. 2011. Two routes to empathy: Insights from cognitive neuroscience. In *Empathy: Philosophical and Psychological Perspectives,* eds. A. Coplan and P. Goldie, 31–44.

Oxford: Oxford University Press.

Gorovitz, S. 1988. Good doctors. In *Ethical Issues in Professional Life*, ed. J. Callahan, 424-43. Oxford: Oxford University.

Grzybowski, A., P. Brona, G. Lim, P. Ruamviboonsuk, G. S. Tan, M. Abramoff, and D. S. Ting. 2020. Artificial intelligence for diabetic retinopathy screening: a review. *Eye* 34(3): 451-60.

Hoffman, M. 2000. *Empathy and Moral Development: The Implications for Caring and Justice.* Cambridge: Cambridge University Press.

Hoffman, M. 2011. Empathy, justice, and the law. In *Empathy: Philosophical and Psychological Perspectives,* eds. A. Coplan and P. Goldie, 230–54. Oxford: Oxford University Press.

Holzinger et al. 2019. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery* 9(4): 1-13.

Kant, I. 1999. *Groundwork of the Metaphysics of Morals*. In *Practical Philosophy*. Cambridge: Cambridge University Press.

Kant, I. 1999. *The Metaphysics of Morals.* In *Practical Philosophy*. Cambridge: Cambridge University Press.

Ledley, R. and L. Lusted. 1959. Reasoning foundations of medical diagnosis. *Science* 130(3366): 9-21.

Ma et al. 2010. Speculative abductive reasoning for hierarchical agent systems. *Computational Logic in Multi-Agent Systems: 11th International Workshop* 11: 49-64.

Macartney, F. 1987. Diagnostic logic. *British Medical Journal* 295(6609): 1325-31.

Marr, B. 2017, Jan. 20. First FDA approval for clinical cloud-based deep learning in healthcare. *Forbes*.

Masto, M. 2015. Empathy and its role in morality. *The Southern Journal of Philosophy* 53(1): 74–96.

Meier et al. 2022. Algorithms for ethical decision-making in the clinic: A proof of concept. *The American Journal of Bioethics* 22(7): 4-20.

Moons, K. and D. Grobbee. 2002. Diagnostic studies as multivariable, prediction research. *Journal of Epidemiology and Community Health* 56(5): 337-8.

Moor, M. et al. 2023. Foundation models for generalist medical artificial intelligence. *Nature* 616(1): 259-265.

Niel, O. and C. Boussard. 2018. Artificial intelligence can predict GFR decline during the course of ADPKD. *American Journal of Kidney Diseases* 71: 911-12.

Paul, L. A. 2021. The Paradox of Empathy. *Episteme* 18(3): 347-66.

Picard, R. W. 2000. *Affective Computing*. Cambridge: MIT press.

Poplin, R. et al. 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering* 2(3): 158-64.

Prinz, J. 2011a. Is empathy necessary for morality? In *Empathy: Philosophical and Psychological Perspectives*, eds. A. Coplan and P. Goldie, 211–29. Oxford: Oxford University Press.

Prinz, J. 2011b. Against empathy. *The Southern Journal of Philosophy* 49: 214–33.

Sherman, N. 1997. *Making a Necessity of Virtue: Aristotle and Kant on Virtue*. Cambridge University Press.

Shoemaker, D. 2014. Emotional lobbying. *Georgetown Journal of Law & Public Policy* 12: 505-20.

Smith, J. 2017. What is empathy for? *Synthese* 194: 709–22.

Sparrow, R. and J. Hatherley. 2020. High hopes for ‘Deep Medicine’? AI, economics, and the future of care. *Hasting Center Report*: 14-7.

Stueber, K. 2006. *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences.* Cambridge: MIT Press.

Toombs, S. K. 2001. The role of empathy in clinical practice. *Journal of Consciousness Studies* 8(5-6): 247-58.

Topol, E. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books.

Torrance, S. and R. Chrisley. 2015. Modelling consciousness-dependent expertise in machine medical moral agents. In *Machine Medical Ethics*, eds. S.P. van Rysewyk and M. Pontier, 291-316. New York: Springer.

US Food and Drug Administration. 2018a, January 12. De Novo Classification Request for IDx-DR. New Regulation Number: 21 CFR 886.1100. Available: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf (accessed November 10, 2022)

US Food and Drug Administration. 2018b, May 24. FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures. Available: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-algorithm-aiding-providers-detecting-wrist-fractures (accessed November 16, 2022).

Vallverdú, J., and D. Casacuberta. 2015. Ethical and technical aspects of emotions to create empathy in medical machines. In *Machine Medical Ethics*, eds. S.P. van Rysewyk and M. Pontier, 341-62. New York: Springer.

NOTES

[1] See, for instance, Toombs, 2001; Torrance and Chrisley, 2015.

[2] See US FDA, 2018b.

[3] See US FDA, 2018a.

[4] See Poplin et al., 2018.

[5] See Niel and Boussard, 2018.

[6] See Fernandez-Esperrach et al., 2016.

[7] See Elfiky et al., 2018.

[8] See also Grzybowski et al., 2020.

[9] Briganti and Le Moine note that “opposing AI and clinicians is, although well represented in the scientific literature, probably not the best way to tackle the issue of performance in medical expertise: several studies are now approaching the interaction between clinicians and algorithms as the combination of human and artificial intelligence outperforms either alone” (Briganti and Le Moine, 2020, 3).

[10] In his book *Dark Medicine*, Eric Topol claims that “the greatest opportunity offered by AI” in medicine is that it would give doctors “more time to come together” with their patients (Topol, 2019, 18). However, as Sparrow and Hatherley point out, “the economics of health care, especially where it is provided in a for-profit context, will dictate that any time savings made possible by a reduction in the administrative burdens on physicians in the course of patient consultations will be used to move more patients through the system rather than to allow practitioners to spend more time talking with, and caring for, their patients” (Sparrow and Hatherley, 2020, 14).

[11] In their recent "Foundation Models for Generalist Medical Artificial Intelligence," Moor et al. claim that "GMAI-based applications will be deployed both in clinical settings and on remote devices such as smartphones, and we predict that they will be useful to diverse audiences, enabling both clinician-facing and patient-facing applications" (Moor et al., 2023, 264).

[12] We certainly acknowledge the distinction between reaching a diagnosis and making treatment recommendations; this is because any diagnosis is going to leave open questions about what is the most appropriate treatment to pursue or whether medical treatment should be pursued at all. However, we also want to emphasize the way that medical diagnoses are inextricably tied up with treatment decisions. This is because reaching a diagnosis leads medical practitioners to restrict what options for treatment would be appropriate to consider. As Moons and Grobbee write, "to set a diagnosis is fundamental in medical care as it offers an indication of the patient's prognosis and directs therapeutic management" (Moons and Grobbee, 2002, 337). Given this, the risks of misdiagnosis are so significant because of the impact that these types of mistakes are likely to have on decisions about treatment for the patient.

[13] Marcia Baron has argued that this is particularly relevant when it comes to figuring out what our contextual *imperfect* obligations are. Unlike perfect duties that yield token obligations, imperfect duties are duties of commission and thus rely on specific cases to be applicable. According to Baron, sympathetic feelings help us "to do things that it is generally impossible, because of the nature of imperfect duties, to do from duty alone" (Baron 1995, 220).

[14] Especially in the 18th century, philosophers' discussion of 'sympathy' focuses on the features that in the contemporary philosophical literature are attributed to 'empathy'. For a discussion of this, see Bailey 2022.

[15] As Bailey points out, the claim that the empathizer's affective experience is one of genuine emotion is widely – though not universally – endorsed (see e.g., Prinz 2011b, 215).

[16] The term 'empathy' has been used in many different ways (Coplan and Goldie, 2011). Alvin Goldman (Goldman, 2006; Goldman, 2011) and others (e.g., Stueber, 2006; Goldie, 2011) have distinguished two species of empathy. 'Lower-level empathy' (in Goldman's terminology) or 'basic empathy' (in Stueber's) is generally taken to be a nonconscious and automatic coming to feel as another feels because another feels that way. Higher-level empathy on the other hand is not automatic but requires some kind of cognitive effort on the part of the empathizer. Like lower-level empathy, successful higher-level empathy generates a similarity in affect between the subject and the target. In addition, these higher-order imaginative exercises, when successful, enable the empathizer to form relevant beliefs about the target's mental states. For a defense of the claim that non-conscious processes like emotional contagion ought not qualify as genuine empathy, see Masto, 2015.

[17] This distinction between a 'mere receptivity' and a 'higher-order empathy' is important for understanding how such a Kantian empathy would differ from a conception of 'empathy' that would require the uncritical acceptance of what the other person believes. That is to say, though it will be important to acknowledge whatever biases the other person holds when understanding that person's perspective, uncritically taking on those biases in one's reconstruction of that other person's perspective would be a form of 'mere receptivity'. In contrast, Kantian 'higher-order empathy' will involve these bias-corrective elements that avoid having to accept other people's mistaken judgments (assuming that these are, in fact, mistaken) when reaching an empathetic understanding of their perspectives. For instance, Kant's form of 'higher-order empathy' would allow one to form an empathetic understanding of a child's fear that there is a monster under her

bed (i.e., one could relate to having formed such irrational beliefs as a child oneself, recognize why believing this to be true would make one afraid, and reach an understanding of how the child is feeling) without having to accept the child's mistaken beliefs.

[18] There is a philosophical question here as to what exactly makes one's body radically different than a human body, but we will not explore this issue here.

[19] Picard writes that "for the most part, the criteria of human emotional intelligence apply directly to computers and to human-computers interaction. One place where they might differ, however, is in the experience of emotions required for empathy. Humans can approximate the emotional experience of each other because we have similar brains and bodies [..] The fact that humans and machines differ in physiology and conscious awareness will cause them to have different emotional experiences. Hence, we cannot expect a machine to really feel what we feel, even if it 'has' emotions. Consequently, the best empathy or understanding we can hope for is at the level of an outsider, who tries to understand, but who has never actually been in our shoes, so to speak" (Picard, 1997, 80). See also Torrance and Chrisley 2015, 303.

[20] As Meier et al. note, "the way in which deep neural networks and FCMs process their inputs by passing activations on through weighted connections is very similar," but like the symbolic methods of LLMs "each node in an FCM has a human-designated semantic meaning" (Meier et al., 2022, 7).

[21] For more about the importance of 'abductive reasoning' to scientific explainability, see Holzinger et al., 2019 and Ma et al., 2010.

[22] Macartney writes about the way that physicians cannot start to focus in on the relevant diagnoses "without a fairly thorough knowledge of the territory to be searched. Students need to familiarize themselves with the normal as well as the abnormal" (Macartney, 1987, 1326).

[23] One can see this type of connection between patients' interests and what doctors identify as symptoms of disease within Norman Daniels' account of diseases as "*deviations from the natural functional organization of a typical member of a species*" (Daniels, 1996, 185) As Daniels writes, "for humans we require an account of the species-typical functions that permit us to pursue biological goals as social animals" (Daniels, 1996, 185); but this will also include a recognition of the "*normal opportunity range* for a given society," which is "the array of 'life plans' reasonable persons in it are likely to construct for themselves" (Daniels, 1996, 187).

[24] Balogh, Miller, and Ball note that "typically, clinicians will consider more than one diagnostic hypothesis or possibility as an explanation of the patient's symptoms and will refine this list as further information is obtained in the diagnostic process" (Balogh, Miller, and Ball, 2015, 32). Ledley and Lusted also summarize this part of the diagnostic process nicely, writing that physicians will "list all the diseases which the specific case can reasonably resemble. Then [physicians] exclude one disease after another from the list until it becomes apparent that the case can be fitted into a definite disease category, or that it may be one of several possible diseases, or else that its exact nature cannot be determined" (Ledley and Lusted, 1959, 9).

[25] For this particular aspect of the diagnostic process, Balogh, Miller, and Ball note that "an overarching question throughout the process is whether sufficient information has been collected to make a diagnosis" (Balogh, Miller, and Ball, 2015, 48).

[26] Balogh, Miller, and Ball write that 'when considering invasive or risky diagnostic testing or treatment options, the diagnostic verification step is particularly important so that a patient is not exposed to these risks without a reasonable chance that the testing or treatment options will be informative and will likely improve patient outcomes" (Balogh, Miller, and Ball, 2015, 34-5).

[27] This need to recognize when patients are likely to have reasonable differences in their interests is captured well by Norman Daniels' discussion of "effective opportunity." As Daniels notes, the "effective opportunity range" will be determined "from the perspective of an individual who has a particular plan of life and who has developed certain skills accordingly" (Daniels, 1996, 215).

[28] On the other hand, our claim does align with views such as the one presented by Barragán-Montero et al., who emphasize the need "to increase the interpretability of the results, which is one of the well-acknowledged limitations of the current ML/DL methods" (Barragán-Montero et al., 2021, 251).