Black-box AI in Medicine:
A Standard of Care Without Interpretability

Abhishek Mishra

Supervised by:
Tom Douglas, Julian Savulescu, and Jonathan Herring

Funded by:
Wellcome Trust

Table of Contents

01 Introduction

Artificial intelligence (AI) systems, especially deep learning models consisting of many hidden layers, have gained widespread attention over the last decade for their seemingly sudden jump in performance. This attention has been followed by widespread experimentation with such models, and ultimately mainstream application, across almost every domain of human effort. They have been used across the sciences, across industry, in medicine, law, finance, and even across the humanities and arts. Their uses have further ranged from being a decision-support systems, an aid to human operators, and autonomous recommender systems across retail, social and traditional media. Over the past decade, there is scarcely an area of human activity that has not seen their uptake, and sometimes even dominance. Medicine has been one domain where their performance, especially in evaluating medical imaging, has seen substantial improvement in the past decade.

This thesis will examine the extent to which the use of such deep learning models in medicine, due to their black-box nature, poses novel ethical and epistemic problems. It will ultimately argue that such problems are either not present, or can be overcome – contrary to the more widespread discussion in the literature (and in policymaking). This introduction will lay the groundwork for this analysis by identifying and discussing some key concepts and themes, and articulating the broad argumentative strategy of the thesis.

01.1 Context for Discussion

This section will briefly lay out some of the background knowledge, and the key concepts that underpin the discussion throughout this thesis.

01.1.1 Deep-learning models in medicine

As mentioned, deep learning models have seen widespread development and adoption in medicine and healthcare more broadly, across diagnosis, treatment, and care management. Proof-of-concept models have been developed for radiological analysis (such as for lung cancer screening (Ardila *et al.* 2019)), to diagnose diabetic retinopathy through optical computed tomography scans (Gulshan *et al.* 2016), diagnosing glaucoma (Bojikian *et al.* 2019), classifying skin cancer (Esteva *et al.* 2017), assessing cardiovascular risk factors (Poplin *et al.* 2018), diagnosing a range of pediatric diseases through electronic health records (Liang *et al.* 2019) (among many other use-cases), and predicting risk of suicide attempts (Walsh *et al.* 2017). Outside of being used as clinical decision-support systems, AI can also be used to provide 'ambient intelligence' in hospital settings (Gerke *et al.* 2020), monitoring for patient mobilization (Yeung *et al.* 2019) and bedside practices such as hand hygiene (Haque *et al.* 2017). The use of AI systems has also been suggested to protect and promote

public health in a more targeted and precise way (Panch *et al.* 2019). AI systems can also be used to yield better outcomes for organ transplantation (e.g. kidney or liver transplants), from assisting in curating transplant chains for incompatible donor-recipient pairs (Sherman, Greenbaum 2019) to predicting survival post-transplantation (Díez-Sanmartín *et al.* 2020). In surgery, deep learning models can be used to assess surgical procedure and performance (Khalid *et al.* 2020) as well as for surgical robotics (Mirnezame, Ahmed 2018). Machine learning models used for radiology are performing at or beyond the level of the average radiologist, matched perhaps only by the most senior and discerning of medical operators (so far) (Liu et al, 2019).

As can be seen, the applications of AI in health and medicine are myriad. Furthermore, while most of the applications now are of an assistive nature, providing aid to clinicians and human operators, many AI applications in the future may instead be autonomous. As our understanding of the best model architectures, our access to medical (and non-medical but medically applicable) data, and their access to compute grows, the performance and ubiquity of such deep learning models will similarly increase.

01.1.2 Black-box nature of deep-learning models

Their uptake, however, has not been without concern. Most of the noteworthy advances in AI for health in the last few years have been driven by deep-learning models – these are models which have multiple 'hidden layers', where as information passes through these different layers, it is cumulatively better 'understood' by the model. A simple example, for instance, concerns an image being processed by a deep learning model (say a convolutional neural network, or CNN) – the first few layers might focus on the image at a pixel level, the next middle few in understanding middle-level complexity such as vertices, contrasts, and finally the last few layers piecing it all together to judge what the image is exactly. Part of the gold rush in medical imaging, for instance, has been due to the success of such CNNs (Krizhevsky et al. 2017).

Such model architectures can be extremely powerful – unfortunately, they have a 'black-box' nature, such that it is hard to understand the decision-making logic that has led to a particular recommendation (say a particular label for an image that the model has been asked to classify). This 'black-box' nature exists because of two key issues that lie in tension: (1) the high dimensions (or variables) that the model is tracking, and ultimately making a decision based-on, for something like an image – for instance, the various pixels, certain combinations of pixels, etc. (where the number of such variables can range from hundreds to now trillions of variables, and (2) the relatively limited capacity of a human mind to understand such high-dimensional decision-logic, given that we can't appreciably track that number of variables and their interactions. The opacity of the model is thus an intrinsic feature that can't be done away with without fundamentally changing the architecture. And unfortunately, the aspect of deep learning that has made it so successful in the last few years, and led to a resurgence of AI, has been this ability to operate over numerous hidden layers, tracking and reasoning across a multitude of dimensions. This produces one of the key tensions that has characterized almost all conversation around risks from, and the ethics of, artificial intelligence: *the very feature that makes them more powerful than other models also makes them less understandable*

01.1.3 Opening up the black-box

It is important that we understand the decision-making logic of agents that we interact with – or so the argument goes. Our inability to do so may cost us a lot. In the case of AI, it is argued that their black-box nature can undermine operators' ability to use them well.

This 'black-box' nature of deep learning models has been seen to be especially problematic in healthcare (Castelvecchi, 2016), and this has been reflected in the development and testing of models which endeavor to be explainable (DeFauw *et al.* 2018, Rajkomar *et al.* 2018) – explanations are increasingly being called for (Rajkomar *et al.* 2018, Topol 2019). Doctors may not be able to determine the extent to which they should rely on the model's recommendation if they cannot access and assess the reasons why the model made the recommendation in the first place. Relying on a model when its recommendation is wrong, or not relying on it when its recommendation is right, can have deep impacts on patients. The opacity of deep learning models therefore raises epistemic issues, which may lead to harmful patient outcomes. It is also claimed to raise ethical issues. In medicine, it may compromise shared decision-making, if doctors aren't able to communicate to patients why certain diagnostic or treatment recommendations are being made to them. It may obscure the ways in which such models are value-laden in their reasoning, if this reasoning cannot be inspected for which values are being prioritized for any given decision. Use of deep learning models may further bias and discrimination for similar reasons. And it may also frustrate our ability to adjudicate responsibility for harmful outcomes, and otherwise use them in accountable ways.

Efforts to counter this 'black-box' nature have focused on two families of approaches: explainability and interpretability. Explainability efforts focus on applying post-hoc methods to the model to analyze it to get insights around what the decision-making process used was. This way of 'opening up the black-box' can aim to provide insights about the operation of the deep learning model generally (global explainability), or about specific classifications that the model has made (local explainability). Global explanations such as partial dependence plots of feature interaction methods can, for instance, focus on understanding model outputs by identifying the average impact of a particular feature (say a particular test value from an EHR chart) on the classification (a disease code), or the relationship between features – this can assist in understanding the average behaviour of the model. Local explanation methods such as surrogate models (LIME) (Ribeiro et al. 2016) or counterfactual explanations (Wachter et al. 2017), on the other hand, help in explaining specific decisions made by the model by attributing it to the relevant features in that case. Some of the most-used explainability approaches, such as LIME, rely on secondary models to approximate the reasoning of the deep-learning model – the explanations then, are in some sense, the 'best guesses' of these secondary models on what the decision-making approach used by the deep-learning model is.

Critics of explainable approaches often argue that such explanations are not reliable, and can be misleading (Rudin 2019). They argue for interpretable approaches, which focus instead on restricting the complexity of the model. Recall that the initial opacity of deep-learning models occurs due to the tension between the high-dimensionality of the model's reasoning, and the relatively lower-dimensionality that humans are able to track and engage with when it comes to explanations. Interpretable approaches focus on reducing the complexity of the former to bring it more in line with the latter. Thus, interpretable models, such as linear or logistic regression models or decision trees, are alternative to deep-learning models that are advocated for because (1) they arguably provide equivalent overall performance – a contested claim in the field – and (2) we can understand the *exact* decision-logic used by the model (as compared to a best-guess one), which offers a host of epistemic and ethical advantages. Generally however, such models are seen as offering lower accuracy, and so there can be a trade-off between interpretability and accuracy (London 2019).

Such arguments are made across domains of AI application, but gain special importance for high-stakes decisions like those made in medicine. These arguments have been repeated by many commentators, for many years, to the point where the importance of understanding the

decisions of deep learning models, and the risks inherent in not doing so, are taken almost as theoretical dogma. As a result of this, across academia as well as industry, there has been an inescapable push towards 'explainability' as a panacea.

01.1.4 The interpretability – performance trade-off

This would be perfectly acceptable, if not for two things. First, certain ways of promoting explainability, such as doing away with deep learning models and their high-dimensional architecture entirely and instead using interpretable models, may have performance costs. As has already been mentioned, it is this very high-dimensional architecture that has given such models an edge as far as accuracy is concerned, and it is improbable that models which are interpretable because their complexity has been restricted would perform at the same level, especially for some tasks such as image classification. This point is sometimes contested by commentators, but it is at the very least a live debate (Rudin, 2019). We need to be open to the possibility that there might be a very real trade-off between accuracy and interpretability. The likelihood of there being such a trade-off rises further when we consider the developments in store in the future. Interpretable models built to predict certain events or phenomena rely to a large extent on reflecting our human understanding of that phenomena into model design choices. To insist that interpretable models will be able to go toe-to-toe in performance with deep learning models is to insist that our understanding of the phenomena we wish to model will scale as quickly as our access to compute and data, and our ability to innovate on model architectures to eke out even more performance gains. This is a hard proposition to defend, especially in light of the very skewed scaling we've seen in the last decade (though that period may have been an outlier) (Kaplan et al. 2020). Regardless, especially in high-stakes domains like medicine, it would be even more crucial that we don't leave performance gains on the table.

Second, even if we were not trying to access explainability by doing away with deep learning models entirely, but were instead using exogenous techniques to guess at the decision-making logic of the model, an uncritical approach to explanations as panacea would not help with that. It is important to know what sorts of explanations are important based on an understanding of if/why they are needed. Scrutinising the epistemic and ethical concerns that are argued to result from a lack of explanations can tell us about why explanations are important, what do they need to capture, and whether there might exist alternatives to them that might serve the same function.

01.2 Thesis Argument and Structure

The goal of this dissertation then, is to push back against this uncritical problematization of opacity, and the seemingly dogmatic acceptance of explainability as a necessary solution. Through this thesis, I will argue that explanations are not necessary to address the ethical and epistemic concerns raised by the use of deep-learning models in medicine. I will do so by

(1) Identifying some of the key ethical and epistemic issues that are argued to arise from the use of deep-learning models in medicine, and are frequently raised as objections to their use – where these objections are used to motivate a need for explanations through the use of interpretable models (Chapters 2 and 3) – and then,

(2) Defusing *three* of the strongest objections, either by arguing that the problematization is misguided and no ethical concern arises at all, or by showing that even if such concerns arise explanations are not *necessary* for their mitigation (Chapters 4-6).

The argument will thus roughly be one of showing that explanations, through the use of interpretable models (and thus the eschewal of 'opaque' models), are not necessary by showing reasonable alternatives. While this thesis will survey a broad array of ethical and epistemic concerns raised around the use of black-box deep learning models in medicine (Chapter 2 and 3), it will select the three problems of adequate epistemic reliance on these models (Chapter 4 - 'How can you know whether to defer to a model if you don't know why it made its recommendation?'), shared decision-making (Chapter 5 - 'How can decision-making be shared and patient-centric if patients can't know why models are making their recommendations?'), and accountability (Chapter 6 - 'How can we have accountable use of AI if we don't understand the model's reasoning?') as major ones. The discussion of each of these three problems will serve as 'case-studies' of sorts - by showing that each of these problems can be defused even in the case of deep-learning models, and thus that good clinical and ethical outcomes can still obtain, this thesis will gesture towards the possibility of an acceptable standard of care even with the use of black-box models. As discussed in the next sub-section, further concerns and objections to the use of black-box, deep-learning models in medicine will still remain – it is beyond the scope of this project to treat each possible objection individually. The strategy (and hope) of this work is that by showing how we might defuse three of the key objections, we might realise the possibility of (and inspire confidence in) justifiable use of black-box models.

01.3 Limitations and Next Steps

This dissertation is merely one part of a larger effort to push back against the argued inexorability of explanations for AI. Much further work remains to be done. Even if the analyses conducted here, and the arguments put forward, are successful, there are still some limitations of the work. First, the above-mentioned epistemic and ethical concerns don't cover all the use-cases where explanations are argued to add value. The ability to understand the decision-making logic of AI systems is also useful to better understand the phenomena that the system is trying to model, or relatedly to better understand how to build better systems. These two major use-cases are not discussed, as are potentially other use-cases where explanations of the model's decisions would be beneficial, and a lack of them harmful. Second, there is still further work to be done even in polishing and building out some of the arguments put forward here. The work in chapter 4, for instance, merely starts on the analysis on how we might build out alternative heuristics for reliance on AI systems if their decision-making logic is not accessible. It does not provide a final solution, but makes progress to it by suggesting an intermediate solution (a modified Bayesian approach to deference) that needs further iterations based on conditions and limitations that have been identified. Further, this is just a theoretical proposal, and before it could be considered a contender for an alternative standard of care (even after it has been theoretically completed), it needs to be empirically validated.

Even so, I believe that the analysis in this dissertation is a robust contribution to the larger project. An update to how we think about the opacity of deep learning models, and even a small walk-back on their disqualification for medical use due to epistemic and ethical reasons can have a large impact on patient outcomes, so long as they are robustly tested and safely used in line with some of the analysis here. It can also have an impact on how we think about adjudicating legal responsibility, through an update to determining medical negligence in the use of AI systems.

I started this project with strong convictions about opacity not being a defeator to the use of deep learning models in medicine. Over the course of my investigations, these convictions have been softened by understanding the realities of patient needs, the difficulties of forming alternatives to explanatory approaches, and the practicalities of deploying such models in health

systems. However, at the end of this journey, I still believe (though perhaps with less radical intent) that the problematization of opacity has been overstated, and that there remains a promising path forward for the epistemically and ethically appropriate use of deep learning AI models.

02. Ethics of Medical AI – An Overview

As discussed in the previous section (Section 1 – Introduction), this chapter will outline some of the key ethical and epistemic issues that arise with the use of deep-learning models in medicine, and more broadly in healthcare. The goal of this survey is to understand the wider array of concerns that have arisen over the use of such models. In subsequent chapters, some of these specific concerns will be individually discussed.

The potential benefit of deep-learning models in medicine is vast – they can lead to better clinical and health outcomes not just because they perform at a superhuman level for some tasks, but also because they may facilitate greater access to healthcare in populations and low-resource settings in which previously healthcare access was limited. Beyond purely medical outcomes, the use of such systems also has the potential to affirm the values we consider important in the healthcare context: enabling greater informed consent and patient-centered medicine, facilitating a more just distribution of healthcare resources, and leading to more unbiased clinical decision-making. However, given the unique features of AI models (their opacity arising from their architecture, as discussed in section 1.1.2), their use might lead to new ethical concerns that undermine these very values in unprecedented ways.

It is important at the outset to highlight some areas we will not be focusing on. Our discussion shall not focus on the privacy and ethical issues around the data that such machine learning models are built on, on health applications and the specific concerns they raise as compared to more traditional channels of health information and advice, or on AI models used outside of healthcare settings that have health implications (such as algorithms used in social media). We will also not discuss models that are used within healthcare settings (like hospitals) but to augment logistical workflows. We restrict our discussion here for reasons to do with space as well as focus. Furthermore, while most of the applications of such models now are of an assistive nature, through providing aid to clinicians and human operators, many AI applications in the future may instead be autonomous. This distinction is an important one and raises very different ethical questions. For the purposes of most of this chapter, we shall focus on such assistive systems, though we will return to the question of ethical issues concerning autonomous systems towards the end.

In this chapter then, we shall consider the main ethical themes that are raised by the use of such AI systems. We will consider in turn the ethical concerns that arise on the basis of (1) bias, discrimination, and fairness, (2) patient-centred medicine, (3) AI and value-based decision-making, (4) responsibility, accountability, and explanation, and (5) broader and long-term societal effects of AI in healthcare, including more speculative ones. In each of the following sections, we will raise the various questions posed under these themes, assemble the main arguments that have been provided, and point to directions for future research. Finally, we shall end with some closing comments.

02.1 Bias, Discrimination, and Fairness

The potential for AI systems to perpetuate bias that is ethically problematic is one of the most well-discussed topics in the domain of AI ethics. Technically, AI systems are intended to discriminate by design – a deep learning classifier trained to detect lung cancer through the analysis of CT scans is meant to discriminate between those scans that sufficiently indicate a positive diagnosis and those that do not. What makes discrimination ethically problematic is when it happens on the basis of factors that are not clinically relevant and so should not have been a basis for assessment, to yield outcomes that are unfair. Such situations are especially problematic when such factors have traditionally been protected categories such as race or gender, though unethical discrimination can also arise in cases in which non-protected categories are used when they shouldn't be. When we write about a factor that *shouldn't* have been used for assessment (independent of such factors being protected), we refer to factors that have not reasonably been shown to affect the outcome in question. A process that is problematically discriminatory in such a way, and yields unfair outcomes, can be said to exhibit bias. Such biased processes can then be argued to violate considerations of justice and equality.[1]

Algorithms have been noted to be problematically biased across many different applications. Algorithmic bias along the lines of race seem particularly foreboding for the healthcare context, considering that there are deep learning models being developed for the classification of skin cancer using photographed images (Esteva *et al.* 2017) in which fewer than 5% of the training images used were of dark-skinned individuals (Zou, Schiebinger 2018). Perhaps the most significant example yet of such discrimination in healthcare is of an algorithm used to provide patient risk scores as the basis for selection in 'high-risk care management' programs: for a given risk score, black patients were considerably sicker than white patients (Obermeyer *et al.* 2019). Selection based on these scores led to a significantly lower representation of black patients in such care management programs across levels of sickness. Given that such an algorithm is deployed nationwide in the United States, and is typical of commercial risk-prediction tools that are applied to roughly 200 million people in that country each year, the scope of discrimination is staggering.

The use of AI systems can yield such discriminatory outcomes due to various potential problems in their development. For instance, while it is unlikely that models will be built for which protected categories like race are features meant to be explicitly factored into the classification, other seemingly unconnected features can serve as proxies for them. This is what happened in the risk prediction case above: the model took as input total past medical expenditure for patients and used it to predict future health expenditure, which was taken to be an indicator of future health risk (Obermeyer *et al.* 2019). Such 'label bias' occurs when the data label used ('medical expenditure') does not mean the same thing for all patients 'because it is an imperfect proxy [for future healthcare need] that is subject to health care disparities rather than an adjudicated truth' (Rajkomar *et al.* 2018). Unequal medical expenditure could arise due to factors such as differential access to care, or differing levels of trust in the healthcare system rather than different risk.

Avoiding such iatrogenic effects requires a deep understanding of the systems that produce the data that AI models are trained on, and how there might be systematic discrimination occurring. Other biases that could lead to such discrimination include 'minority bias', when protected groups are underrepresented in the training data (as in the case of the dermatological classifier above), 'missing data bias', when data is missing for protected groups in a non-random fashion, and 'informativeness bias', when chosen features are less informative to render a prediction in a

[1] We will return to the question of what constitutes bias or unfair discrimination at the end of the section.

protected group (Rajkomar *et al.* 2018). Ensuring that unfair discrimination doesn't occur requires that all these (and more) biases be considered when building AI systems.

But how do we determine if a model is discriminating unfairly? When it comes to measures of fairness, there are multiple viable candidates. Earlier we noted that ethically problematic discrimination can happen when a model's output is based on protected attributes (like race, gender) or their proxies – this is known as *anti-classification* (Corbett-Davies, Goel, 2018). Other candidates for operational measures of fairness are *classification parity*, which requires common measures of predictive performance (false positive and false negative rates) to be equal across groups, and *equal calibration*, which requires that the for a given model output (such as a particular health risk score) it is equally likely that the output is correct regardless of which group the individual is from. Contextualized to the care management program risk prediction case earlier, anti-classification requires that risk scores are not calculated by explicitly factoring in race or any of its clinically irrelevant proxies (which can easily happen, given that medical expenditure *is* correlated with race, for instance). Classification parity requires that false positive and false negative rates are equal across white and black patients, and equal calibration requires that for a given risk score the likelihood of needing high-risk care management remains the same regardless of whether the patient is white or black.

Considering that these measures all seem intuitively to capture different parts of various conceptions of fairness, the question of which of these measures to use becomes an important one for AI systems. Furthermore, lest we think that we might just be able to use them all in some aggregated fashion, there is a further twist – classification parity and equal calibration are incompatible with each other and cannot be fulfilled simultaneously, and so must be 'traded-off' against each other (Kleinberg et al. 2016). Given that different groups have different base rates of the property being classified (e.g. high-risk care needs, measured perhaps by a comorbidity score), a model that is calibrated would *necessarily* produce different false positive and false negative rates. Such conditions necessitate a choice between classification parity and equal calibration as to which is the more desirable measure of fairness. While there have been some arguments put forward in favour of selecting one of these measures to the exclusion of the other (Long 2020), this debate is still quite new and more work needs to be done.

So far, we have been assuming that the various protected categories mentioned here are actually clinically irrelevant when it comes to medical decision-making. However, there might be circumstances in which such features have been reasonably shown through research to be clinically relevant. For instance, it has been shown that the base rate of a disease is different across sub-populations (e.g. prevalence of breast cancer in men vs. women). Furthermore, there are other complicating factors. For example, we might imagine cases in which a model might not display fairness, but might otherwise yield significantly better performance from existing methods. Whether such a model should nonetheless be used would depend on which theory of justice one subscribes to, and how such a theory balances total good (measured in terms of aggregate health outcomes for instance) and fairness.[2] Additional complications might also arise when we consider whether the unfair outcome disadvantages those who are already well-off in society, or those who are worse off (see Herzog).

Another thing to note is that so far, the measures of fairness discussed are measures of model performance. Ultimately, the ideal is to aim for fair *outcomes*, something that fair model performance (however that might be measured) doesn't necessitate (Rajkomar *et al.* 2018). Fair model

[2] To err on the side of stating the obvious, it is always important to be as sure as possible that under these conditions the unfair behaviour is actually because of differing base rates rather than because of historically different treatment that feeds back into the data we use to build the model.

performance might get disturbed by human bias in implementation, through clinician or patient bias. Thus, while it is important to focus on finding appropriate measures of fairness in model performance, it is equally important to ensure that those using the model are appropriately trained and do not succumb to 'automation bias' – over or under-relying on model output to produce errors (Parasuraman, Riley, 1997). As has been noted, '[t]he more advanced a control system is, so the more crucial may be the contribution of the human operator' (Bainbridge 1983).

02.2 AI and Patient-centred Medicine

The core of the concept of patient-centred medicine can be expressed as follows:

> First, healthcare should treat patients as *people* whose values, beliefs, and psychosocial context all play important roles in establishing and maintaining their health… Second, healthcare should treat patients as equal partners in medical decision-making: their wants should be heard, their wishes respected, and their knowledge considered. (Bjerring, Busch, 2020)

This collaborative and shared model of decision-making is a departure from prior paternalistic models, in which the clinician's recommendations were taken as decisive without substantive input from the patient. This new approach of shared decision-making and patient-centred care is increasingly taken to be the gold standard of clinician-patient interactions, in which 'the clinician offers options and describes their risks and benefits, and the patient expresses his or her preferences and values' (Barry, Edgman-Levitan, 2012).

Reliance on AI (and specifically deep learning) models has been argued to be in tension with the ideals of patient-centred care in two ways. First, it has been argued that AI models used as clinical aids, especially for treatment recommendation, are not sensitive to patient values and instead inflexibly operate based on fixed values. Taking as example IBM's Watson for Oncology system, Rosalind McDougall argues that the ranked list of treatment recommendations produced by Watson produces two harms: it bases its ranking solely on the particular value of maximizing lifespan, and in doing so does not 'encourage doctors and patients to recognize treatment decision making as value-laden at all' (McDougall 2019). This argument can be extended to any AI system that, like Watson for Oncology, prioritizes certain treatment recommendations over others, as doing so requires a metric (like longevity) to make that prioritization. The very act of picking such a metric loads the AI's output with a particular value that might not be prioritized by the patient.

The second way in which reliance on AI is seen to be at odds with the ideals of patient-centred care is grounded in the opacity of deep learning systems. As has been mentioned, deep learning systems are opaque in that it is often not possible to determine why a particular classification was made. Such opacity has been argued to compromise informed consent, a key requirement of patient-centred care (and perhaps a value accepted even by those who reject the model of patient-centred medicine). Bjerring and Busch note that if clinicians relying on opaque deep learning systems cannot understand why certain decisions are made, they are unable to communicate this information to the patient, which in turn doesn't allow the patient to make an informed decision (Bjerring, Busch, 2020). Unlike the previous argument, this applies not just to treatment recommendation systems but also to those that aid diagnosis.[3] Using the extreme case of a deep learning system that is advanced enough to draw unexplained correlations between hitherto unconnected medical variables, they argue that the clinician will be unable to convey information

[3] The distinction between diagnosis and treatment is not always a clean one, especially when one considers cases where diagnosis is done *by* administering treatment for potential ailments. Diagnostic decisions can thus be value-laden as well. However, for the purposes of this chapter, we shall continue to uncritically distinguish between these two processes.

relevant to the diagnostic process to the patient. According to this argument, the opacity of such systems compromises informed consent.

On the basis of these (and other) reasons, some have argued that patients have a right to refuse diagnostics and treatment planning by AI systems in favour of a human-only alternative (Ploug, Holm, 2020; de Miguel Beriain 2020). However, it is not obvious that these reasons are strong enough to support such a right. Even if we accept these reasons, they would support only a right to refuse *some* applications of *some* types of AI models. Considering the first argument above, when it comes to treatment recommendation and other value-laden decision-making, we may either reject such systems outright (which still allows their use in non-value-laden clinical decision-making) or we may build such systems to be adjustable when used, such that alternative or multiple values can be selected based on the patient's preferences and values (what McDougall calls *value-flexible design*). Furthermore, realizing patient-centred care requires that the *system* of clinical decision-making accommodate patient values, and this can be achieved in ways other than ensuring each element of the system (such as the AI model) is value-flexible. Given the complexity of patient values and preferences, it might not always be possible to represent them among the model's input parameters. Instead of rejecting such systems, we might instead compensate for them by having the clinician discuss with the patient and reorder treatment rankings based on the patient's alternative values and preferences.

While McDougall does entertain this option as a possible way forward, she argues that 'such an approach diminishes the patient's role and represents a backwards step in respecting patient autonomy' (McDougall 2019). However, it is unclear why this is so – hypothetically, if the ordered ranking in the case in which a patient's values are driving the ranking process from the start can be arrived at through compensating for the AI's approach, then there seems to be no ethically relevant difference in the two processes. In fact, this method arguably mirrors the one used in conventional non-AI settings, in which the doctor comes up with a diagnosis and an initial understanding of which treatment options are most feasible, and prioritizes them subsequently based on patient input. While it is true that 'current clinical practice… does not always meet the ideal of shared decision making', an AI model's fixed-value approach does not need to ground a right to refusal if traditional clinical decision-making does not. What *would* further promote patient-centred care by clinicians relying on AI decision-aids would be an explicit understanding of what the default values encoded in the model are (e.g. maximizing lifespan), and how this might compromise the values their patients might have.

It is similarly unclear that opacity of deep learning models can ground a right to refusal by the argument that it precludes informed consent. The predominant justification for informed consent rests on the value of autonomy and self-governance for the patient (Beauchamp 2010), and it is unclear that a patient's autonomy or ability to self-determine is reduced in the absence of knowledge of how a deep learning classifier generates a particular output. Even in a conventional, non-AI case, while patients may exercise their access to the chain of reasoning by which their doctor arrived at a clinical judgment, for the vast majority who are not medically trained this information does not contribute to autonomous action. As Beauchamp notes, "persons understand only if they have acquired *pertinent* information and have *relevant* beliefs about the nature and consequences of their actions" [emphasis ours], and it is hard to recognize information that one is not trained to assess as reasonably pertinent or relevant. The argument against deep learning models as being singularly problematic because of their opacity is further strengthened when we consider that many other cases of traditional medical practice are quite similar to them, in that they are 'atheoretical, associationist, and opaque' (London 2019). Whether we consider the majority of drug development that is never approved for any indication, half of phase III drug trials that fail, or the historical prescription of aspirin for over a century without understanding the underlying

mechanism, medicine is a field where intervention is often made on empirical grounds prior to understanding through causal explanation. To charge deep learning models then as being problematically opaque would be to level the same charge against many other aspects of traditional medical practice.

Similarly, even if a deep learning model could be made explicable using medically invoked concepts such that the clinician might sufficiently understand the recommendation, it is unclear how this contributes to autonomous choice for most patients and thus compromises informed consent. One possible rebuttal here is that even if information about how the AI model arrived at its recommendation is *functionally* irrelevant for a patient (it doesn't contribute to a different decision), as long as the patient *feels* the information relevant, then it bears on whether informed consent obtains. However, this would make the requirements for informed consent too strict, as many different types of information unrelated to the decision would then be required for informed consent so long as they are *felt* by the patient to be relevant.

That being said, there do exist patients for whom such knowledge would be reasonably pertinent and would contribute to a greater understanding of the situation, and for whom a case could be made that opacity has compromised informed consent. In such situations, it might perhaps still be plausibly argued that although AI systems do compromise patient-centred care, they might be justified on grounds of other benefits that they bring (e.g. superior accuracy, reliability, and scalability). Under such circumstances, the different ethical values of respect for patient autonomy and overall patient benefit would need to be weighed and traded-off against each other at the aggregate and individual levels.[4]

Beyond the arguments above, there have been other grounds posited for a right to refuse AI-assisted diagnostics or treatment planning. Ploug and Holm argue that since AI systems are known to be biased in various ways, and since their opacity may obscure proof of such bias, patients have a refusal right (Ploug, Holm, 2020). While they admit that such bias is not exclusive to AI systems, since clinicians also suffer from bias, they believe that there are adequate corrective measures and mechanisms for accountability for clinician bias that are not available in the case of model bias. For instance, clinicians may be corrected by other members of the health care team they operate as part of, through shared deliberation and discussion, and may be held accountable by such peers as well – all of which is harder to do for opaque systems. Another posited ground for a right to refuse is to permit patients to act on their 'rational concerns' such as reliance on medical AI leading to a society-wide de-skilling of medical professionals, or that 'AI diagnostics and treatment planning become monopolized with a number of negative effects' (Ploug, Holm, 2020). While we will not evaluate these arguments here, they do highlight the importance of future research on the question of what can justify a right to refuse AI-aided diagnosis or treatment by patients, where such a right would imply a right to insist on alternatives to the AI-assisted process. Another related important question is: do clinicians have an obligation to disclose reliance on AI systems to their patients? It is important that these issues be more fully explored to get a better understanding of other ethical concerns relating to patient-centred care that are raised by clinical reliance on AI systems.

02.3 AI and Value-based Decision-making

In the previous section we briefly discussed how the outputs of AI systems can be value-laden, such that patient-centred care might require a compensatory adjustment to the outputs. While our

[4] Arguing for this claim and fully fleshing it out is beyond the scope of this chapter, and so for now we shall just raise it as a possible but not yet defended position.

argument has been that this need not decisively compromise patient-centred medicine, care must still be taken in deciding how such systems ought to be built vis-à-vis the values embedded in them. In this section, we focus on questions relating to the values embedded in AI systems. Specifically, we focus on two questions: (1) Under what circumstance should we be especially careful about the values embedded in AI systems, and (2) how should we select these values.[5]

There have been a few suggestions in the literature for when it is especially important that we get right the values that are embedded in health AI systems (Freedman *et al.* 2020). First, when decisions need to be made quickly, there may not be time for a human operator to be in-the-loop to assess and compensate for the value-ladenness of such systems. There have been several AI models developed for intensive care usage (Gutierrez 2020), for instance to predict length of stay for patients (Sotoodeh, Ho, 2019), ICU mortality (Awad *et al.* 2017), and critical risk (Flechet *et al.* 2019). Unlike in the aforementioned case of the Watson for Oncology, clinicians may not have the time to adjust for values embedded in the model.

Second, in cases in which AI models are being used to solve computational problems that exceed human capabilities, the ethically-relevant decision might be hard to decouple from the computational decision. We have already encountered one such example : healthcare resource allocation models like risk predictors for high-risk care that, if not configured appropriately, perpetuate unfair discrimination. Another example is kidney exchange algorithms, which match prospective recipients for kidney transplants who have willing but incompatible donors with other similar pairs to facilitate a trade (Roth *et al.* 2004). Several AI models have been developed for this highly computationally demanding task (Dickerson, Sandholm, 2015).

Third, when AI systems need to be deployed autonomously (or mostly-autonomously), such that there may not be human operators to offset value assumptions in the system. For instance, in low-resource settings in which specialists are in short-supply, AI system performance might be at an adequate enough level that it would be unethical to withhold autonomous deployment (Schönberger 2019).

Under such conditions, special care needs to be taken to ensure that the values embedded are ethically robust and defensible. Given that by our very characterization of such circumstances, a human-in-the-loop arrangement for value mediation is not possible, it will be difficult to factor individual patient values into such systems.[6] We would thus need to consider values not specific to a single individual, more akin to the values that guide public health decision-making. For instance, in kidney exchange scenarios, decisions need to be made about the relative weights assigned to patient categories (such as young and old) in cases in which prioritization is necessary, such as for tiebreaking purposes (Freedman *et al.* 2020). In the United States, it is also required that considerations of social justice be taken into account for allocation solutions, such as to consider an 'assessment of their cumulative effect on socioeconomic inequities' (42 CFR § 121.4 - OPTN policies). Other social justice concerns might also arise, such as acceptance of minority donors given that in countries like the United States and Australia, 'white, young, wealthy, privately insured, and well educated' patients with kidney failure are more likely to receive a transplant (Reese *et al.* 2015). Appropriate answers to such ethically-charged questions would need to be represented within AI systems – if we do not explicitly encode the values we care about, alternative values will be implicitly encoded instead by omission.

[5] To have AI models that operate in ethical ways, there also remains the further question of *how* such values should be encoded in AI models.

[6] This remark precludes more speculative scenarios in which models can be prepared for patient-specific application before the application itself, perhaps by pre-emptively being trained on patient data or factoring in pre-articulated patient preferences (perhaps like a version of an advance directive for use under such settings).

How are these ethical concerns to be addressed? Whose responses to such ethical questions are relevant? Traditionally, societies have been able to reach consensus for such questions, despite widespread disagreement, through social and institutional structures such as courts, voting, mediation, public consultation processes, etc. (Cave *et al.* 2019). Recently however, one new approach has been to 'crowdsource' the ethical solution through gathering enormous amounts of data about public preferences. In the Moral Machine study, Awad and colleagues collected 40 million public responses, across 233 countries and territories, to moral dilemmas faced by autonomous vehicles (Awad *et al.* 2018). The moral dilemmas presented were characterized by unavoidable accidents that required a choice between swerving or staying on course to spare one of two parties – humans vs. pets, passengers vs. pedestrians, more lives vs. fewer, the young vs. the elderly, etc. Preferences (or moral intuitions) in these cases were registered and collated to identify cross-cultural variation, which was then suggested as relevant to policy decisions (though the exact way in which such data should be used was left open).

The Moral Machine experiment is unprecedented in scale, and one can see how similar approaches might be used to gather public intuitions about the values and decisions that should guide ethical public health policy. When considering health AI systems that will affect large numbers of people (such as a kidney exchange algorithms), gathering such information would clearly be relevant to the decision of what values such systems should be aligned with. The question is how we might use such information. As Savulescu et al. argue, it would be a mistake for such decisions to blindly follow people's moral intuitions – public views on moral questions can be deeply mistaken, such as when there is low support for organ donation despite it being rightly seen as a problem to be overcome (Savulescu *et al.* 2019). A 'reflective equilibrium' approach – where such intuitions are first screened for bias and uncritical reflection, then compared with the intuitions of professional ethicists, and subsequently evaluated through our more general ethical values and theories – would yield a better foundation for decisions on how the values of such AI systems should be determined.

02.4 Responsibility, Accountability, and Explanation

When we speak about actor being responsible for the outcomes resulting from the use of AI systems, we may mean that it would be deserved for them to be praised or blamed, or rewarded or punished, for these outcomes (Strawson 1962). When we speak about accountable use of AI systems, we may invoke the following widely accepted conception of accountability:

> a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgment, and the actor may face consequences. (Bovens 2007)

Thus, we can connect responsibility and accountability as follows: mechanisms of accountability (the relationships between an actor and a forum) provide the practical channels through which responsibility can be adjudicated and attributed. Based on the conditions through which we assess and attribute responsibility, the forum may pose questions, pass judgment, and subsequently decide the consequences for an actor.

Providing a systematic analysis of all the different conditions of responsibility, all the actors involved, and other elements of accountability mentioned above will be beyond the scope of this section (Wieringa 2020). We endeavour instead to raise what we consider to be the most interesting and pertinent ethical issues that occur in adjudicating responsibility and pursuing accountability.

First, what are the harms that we are to hold actors responsible and accountable for when it comes to the use of AI systems in healthcare? There are of course harms to the individual, when an AI

system that allocates scarce healthcare resources (such as ICU beds) errs and misallocates. There might also be injury that is sustained by patients when a clinician uses AI systems to help in diagnosis or determining treatments for them, either through an error in the system or on part of the clinician relying on the system (Parasuraman, Riley, 1997). Harm might also result from discriminatory behaviour (or behaviour aligned to misguided values in other ways) that we have discussed above; such harm would not just affect a single individual but would also have ramifications at the level of access to care to minority patient populations (Obermeyer *et al.* 2019).

Across these different cases, responsibility for the harm would be allocated to different actors, from the health care professionals using AI systems as clinical decision-support systems, to developers designing and building the systems, to auditors who are to review the system before deployment, to public health officials who are in charge of specifying the values which such systems would be built to operate by. A precondition for adjudicating responsibility in any of these cases is the presence of an established standard of care to prevent each of the harms.

One condition for attributing responsibility is the 'epistemic' condition, wherein moral culpability requires that the actor was not ignorant of the significance and consequences of their actions, or if he was, then this ignorance is itself culpable (Wieland 2017). In the absence of an established standard (or multiple established standards) of operationalizing fairness, for instance, it would be difficult to hold responsible an AI designer that ensured a model exhibited fairness as equal calibration rather than classification parity (Angwin et al. 2016). The effective attribution of responsibility, and thus the pursuit of accountability, for AI systems in healthcare hence requires standards for most of the issues that have been discussed in this chapter so far. Given that most of these issues are currently 'live' ones, such standards are still being debated.

One common mechanism to hold clinicians accountable is the adjudication of clinical negligence in cases in which injury is sustained by patients allegedly due to a clinician's actions. Traditionally, in the United Kingdom, such negligence is assessed by the Bolam test, which tests for negligence by checking if the clinician has 'acted in accordance with a practice accepted as proper by a responsible body of medical men skilled in that particular art' (*Bolam v Friern Hospital Management Committee* [1957]). If this test is failed, as assessed by expert witnesses, the clinician is considered to have not performed to the standard of care and been proven negligent. When it comes to the use of AI systems in clinical settings, some have argued that up to the point in which AI systems become undeniably superior in performance to clinicians, existing standards of care and means of assessing them are sufficient (Price *et al.* 2019; Schönberger 2019). However, using AI systems in clinical practice requires skill, not just in medical matters of fact, but also in the heuristics and methods for appropriate reliance on such systems, which might not be something that existing medical professionals possess (as evidenced by the various calls to revise medical education in light of AI use (Wartman 2019)). Given that Bolam requires that clinicians act in accordance with what skilled medical professionals consider reasonable, and given that medical professionals traditionally might not be skilled on how best to rely on AI models, this might introduce difficulties in the use of Bolam to judge negligence for injury caused by AI use.

As mentioned above, once there is consensus around an established standard of care for how clinicians should rely on AI systems, across varying performance metrics and system configurations (level of opacity, etc.), these practices can be encouraged across the field and used as the basis for judging negligence. Some authors present a model of AI use in which either the AI's performance is low enough that the standard of care would persist as it has traditionally been, or high enough such that full use of AI would just *become* the standard of care (Price *et al.* 2019). Whether AI performance in clinical setting will indeed follow this 2-step trajectory is an empirical issue. However, considering the number of deep learning proof-of-concepts that display initial

performance comparable to or slightly beyond the average clinician, it would serve us well to think about adjudicating responsibility through an appropriate standard of care for the use of models that are neither obviously inferior or obviously superior to their user. Looking to work on automation bias, social epistemology, and group psychology would be beneficial in designing appropriate reliance strategies (Mishra 2019).

One of the recurring features of deep learning models that is invoked in discussions of responsibility and accountability is their opacity. The fact that such systems are not explainable, that their recommendations cannot be explained in terms of the exact logic and weighting of the factors that yielded them, is taken to be in tension with their accountable use. As Doshi-Velez et al. note:

> By exposing the logic behind a decision, explanation can be used to prevent errors and increase trust. Explanations can also be used to ascertain whether certain criteria were used appropriately or inappropriately in case of a dispute. (Doshi-Velez *et al.* 2017)

The explainability of AI models is thus taken as a necessary, if insufficient, condition for accountability. As the argument goes, a user of an opaque system, such as a clinician relying on a diagnostic aid, needs to be able to understand the system's recommendation if she is to use it appropriately and subsequently provide an account of his decisions. AI developers also need to be able to understand why the models they develop function as they do, so as to be able to make sure that bad outcomes such as discrimination or value misalignment do not occur.

However, we believe that such characterizations of the necessity of explanation for accountability are too quick. Explanations here are seen as necessary for accountability because it is presumed that the development and use of systems which offer such explanations would always lead to better health outcomes (compared to the development and use of comparatively opaque systems) – in other words, because norms governing their development and use require them to be explainable. However, if there are acceptable norms for the development and use of deep learning models that don't require access to local explanations to produce the best clinical outcomes or operate in line with the ideals of justice and non-discrimination, then it seems unclear why explanations would be necessary for accountability. In such cases, accountability would merely require judging whether the various actors acted in line with these non-explanation norms – the local explanation for the model's recommendation would be irrelevant. Explanations are thus *instrumentally* necessary rather than *intrinsically* necessary – they are necessary to the extent that they feature in the norms against which accounts of the development and use of such systems are judged. If the best results from the development and use of such systems can be achieved despite them being opaque, accountability does not require an insight into their inner workings. Allow us to illustrate with a few examples.

Algorithms like the aforementioned critical-care risk-prediction model (Obermeyer *et al.* 2019) may face a charge of being discriminatory against a population group. One way of assessing such a charge is to gain some insight into the exact chain of causal reasoning that led to black patients being mislabelled as low-risk when white patients with similar profiles were labelled as high-risk. However, this is not the only way to assess the charge. As we have seen, measures of fairness like classification parity and equal calibration (as opposed to anti-classification) can also indicate to developers that their algorithm exhibits problematic discrimination. Further, such measures are completely derived from the outputs of such systems, rather than by examining their inner workings. As such, if these measures of fairness are accepted to be appropriate ones, then the account of an AI developer can be assessed and responsibility attributed (or not) without invoking the need for such systems to be explainable. Similarly, if there exist acceptable standards on how clinicians might use opaque decision aids, such as by comparing their approximate accuracy with

the model's and updating for whether in a particular situation a model likely outperforms or underperforms them, they could plausibly justify their decisions without invoking an explanation for the model's recommendation.

This line of reasoning is of course compatible with the possibility that for a given application, the development and use of explainable systems leads to better health outcomes and greater non-discriminatory behaviour than for opaque systems. This is what Doshi-Velez et al. meant when they wrote that 'explanation can be used to prevent errors and increase trust' (Doshi-Velez *et al.* 2017). A developer of risk-prediction models might create models that discriminate even less and a clinician might make even better clinical judgments if the system were not opaque. However, this is an empirical question, an answer to which cannot be presumed in favour of explainable systems. If it turns out that we can construct standards and norms for the development and use of non-explainable systems that produce better results, then as argued there is no recourse to the necessity of explanations through accountability. Furthermore, this empirical question is still open – research in human factors has shown that the presence and type of explanation can cause clinicians to over or under-rely on decision aids (Bussone *et al.* 2015). Similarly, it is an empirical question whether, for instance, risk prediction models will reliably be less discriminatory if they are built to be explainable rather than opaque. It is important that the question of whether accountability requires AI systems to be explainable boils down to empirical questions rather than normative ones, as it prevents us from seeing explainability as conceptually necessary for accountability, and frees us to look for alternative solutions.

02.5 Societal and Speculative Effects

In this section, we consider some potential impacts of the use of AI systems within healthcare that are a little more speculative, in that there have been fewer concrete discussions in the literature pertaining to them, and these effects will likely be felt further in the future than the topics we have considered thus far. Nonetheless, we believe that the impact may be sufficiently large as to merit discussion.

First, there is a concern that the use of AI in healthcare can widen health inequality in societies for reasons beyond the discriminatory potential of algorithms that we have already discussed. Specifically, the concern here is that the potential of AI in healthcare might be utilized primarily for tackling conditions that are more prevalent amongst those who are already healthy and less for conditions, such as infectious diseases and antimicrobial resistance, that have a higher burden globally or are present amongst those who are the least healthy, (Topol 2019; Joshi 2019) or most disadvantaged in other respects (e.g. income). As the capabilities of AI systems increase, the areas in which they are deployed will see the most gains in terms of healthcare outcomes, and so a disproportionate deployment to serve those who are already relatively healthy (perhaps because of disproportionate funding) can worsen the already unequal distribution of health. It is important to note that such inequality would violate two competing approaches to health care allocation, and so would be a concern regardless of which of these two approaches are believed to be right. First, it would violate a utilitarian approach to health allocation according to which the appropriate deployment of AI would be to those conditions for which the burden would be highest. Second, it would violate considerations of distributive justice in healthcare according to which the focus should be to prioritize improving the health of those who are worst off in society.

The clash between a utilitarian and a prioritarian approach is an existing one in global health, and not one which we will explore here deeply. However, so far, approaches to the use of AI in health have been driven primarily by the availability of data and funding, neither of which ensure that either a utilitarian or a prioritarian approach is fulfilled (and in fact might ensure that neither of

them is fulfilled). Ensuring that health inequalities do not worsen as a result of the use of AI requires the formulation of guidelines for ethical AI deployment that would direct such development towards either those conditions with the greatest burden or those affecting the worst off (Winters *et al.* 2020).[7]

The second issue to consider is the impact of extensive AI use in healthcare on healthcare professionals (HCPs). There already exists some empirical evidence for the proposition that reliance on computer-aid or other innovations reduces skills of HCPs (Povyakalo *et al.* 2013; Tsai *et al.* 2003; Hoff 2011), and the worry is that as AI systems become more capable and more pervasive, such de-skilling will increase significantly. While it seems that being able to provide healthcare through higher-performing processes would be good for health outcomes, there is the concern that in the long term, the quality of medical diagnostics itself will be lower than what they would have otherwise been (Froomkin *et al.* 2018). However, such analysis rests on fixed assumptions about the opacity of machine learning systems, assumptions which may have to be revised given the new directions research on deep learning might take (Bengio et al. 2019).

Beyond the question of de-skilling, there is also the question of HCP burnout. It is well established that burnout amongst clinicians has been accelerating, and in turn leads to medical errors (Shanafelt *et al.* 2010). Many have proposed that use of AI in clinical settings could help reduce some of the stress that is placed on clinicians and ameliorate the burnout, improving their quality of life and allowing better performance. However, the alternative is also possible – that adding AI use to clinical workflows increases the cognitive load facing clinicians (Maddox *et al.* 2019). The use of clinical decision support systems in the past has often added to the information clinicians need to process, often leading to instances of automation bias through over-reliance given the difficulty of verifying the outputs of the decision-aid (Lyell, Coiera, 2017). This possibility is especially pertinent given past instances in which adoption of clinical innovations, such as electronic medical records, backfired and led to additional burden being placed on front-line HCPs (Verghese *et al.* 2018). While such concerns might be addressed by the evolution of medical education and medical practice to rely more easily and substantively on AI systems, for now it is an issue worth watching out for and considering.

02.6 Conclusion

In this chapter, we have attempted to provide an overview of the landscape of ethical issues for the use of AI systems in healthcare. In many ways, the use of AI systems in healthcare shines a light on the same ethical concerns that can be observed with AI use in other domains such as predictive policing, criminal sentencing, etc. – the importance of explainable systems, questions around their accountable use, and about the values imbued in such systems such as fair treatment in recommendations. However, the discussion for the healthcare context also departs from these general concerns in other ways, in terms of both the specific instantiations of the more general concerns – medical negligence for accountability, historical public health consensus around what constitutes fairness – as well as the entirely new challenges, such as the necessity of upholding the ideals of shared decision-making and patient-centred medicine. It is important to take the healthcare context seriously, and not just assume that solutions worked out for AI elsewhere (such as techniques to make models more explainable) will suffice for transplantation.

[7] For the purposes of preliminary analysis here, we have utilized only two accounts of justice that might apply – there are many others that also need to be considered, including those according to which a just arrangement might not preclude a given level of inequality. [Unclear – rephrase]

There is plenty of further ethical research to be done, not just in answering the open questions that have been flagged in this chapter but also in addressing the adjacent topics that have not been covered, such as the health impacts of AI deployed in non-health settings (De Andrade *et al.* 2018). Further, any discussion of the ethics of AI in healthcare would be incomplete without a complementary discussion of data ethics, which similarly brings novel challenges in the healthcare domain.

However, for the purposes of this thesis, such a discussion suffices to lay out a somewhat broad sweep of the ethical discourse around these models. In the next chapter (Chapter 3), I will focus on setting up the linkages between two interconnected questions – what does an epistemically appropriate use of such classifiers in the clinical context require, and how does that impact their responsible use. This initial framing will usefully set up a deeper discussion later in the thesis on the three key epistemic and ethical issues outlined in the introduction: epistemically appropriate reliance on deep-learning models (Chapter 4), shared decision-making (Chapter 5), and accountable use of such models (Chapter 6).

03. The Epistemic Conditions for Clinician Responsibility

03.1 Introduction

This chapter aims to sketch an approach to what constitutes a responsible use of 'black-box' deep-learning models in clinical settings. Specifically, it aims to sketch an approach to thinking about *physician responsibility* in the use of *machine learning classifiers* (MLCs) for the purposes of *clinical practice* to determine diagnosis and treatment recommendations. Further narrowing the scope of the chapter, it will consider only the question of what are the *epistemic* conditions for physician responsibility in such a setting.

As mentioned in the introductory chapter (Chapter 1) and enumerated in the previous chapter (Chapter 2), there are of course necessary conditions for responsibility other than the epistemic condition – for instance, whether the physician is free to choose his actions and whether these actions actually cause a morally bad outcome – and many other ethical and epistemic issues even apart from that. However, given that the goal of this thesis is to illustrate a workable alternative to interpretable systems for some ethical and epistemic challenges, these will be left for separate consideration elsewhere. Physician responsibility also includes a responsibility towards the patient beyond just determining accurate diagnoses and efficacious treatment options, such as ensuring that the overall care is sufficiently informed by the patient's desires and values. However, this too shall be left for separate discussion in subsequent chapters (chapter 5, and to a lesser extent, chapter 6). This chapter shall focus only on the conditions for physician responsibility that concern the generation of accurate diagnoses and efficacious treatment options.

My approach here will then be as follows. In section 3.2, I will first introduce the state of development of MLCs for clinical practice, and importantly, repurpose a recently developed model as a potential example for an MLC deployed in clinical settings. Such a use-case will be important for our discussion, as the details of the model will ground our moral and epistemic analysis in subsequent sections. In section 3.3, I will sketch out an approach to determine how physicians can use such classifiers in epistemically rational ways, with the hope that epistemically rational usage will produce more accurate clinical judgments, which will produce better clinical outcomes. I will do so by first outlining what sorts of questions are necessary to answer to arrive at a decision about how one should use a classifier in a clinical setting, and then sketching how we might answer those questions in an epistemically justified manner. The aim here is not to provide such appropriate answers, but instead just to sketch the main issues that need to be considered for how to solve these questions, and what steps might need to be taken to arrive at an answer. Such a preliminary sketch will also be helpful for section 3.4, where we will consider the epistemic conditions for responsible use of such models by physicians in clinical settings. The epistemic landscape sketched in the previous section will be helpful in identifying what sorts

of beliefs physicians will need to have to responsibly use such systems, and we will then consider the issue of when ignorance about such beliefs can be blameworthy. I will also raise two arguments from the literature on culpable ignorance, reasonable foresight grounding blameworthiness and difficulty of acquiring information mitigating blameworthiness, and argue that they are especially relevant to clinical use of classifiers. Finally, I will raise a general puzzle about the epistemic conditions for moral responsibility that surfaces when we consider whether ignorance about certain beliefs about how best to use models can be culpable. In section 3.5, I shall then end by indicating how this overall preliminary sketch that will need to be developed further. I will also discuss the ways in which this project can be expanded further by looking at the other determinants of physician responsibility, including freedom and causality as well as patient-centered decision-making, and considering which are the other agents whose responsibility might bear on physician responsibility.

03.2 Machine Learning for Clinical Use – Towards a Potential Use-case

As outlined in chapter 1, clinical practice has traditionally been a great fit for the application of machine learning models because of how well-suited some medical tasks are to the specific areas that AI has emerged as competent in over the past decade. However, while these advances for ML applications in clinical use are promising, it is also important to put them in context. Most of the ML models that have contributed to the hype around ML in medicine have been trained on retrospective data, and have either not been clinically validated yet or are still in early stages of clinical trials. While regulatory bodies like the FDA have approved some AI algorithms for image interpretation in healthcare, most of these products have not been backed-up by peer-reviewed publications (Topol 2019). There have also been note-worthy cases where advertised AI applications in healthcare have failed to perform up to expectations in clinical settings (Ross *et al.* 2017). While these incidents shouldn't reduce our enthusiasm for the possibilities of ML in medicine, they should make us more careful about what we can expect from such models, how we should think about implementing them in clinical practice, and how we should understand what their responsible use would require.

For the purposes of our discussion about clinical responsibility for the use of ML models, it would be helpful to have a more concrete idea of what such a potential system might look like, and how it might be used in clinical settings. A case-study of an ML model currently in practice would be ideal, but given that we're still in the early stages of such technologies, such case-studies are hard to come by. To that end, this chapter will use as a prospective clinical use-case of an ML model one of the recent and well-known models developed: the use of an MLC to evaluate and accurately diagnose pediatric diseases (Liang *et al.* 2019). The main features of the model, and a potential extrapolation to its use in clinical practice, can be briefly described in the following table.

Model aim & description	This model aims to diagnose a range of pediatric diseases, across multiple organ systems (e.g respiratory diseases, gastrointestinal diseases, neuropsychiatric diseases, etc.), by evaluating the various types of clinically relevant information present in electronic health record (I) data.
Training data	The model is trained IEHR data, which was initially annotated by clinicians to select the clinically relevant features foIhe EHR charts, labeled with the initial diagnosis of the attending physician of the pediatric disease.

Model outputs	The model output in any particular case was the relevant, specific diagnosis.[8]
Explainability	The model is able to 'explain' its recommendation by highlighting which are the key clinical features (such as 'abdominal pain' or 'vomiting'), from which specific categories of the EHR (such as history of present illnesses, or physical exam), that drive the ultimate diagnosis prediction.
Model performance	The model is said to demonstrate "high diagnostic accuracy across multiple organ systems and is comparable to experienced pediatricians in diagnosing common childhood diseases".**Error! Bookmark not defined.** This accuracy was measured primarily using an F1 score, which is composed of the precision and recall of the model, which are computed using the true positive rate and false positive rate of the classifier.
Prospective clinical usage	We can thus imagine a clinical use of this model, based on the above features. Such a model of 'comparable' accuracy can be used to give the attending physician an added perspective on her young patients, on what diseases they might have. The physician can even, if she so chooses, gain some measure of insight into why the model selected a particular diagnosis based on the features of thIatient's EHR chart that it found 'relevant'.

Table 1 – Hypothetical clinical use case of an MLC for diagnosis

These details paint a slightly more concrete picture of the kind of MLC we might expect to find in clinical settings, and so will serve to ground the moral and epistemic analysis in the coming chapters better. While there have also been other models which we might have used as a potential use-case, this particular model (henceforth, the pediatrics classifier) was selected for a few reasons. First, the pediatrics classifier is a model that is more complex, in that because itI trained on EHR data, it takes as inputs a greater diversity of information types (patient history, physical exam, scans, tests, etc.). It also aims to diagnose a range of diseases, in comparison to some other models which aim to diagnose just one. This diversity of inputs and outputs, I think, is indicative of the more complex classifiers that we can expect to see in the coming years, and so an analysis that is based on such a classifier may be more applicable in the future than one of simpler classifiers.

Second, this model is at least partly explainable, as seen from the highlighting of the 'key' clinical features that drive a decision – this is based on an exogenous model providing explanations over the original model, as discussed in Chapter 1. Finally, this model is one that is touted to have diagnostic accuracy comparable to experts. It is unlikely that we will see models being introduced in clinical settings that have sub-human accuracy (DeFauw *et al.* 2018), and models which have significantly superior accuracy might be philosophically less interesting given that deference to their judgment might be more intuitively called for. A model which is said to have 'comparable' performance, on the other hand, raises interesting questions for moral responsibility. This decision isn't just for its philosophically interesting characteristics, however, given that there are also other well-known models which have claimed comparable performance (Esteva *et al.* 2017,

[8] This diagnosis seems to have been presented according to the ICD-10 diagnosis codes, though the paper doesn't explicitly mention so.**Error! Bookmark not defined.**

DeFauw *et al.* 2018), and so this feature of the model is also representative of other MLCs being developed.

03.3 Epistemically Rational Use of MLCs in Clinical Practice

In this section, I will sketch one approach to determine how physicians can use MLCs like the pediatrics classifier in maximally epistemically rational ways. The reason why a maximally epistemically rational use of MLCs is important is because such a use will maximize good clinical outcomes, though I will not argue for the details of this relationship here. I will instead assume (what it seems to me is uncontroversial) that if doctors can more reliably arrive at true beliefs rather than false ones about the tools they use and the clinical decisions they make, this is one determinant for maximizing good clinical outcomes. As mentioned, the goal here is not to provide *any* final answers to questions about rational use of such systems – it's merely to map out and establish the various beliefs that will come into play when thinking about their rational use, such that we can subsequently use such a map to consider what epistemic conditions need to obtain for the responsible use of such systems.

The approach I will take here then is to first identify the sequence of steps that physicians will take and the inherent questions they will have if they use such models in a particular way: as providing them with recommendations that compete with their own. I will then suggest ways in which the physicians can more rationally carry out this process, by having more justified answers to these questions. Importantly, the claim is not that the physician should engage in answering these questions through the methods suggested here – rather, the claim is that these methods should allow for better answers to these questions, and these answers should be accessible to the physicians for the purposes of using these models.

In general, I will phrase this discussion of how to most epistemically rationally use such classifiers as one particular (yet unexplored) practical application of the philosophical discussion on higher-order evidence (Christensen 2010, Christensen 2016). While first-order evidence, such as the test results a physician may consult to aid him in forming a diagnosis, contributes to an agent forming a particular belief (about what the medical condition of a patient is), higher-order evidence bears on the reliability of an agent's thinking or of their evidence on the matter. As an example, for a physician making a diagnosis, the fact that another physician disagrees with him on the diagnosis can be considered higher-order evidence towards the diagnosis he initially held. Similarly, if the physician has had a poor track record of clinical judgments in a particular domain, that would constitute higher-order evidence against the diagnoses he produces concerning that domain. We can thus similarly consider MLCs like the pediatrics classifier and the outputs they produce as constituting higher-order evidence for the physician who is producing outputs of a similar sort, for instance diagnoses in pediatric cases. In what follows, we will map the sorts of beliefs that the physician would have to hold, explicitly or implicitly, about the classifier and himself, and will then consider how this can be assessed as higher-order evidence so as to be epistemically rational in the classifier's usage.

Given that the function of models like the pediatrics classifier is to help physicians make (epistemically) better diagnostic decisions, the information provided regarding the model in *Table 1* is insufficient and so underdetermines the physician's decision. To understand why, consider again the information provided. The physician has access to (1) the model's recommendation on what the diagnosis should be, (2) an 'explanation' of that decision based on the model highlighting the parameters it looked at to make the decision, (3) a claim that the model exhibits 'comparable' performance to experienced pediatricians. Assuming for now that such claims are all justified, there are 3 further steps that are *necessary* for the physician to make a final decision on

the diagnosis. First, given that the claim of the model having comparable performance is established independently of any explanation that the model provides and is based solely on its accuracy metrics, how if at all should the provision of an explanation (and the contents of the explanation) change the epistemic status of the model? Put another way, how do the accuracy metrics of the model and the explanations it can provide contribute to the overall epistemic status of the model? Second, although we may know that the accuracy of the model is comparable to the physician's, how does the overall epistemic status of the model compare to the physician's epistemic status? For instance, in the unlikely case that explanation has no epistemic value (and merely some moral value), the epistemic status of the model and the physician will be measured entirely by their accuracy, and so be as comparable as their accuracies are. Third, how much weight should be given to the output of a model that exhibits an epistemic status comparable to the physician (assuming the epistemic status is comparable)? Without this crucial belief, the physician will be unable to decide how to use the model's diagnosis recommendation.

Thus, in using an MLC like the pediatric classifier to inform his final diagnosis, the physician has to answer the following questions:

(1) What is the recommendation provided by the model?

(2) What is the explanation provided by the model for this particular recommendation?

(3) What is the accuracy of the recommendation provided by the model in this specific case?

(4) How should the explanation the model provides modify or otherwise contribute to its overall epistemic status, initially established merely by its accuracy?

(5) How does the epistemic status of the physician compare to the epistemic status of the model established in (4)?

(6) Given the epistemic status comparison established in (5), where the epistemic status of the model is either superior, comparable, or inferior to the physician, how much weight should the physician give to the model's recommendation?

The MLCs that are being developed nowadays, such as the ones mentioned in the previous section, will provide the answer to (1) and (2). For instance, the pediatric classifier will provide a diagnosis recommendation for what disease the young patient may have, and will also provide an explanation (in one way) for this selection by highlighting the information that it considered relevant in the patient's EHR graph in making this diagnosis. Furthermore, information about the overall diagnostic accuracy of the classifier may go some way in providing an answer to (3). In the case of the pediatrics classifier, not only is there an indication of what the model's overall accuracy is in its diagnosis across most pediatrics diseases, it can also provide more granular accuracy metrics about its diagnostic ability for diseases per the relevant organ system (for instance, respiratory diseases vs. gastrointestinal diseases) but also accuracy metrics for its diagnoses of specific diseases (for instance how accurate it is in diagnosing encephalitis). These numbers can then be used to further triangulate the prospective accuracy of the model for a given clinical scenario.[10] However, this still leaves answers to (4)-(6), and these questions have so far been relatively unexplored.

The point here is not just that these questions need to be answered for the physician to evaluate and subsequently use the model's outputs in an epistemically rational manner. Rather, the point is stronger – for systems like the pediatrics classifier, the physician cannot arrive at a final

[10] Some classifiers actually have been built to provide a confidence level for each recommendation they make, rather than the physician just being forced to rely on classifier-level accuracy metrics.**Error! Bookmark not defined.**

decision without taking *some* stance on the answers to all these questions, regardless of whether this stance is justified, or even considered explicitly. For instance, imagine a case where the physician using the pediatrics classifier just has information about (1)-(3) and no explicit beliefs about (4)-(6). In such a situation, the physician is faced with a classifier that suggests a particular diagnosis, highlighting which parts of the patient's EHR chart led it to such a diagnosis. The physician would also have access to an approximate accuracy measure for that particular decision. The options facing the physician might be to offer a particular treatment based on his analysis or offer a different treatment based on the model's analysis.[11] In such a situation, the physician might decide ultimately to go with his own analysis since the model's recommendation, explanation, and predicted accuracy are not intuitively convincing to him. However, while this might seem to be a decision made just on answers to questions (1)-(3), the intuitive evaluation by the doctor actually reflects a stance on questions (4)-(6) as well. Roughly, for the doctor to go with her own analysis she would have to have at least an implicit belief that the answers to (4)-(6) support the final claim that going with her own analysis, without updating it for the model's recommendation, will maximize the final diagnostic accuracy. She might believe this, for instance, by believing that explanation adds nothing to the epistemic status of an agent (be it the model or herself) beyond the predicted accuracy of their diagnosis, that the predicted diagnostic accuracy of the model is comparable to her own, and that in such cases of comparable epistemic status the rational thing to do is go with one's own analysis.

Establishing that these are the necessary steps that a physician would take to finally decide on her decision when using such classifiers, the next task is determining how the physician *should* use such classifiers such that their use is maximally epistemically rational. To answer this, we need to know what the right answers to (1)-(6) are. We also need to determine how the physician can come to know these right answers. We shall leave aside the latter question for our current purposes, but we will return partially to it in section IV.

I will now sketch out how we might think about what the right answers to these questions might look like. (1)-(2) are merely factual questions, the answers to which will be provided by the model itself (as it is by the pediatrics classifier). So more specifically, we will look at how we might get the right answers to (3)-(6). My aim here will not be to arrive at these final answers, but instead indicate briefly how we might approach the task of determining them.

First, let us consider (3): what might constitute justified (or reasonable) belief about the accuracy of the model's recommendation, in any particular case. Model accuracy measurements are usually derived by testing them on a data-set, and by using metrics such as the F1 or AUROC scores to gain an understanding of the model's performance by measuring its true positive and false positive rates. Here, there is some debate about which particular accuracy metric for the model's performance is appropriate under which conditions (Saito *et al.* 2015, Lobo *et al.* 2007). Even if we are to settle on a particular accuracy measure, an evaluation of the accuracy of a model in a particular situation requires not just average, overall accuracy but more specific accuracy, perhaps the model's accuracy in making recommendations relating to the organ system that seems to be affected. Thus, to be maximally epistemically rational, the physician's belief about the accuracy of the model in a particular case has to be informed by the narrowest reference class for which model accuracy is available, regardless of whether that is the model's accuracy when diagnosing that particular disease, or diseases within that organ system.

[11] For the sake of argument, let us assume that there isn't enough time to collect further information which might allow him to adjudicate between the two potential diagnoses before suggesting a treatment. We might imagine, for instance, that this pertains to a classifier being used in time-sensitive situations like the ICU or A&E ward.

Second, with respect to question (4) ("How should the explanation the model provides modify or otherwise contribute to its overall epistemic status, initially established merely by its accuracy?"), we might think that explanation contributes to epistemic status in the following way. Imagine two deep learning models, both performing with the same accuracy. However, one model is able to provide some explanation for its classification, while the other can't. In such a scenario, would we say that both models have the same epistemic status, that one would be equally justified in relying on either of the two models? In one sense, we might say that they are, because it seems that they're equally reliable in their ability to produce true classifications. However, a model's epistemic status is important insofar as we can use it to guide our ability to rely on the model in the future – just because two models have produced similar or comparable accuracy rates in the past (especially so if these rates are produced in pre-clinical-trial settings, as they have been for most of these models), doesn't mean that they will perform so in the future. This is especially true in machine learning, due to the problem of 'over-fitting': a classifier gains high accuracy in predictions when trained on a dataset by 'memorizing' the dataset, but has significantly lower accuracy when it's tested on new samples the likes of which it wasn't trained on (Ravi *et al.* 2017). An explanation can serve to illuminate why the classification was made as it was, and by assessing the logic inherent in the explanation (such as whether the model considered all the relevant or any of the irrelevant determinants of the classification), a physician can better assess the model's likelihood of making a true classification in a future case.

Further, the epistemic status of the 'explanation' provided by the model would also change depending on whether the explanation is due to the model's interpretability, or if it is due to post-hoc explanability techniques applied to a non-interpretable model. An interpretable model is one where its decision-logic is accessible because the model's complexity has been constrained enough for it to be understandable to humans – for instance with non-deep MLCs such as decision-tree models or linear regression models. Such lower-complexity models, *ceteris paribus*, tend to generalize better to unseen data than higher-complexity models, and so would be more reliable if the accuracies were comparable. This is in contrast with cases where post-hoc explanability is obtained – a non-constrained, complex model, such as a deep-learning model, has an explanation for its decision-logic provided through a result of a secondary, more interpretable model being trained to approximate it and thus provide an explanation. It is important to note, however, that here it is the complexity of the model that contributes to its epistemic status rather than the explanation – the explanation is a necessary but subsequent product of what actually alters the epistemic status.

The classifier providing an explanation can thus allow the physician to better assess the model's reasoning, and further measure the quality of the higher-order evidence being offered if such a model's output were to conflict with his own assessment. However, while this makes sense theoretically, it may not always bear out practically, as some cases of automation bias have shown – where the mere fact of reliance on automation produces human errors. Clinical decision support systems (CDSSs) were sometimes under-relied upon by physicians, subsequently causing errors, for the mere fact that these systems provided explanations (Bussone *et al.* 2015). This topic thus merits further attention.

Third, as per (5), physicians need to have calibrated beliefs about their own epistemic performance, in comparison to the classifiers they use. In such a situation, if one of the ways in which epistemic status is determined is a combination of accuracy of classification modified in some way by an ability to provide explanation (or any other factors that we haven't considered so far), then physicians need to be able to assess themselves in those ways. Perhaps clinicians could classify themselves according to the categories of experts the model was tested against. For instance, the accuracy of the pediatrics classifier was tested against a panel comprising both

junior and senior doctors, and while the pediatrics classifier outperformed junior doctors, it was outperformed by the senior ones. One way in which doctors might determine their own accuracy in comparison to the classifier would be to find ways of reliably classifying themselves in such a case as either a junior or a senior doctor – noting that this would be a perhaps imperfect proxy for their own accuracy. Alternatively, more direct accuracy testing of physicians could be done to set a more commensurable baseline. More granular classification of the experts MLCs are tested against, where such granularity produces differing comparable accuracies, would definitely aid in physicians having more justified beliefs about (5).

However, this still does not fully solve the issue. On top of the availability of historical model performance, models can also be (and often are) designed to provide a quantified uncertainty level for any given classification. It would be difficult, but perhaps necessary, then for clinicians to start articulating the confidence levels they would attach to any diagnoses they had to make, and for those articulations to be calibrated to the extent model uncertainty quantification is. Perhaps this task would be eased by more coarse-grained categories of subjective confidence that the clinicians could invoke – certain, somewhat certain, borderline certain, not certain, etc. Either way, further thought on this is needed.

Finally, we come to (6): determining how the physician should use the model's output as per the results of the comparison in (5). There are (at least) two possible approaches we can take here – a theoretical approach and an empirical approach.

For the theoretical approach, we can look to the literature on social epistemology, and especially to the sub-literature on epistemic peer disagreement. The epistemological problem of disagreement essentially asks how (and how much) we should update our beliefs when we disagree with others about them. The problem of epistemic peer disagreement further asks how (and how much) we should update our beliefs when we disagree with epistemic *peers*. There can also be similar problems of disagreement constructed not just for one's peers, but also for epistemic agents who're superior in different ways, such as *experts* (agents who are superior in the information they possess and also their judgment about the information) and *gurus* (agents who are superior in their judgment but not necessarily in information) (Elga 2007).[12] If we are to look to this literature for guidance on how physicians should satisfy (6) and assign the appropriate weight to the recommendations of a classifier, our first task would be to see how we might be able to identify classifiers as an epistemic peer, superior, or inferior. In the literature, other human agents have been categorized based on features including their cognitive ability, evidence bought to bear, biases, background knowledge (Bryan 2014), but it's unclear whether and how this might apply to MLCs. Instead, we might resort to relying on their historical diagnostic accuracy, which is a common-sense and accepted measure of recognizing peers, superiors, and inferiors (Goldman 2001), though we may have to supplement it with other measures (such as perhaps explainability, as we have seen). However, if we can categorize these models as such, theoretical avenues of providing rational and justified answers to (6) open up. For instance, if a model is identified as an epistemic peer to a physician (based on all relevant factors including accuracy), there are some potential recommendations in the literature on how physicians should use them. Conciliatory approaches recommend that the rational thing for the physician to do is to reduce his credence in his own recommendations in the direction of the model's recommendations. On the other hand, steadfast approaches recommend that rationality does not require the physician to reduce his credences by any amount. Such theoretical approaches might provide one avenue through which might emerge justified beliefs about how physicians should use classifiers, having determined their vis-à-vis their own.

[12] The literature here also distinguishes more generally between rationality peers and accuracy peers, and how the two notions interact.**Error! Bookmark not defined.** This is something to consider further, moving ahead.

The empirical approach on the other hand would require gaining accuracy measures not just of the classifier performance and expert performance on a clinical task in isolation, as has traditionally been done for recent models developed, but also for different ways in which a physician may utilize these classifiers. For instance, physicians can use models in conciliatory and steadfast ways, and also in ways that (through principled application) are conciliatory in some cases and steadfast in others, and the accuracy of these different methods can then be established (by testing on a data-set, for instance). This would allow us to determine which of the competing methods of engaging the models are actually most reliable in practice. This also has the added benefit that technically, answers to questions (4) and (5) may not be needed – with enough resources, all of the most important conceivable ways of engaging with a model can be tested without first determining whether the model is an inferior, a peer, or a superior in its performance.[13]

These are some ways in which we might arrive at the right, or at least reliable, beliefs about the answers to questions (3)-(6). As is obvious, the analysis here is merely an initial sketch of how we might do so – the bulk of the work remains to be done. However, such a preliminary analysis of the epistemic requirements for a rational and reliable use of such systems is sufficient to ground our discussion of the epistemic conditions for the responsible use of such systems. This is what we turn to now.

03.4 Exploring Responsibility for Clinical MLCs

Exploring what responsible use of clinical MLCs would look like yields multiple rewards. Not only is it intrinsically ethically valuable to understand what users of such models need to do to behave in an ethically responsible manner, it also provides a building block to then examine other connected concepts, such as accountability in clinical settings, explainability of ML models, and clinical negligence. In this section, we shall thus explore the concept of moral responsibility first philosophically, and then fit it to our current use-case by considering existing general clinical norms and values, and the specifics of using models like the pediatrics classifier. Here, the type of responsibility we will be considering will be what is known in the literature as 'negative' responsibility – the adjudication of blameworthiness or praiseworthiness for actions taken by agents that may result in harm or good. I will also introduce one puzzle for moral responsibility in the use of MLCs which I will argue is a novel one, and isn't easily dissolved based on existing discussions of moral responsibility in the literature.

The goal here is ultimately to assess morally responsible behaviors for a clinician using an MLC. A problem case would be if a physician uses an MLC to inform his clinical decision, and this decision subsequently causes harm by causing an injury to the patient. The question then is, under what conditions would the physician's actions be blameworthy in such a scenario? How should the physician have instead behaved? These behaviors can include, for instance, the ways in which a clinician should use the MLC, and how they should act in cases of disagreement between their own assessment and the MLC's output.

Traditionally, there have been two conditions offered to assess if an agent is morally responsible in any particular situation (Jonas 1984, Eshleman 2016). First, there needs to be a connection between the agent and the outcome of their actions, such that the consequences must be traceable to the agent (call this the control condition). Second, the agent needs to be *aware* of

[13] Going through answering (4) and (5) might still be beneficial from the point of view of limiting the number of tested methods of using the model, considering that this would reduce the resources needed.

their actions, the consequences, and the moral significance (call this the epistemic condition). If both of these conditions are met, then the agent can be said to be morally responsible under these circumstances. For the purposes of this chapter, we shall concern ourselves with the epistemic condition. While the control condition also poses novel problems in the context of the use of MLCs in clinical practice, we will not explore it in this chapter.

In what follows, I will use interchangeably the phrases 'being held responsible for', 'being culpable for', and 'being blameworthy for'.

03.4.1 The Epistemic Condition for Moral Responsibility

If a physician can cure a patient's illness, but ultimately doesn't, we would consider him blameworthy for his actions. Unless of course he doesn't cure the patient because he doesn't know how, and so is ignorant of the means through which he can do so. However, is this ignorance itself culpable? Perhaps he is ignorant because a treatment for that illness hasn't been discovered yet. In such a case, it would seem that he is not blameworthy for his ignorance – there is nothing that he could have been reasonably expected to do that would allow him to have knowledge of the treatment. On the other hand, he could also be ignorant because he wasn't attentive in medical school when this disease and its treatment were being discussed. In such a situation, he seems to be culpable for his ignorance. There were things that he could have been reasonably expected to do in the past that would have resulted in him not being ignorant in the present case, and he is blameworthy for not doing them.

This is a simple example that illustrates what the main question concerning the epistemic conditions for moral responsibility is: to be blameworthy, what are the facts that an agent needs to be aware of in acting to produce the negative results as he does, and when does ignorance about these facts exculpate the agent? This can then be disassembled into two separate questions. First, what are the facts about which the agent needs to be aware for him to be blameworthy? Second, when does ignorance of these facts exculpate? We will consider each of these questions in turn.

Let us begin with a discussion of what the relevant facts are, and for the sake of the discussion let us assume that ignorance of these facts in the cases discussed is not culpable, so that the ignorance exculpates. For the purposes of our discussion, there are two main types of facts that an agent needs to be aware of for them to be held morally responsible: an awareness of the consequences of their actions, and an awareness of the alternatives available to them. For the physician in the cases that we're interested in, this would translate to him needing to be aware of the ways in which he's using the MLCs and the consequences that will have, as well as being aware that there are alternative ways in which the MLCs can be used (including not being relied upon at all).

Based on our earlier discussion in section III, the physician can fail to be aware in these ways by having false beliefs about the answers to any of the questions (3)-(6).[14] The physician might be unaware that the accuracy of the classifier for a specific recommendation may not be adequately represented by the model's overall average accuracy (or accuracy for a stated reference class), which is what he might be taking to be relevant. This can happen in cases where for instance the pediatrics classifier is used for older patients as well – something which the physician may not

[14] It seems that it would be hard for the physician to be mistaken about the answers to questions (1) and (2), given that they will be produced by the classifier for him. Depending on the classifier, the answer to question (3) may also be presented, as has been done with some classifiers that produce confidence levels with each recommendation as well.**Error! Bookmark not defined.**

know exceeds its scope of operation. The physician might alternatively be unaware that explanations should be used in a certain way to further refine the epistemic status of the classifier in a given situation, instead discarding them and believing that the classifier's epistemic status is higher than what it actually would be if he had included the explanation as part of the overall evaluation of epistemic status. Similarly, the physician can overestimate his own epistemic status vis-à-vis the classifier (or underestimate it), leading to an injury through a inappropriate use of dismissive (or deferential) attitude towards the classifier's recommendation. This overestimation or underestimation of his own capabilities can happen because he mistakenly, introspectively tries to attach a confidence level to his own analysis to compare to the model's, unaware that there are alternative methods to establish his own epistemic status compared to the model (as discussed earlier). Finally, the physician might appropriately believe the model to be one of comparable status in the specific case under examination, but mistakenly think that the only rational way to engage with models of comparable status is to be steadfast in one's own analysis, and discard the model's analysis completely. To determine what are the appropriate beliefs to hold in all these cases requires having sufficiently justified answers to questions (3)-(6) as mentioned in the previous section, but the important point is that being mistaken or ignorant about these answers will render the physician not blameworthy for his actions and their consequences, so long as this ignorance itself is not culpable.

This is then what we must consider now – under what conditions is ignorance about one's actions, their alternatives, and their consequences culpable? The consensus in the debate seems to be (at least) that an agent is only blameworthy for his ignorance if he is blameworthy for a prior act due to which he omitted to inform himself in an appropriate way, which had he done would not have led to his ignorance on the relevant facts (Wieland 2017). Such a prior act is known as a 'benighting act', which produces the ignorance or mistaken belief that subsequently leads to the 'unwitting act', which then produces the harm (Smith 1983). Thus, an agent is only blameworthy for his unwitting act if he is blameworthy for the ignorance that caused the unwitting act, and he is only blameworthy for his ignorance if he is blameworthy for the benighting act.[15] In the case of the physician using MLCs to inform his clinical decision-making, based on this we can observe that if harm is produced by the physician using a model where the physician was reasoning based on ignorance or false beliefs about (3)-(6), then the physician is only blameworthy for this outcome if he is blameworthy for the benighting act that produced this ignorance or false beliefs.

However, this just pushes the buck back by tracing blameworthiness about unwitting acts to blameworthiness about benighting acts. The crucial question now is how might we assess blameworthiness for benighting acts? Here, there is significantly less consensus. *Internalists* hold that blameworthiness for a benighting act is assessed in exactly the same was as it is for an unwitting act: by assessing if the agent had the appropriate beliefs (in the same way as answers to (1)-(6) are for the unwitting act) about his actions and their consequences and acted despite them, and if not whether their ignorance about these beliefs is culpable (Zimmerman 1997, Zimmerman 2008). However, as noted, this leads to a regress with a revisionist implication that we can only find agents blameworthy if at some point, their ignorance bottoms out in an *akratic* act: an action taken despite knowing that it would lead to harmful, negative consequences. This, it is then argued, either makes the landscape of morally blameworthy acts significantly sparser than what we ordinarily take it to be given that such akratic acts are not common (Zimmerman

[15] It's worth noting here that this kind of fine-grained philosophical analysis, even if appropriate, is unlikely to be used in a courtroom for actual rulings around negligence. Under those circumstances, a somewhat 'cruder' test might be necessary. However, considering the underpinning philosophical questions around responsibility will provide us with the conceptual apparatus to then consider the more approximate tests that might be directly useful in such legal scenarios.

1997), or we face an epistemic problem of not knowing in many cases whether the action taken was akratic in such a way and thus whether the agent was blameworthy (Rosen 2004) - thus, a sparsity of morally blameworthy acts that is either metaphysically or epistemically motivated. Since internalism about benighting acts has such revisionist implications, there have been alternative proposals for assessing blameworthiness for such acts, or even directly about the ignorance itself. One such proposal, for instance, posits that ignorance is blameworthy if it is produced by the exercise of epistemic vices such as "overconfidence, arrogance, dismissiveness, laziness, dogmatism, incuriosity…" (FitzPatrick 2008). Thus, it is sufficient for the physician to be blameworthy for his ignorance about appropriate answers to questions (3)-(6) if this ignorance is a result of him being overconfident or dismissive about what he needs to know, even if these epistemic vices were not exercised knowingly by him.

A full discussion and adjudication of the various ways in which ignorance can be found to be culpable is beyond the scope of this chapter. However, I will note here a few, preliminary thoughts on how this discussion can be applied to our case of clinically-used classifiers. First, there are some cases where blameworthiness seems to be clearly, intuitively, established. Imagine that the physician was educated that (a) he would be making beliefs about questions (3)-(6) when using a classifier, and on (b) what the correct beliefs or belief-generating processes about these topics were. It seems that if subsequently he was still ignorant or held mistaken beliefs, where these beliefs led to harm, then his ignorance would indeed be culpable. This seems intuitively compelling as a general policy pertaining to determining physician blameworthiness, even without needing to outline an account for why the physician is blameworthy for his failure to act according to (b) despite being told to do so (as the internalist would have to provide). Perhaps the intuitive force comes from a tacit acceptance of the 'reasonable foreseeability' condition that is often quoted in the literature (Vargas 2005, Fischer 2009), where what matters for blameworthiness is not whether the agent *actually foresaw* the consequences of his action but whether the consequences were *reasonably foreseeable*. Being told about (a) and (b), it seems that the physician should have been able to reasonable foresee the harm he might produce, and so would be culpable. This more pragmatic approach to assessing blameworthiness, and thus responsibility, cuts off the regress we've outlined above.

This is further supported by the fact that when we look at existing methods of determining physician responsibility (albeit negative responsibility), something like reasonable foreseeability is required. For instance, when it comes to rulings of medical negligence, to determine if a physician is responsible for an injury to a patient, it is necessary to establish that the physician has fallen short of the standard of care expected of them. Assessing whether the physician has fallen short of the standard of care (and thus has breached his duty of care) requires assessing whether their "error was one which would be made by a professional exercising *reasonable* skill and care" [emphasis mine] (*Muller v Kings College Hospital* 2017, Herring 2018). The rough idea here is that the physician's responsibility for the harm produced is assessed according to whether it fell short of the consequences it would have been reasonable to foresee, and thus the actions it would have been reasonable to take, based on the medical learning and experience that he had (*Montgomery v Lanarkshire Health Board* 2015, *Rogers v Whitaker* 1992).[16]

Second, it seems that the difficulty of correcting the ignorance should matter to the extent to which an agent is held blameworthy for it. This is a position which has been echoed in the

[16] The fact that existing methods of determining physician responsibility in the field echo the test of reasonable foreseeability need not be a decisive argument in favor of this method of assessing moral responsibility for ignorance – one might believe that in such a case, the legal understanding of responsibility and its requirements is ethically shortsighted. However, to the extent that we see proximity to existing practice as an advantage (even if just a tie-breaking one) in practical ethics, this resemblance has some force.

literature as well, with some claiming that difficulty which arises from context or lack of skill, where it's hard to determine what the appropriate belief is that one should act according to, mitigates blameworthiness even when other types of difficulties (such as requiring a lot of effort) may not (Bradford 2017, Fuerrero 2017). This becomes relevant to our case if we imagine that doctors would be considered blameworthy for not having reasonably appropriate beliefs about (3)-(6), even if they would have to research for and establish these beliefs purely by themselves. If the approaches specified in section III about how one might establish reasonably appropriate beliefs about these questions are true, this might take physicians far outside the kinds of arguments they're normally skilled in evaluating, making their judgments and reasoning about these domains less reliable than native practitioners in those domains. We might imagine that under such considerations, the physicians might at least be less blameworthy than compared to when they fail to determine appropriate beliefs about some other purely medical topic.

03.4.2 A Puzzle for the Epistemic Condition

This brings us to one specific case that raises a puzzle for the assessment of the epistemic condition for moral responsibility. This puzzle is raised because of a tension between (1) certain assessment criteria for whether or not the epistemic condition is met (i.e. that the ignorance is culpable), and (2) certain topics about which moral responsibility, and thus the epistemic condition, needs to be adjudicated. The case goes as follows.

Imagine a case where the physician is to use the pediatrics classifier. Further assume that the pediatrics classifier has been evaluated to have epistemic status comparable to the physician for a particular case, and so is the physician's epistemic peer. To use the classifier in a way that is rational, the physician needs to be aware of and act according to appropriate beliefs about (1)-(6). As it happens, the physician is not consciously aware that he needs to know the facts of the matter about anything apart from (1)-(3), and so ends up acting on implicit beliefs about (4)-(6). For our current purposes, we won't specify the exact mechanics through which this belief is an implicit one – we'll merely state that these beliefs are hastily and uncritically formed, and are perhaps intuitive. One of those implicit beliefs is about (6), where the physician implicitly accepts that the appropriate response to an epistemic peer is just to persist with one's own analysis and beliefs, as far as further action is concerned. Thus, when the physician encounters a disagreement between his analysis and the model's, he goes with his own analysis steadfastly and administers the relevant treatment. This diagnosis ends up being wrong and causing injury, such that if he had updated in favour of the model's recommendation this would not have happened. Is the physician blameworthy in such an instance?

The key aspect of this scenario that might prompt the intuition that the physician is indeed blameworthy is that the physician's belief about (6), that the optimal strategy in cases of peer-disagreement is to stand by your belief, is uncritical and intuitively-motivated. They are not derived in any way from expert reasoning or belief-forming processes, if we assume that deliberations on question (6) admit of distinction in reasoning employed by laypersons and experts. Insofar as these beliefs are formed hastily, intuitively, and with no evidence of expert-reasoning we might argue that they are unjustified. They are unjustified not in an *internalist* sense, for the physician might not be aware of any defeaters to forming that particular belief, such as an awareness of the fact that such beliefs are formed hastily and intuitively. Rather, they are unjustified in an *externalist* sense, where the external criteria is something like 'beliefs should be generated based on reasoning employed by skilled / specialist / expert operators in the domain". Such operators in this case might thus be epistemologists, or others, who're trained on how peer disagreement is to be managed. We might thus think that insofar as these beliefs are implicit, hastily and intuitively formed by the clinician, they are unjustified, and the physician is culpably

ignorant in such an externalist sense. Thus, the clinician would be blameworthy for the consequences of his action.

This remains unaffected by the fact that some of those who're competent when it comes to the domain of decision-making when faced with epistemic peers, such as various social epistemologists, would also stand by and advocate such steadfast views. We might explain why this does not affect our intuition of the physician being blameworthy by arguing that these experts hold their beliefs in a justified manner, because the process by which they arrived at their belief – extensively studying the literature, considering counterexamples, conversing with other experts, running experiments – is much more reliable in producing an acceptable ratio of true to false beliefs. The physician's process is thus inferior to the expert's in producing judgments about the domain of higher-order evidence about epistemic peers, as far as justification is concerned. That both processes produced the same belief does not mean that the belief is equally justified in both cases, and to the extent that it is less justified (or totally unjustified) for the physician, he would be blameworthy in a way that the expert may not be (absent defeaters, such as the physician's ignorance not being culpable).

We might illustrate this with another example. Imagine if a layperson were to come across an individual having certain breathing difficulties and coughing, and it seems intuitively obvious to the layperson that what would be helpful is patting this individual with some force on the back, to relieve his difficulties. The layperson proceeds to do this, and this ends up worsening the situation, and causing some harm to the afflicted person. Now, it might turn out that even among medical practitioners, there is genuine disagreement about such a case, where some believe that what should be done is exactly what the layperson did, while others believing that some other treatment option (such as the Heimlich manouvre) is more appropriate. However, the mere fact that the layperson arrived at the same belief about what should be done as the competent practitioner shouldn't exculpate him for the harm that he caused, precisely because his decision was not justified (in the externalist sense) while the medical practitioner's was. This is why if a competent practitioner had held the steadfast view, or had gone for the forceful patting approach, and the injury was still sustained, they would be less (if at all) blameworthy given that their beliefs were held justifiably, despite the state of genuine disagreement among competent practitioners on what the right way to go is. This sort of reasoning is also embedded in practical determinations of responsibility, for instance in cases of medical negligence. In deciding whether a physician has been negligent, the judge is sensitive to genuine disagreement among practitioners, such that if there are competent practitioners who find the conduct of the physician reasonable, even if there are others who don't, the physician is seen to not be negligent (Herring 2018).

So far, it seems that the physician, like the layperson, is blameworthy for acting on steadfast beliefs even if such beliefs are held by competent practitioners as well. However, there is a further twist here: when it comes to disagreement with epistemic peers, some of the competent practitioners defend the steadfast view against opposing views *due to its intuitive appeal*. For instance, the steadfast view is seen as superior to conciliatory views due to it being more in line with intuitions about self-trust and not being spineless when it comes to one's own analysis (Elga 2007). If fitting more with such intuitive considerations justify the steadfast view when it comes to epistemic considerations, then to what extent is the physician really blameworthy for relying on beliefs which were formed by him hastily but *intuitively*? For even if he may not have as much familiarity with the relevant arguments and paradigmatic cases as the competent professionals, the intuitions that he may be relying upon might be the same ones that epistemologists use to justify the steadfast view. There is a clear disanalogy here with the layperson and his treatment through forceful patting – competent medical practitioners do not find the forceful patting

method justified (even partly) because of its intuitive appeal, the same intuitive appeal that is also accessible to the layperson. Defenders of the steadfast view, on the other hand, appeal to its fit with such intuitions as justification. The intuitive appeal of steadfastness then is an argument for, even if not exculpating him completely, mitigating the physician's blameworthiness.

Thus, on the one hand, the physician's beliefs about steadfastness as an appropriate attitude to peer disagreement seems unjustified as it's not produced in a way that seems acceptable based on an externalist criteria – it is produced hastily, uncritically, intuitively and not based on expert belief-forming processes. On the other hand, perhaps there is an argument to be made for the physician's belief being justified precisely because it is intuitive, as experts might be appealing to the same intuitions for justification. On the former view, the physician is culpably ignorant while on the latter view, he isn't (or is less so than it initially seems). This clearly has implications for whether the physician can be held responsible for the consequences of his action, as we've seen in the previous section, and so this puzzle seems crucial for our task.

To answer to this puzzle, we will have to better understand what precisely it means for a belief to be formed implicitly in such a way, as well as what is sufficient for a belief about what is rational to do to be justified. If the belief is formed hastily and uncritically, is it a matter of luck that the physician ends up with the steadfast view? If it is so, then this would imply that the physician's process of belief formation is less reliable than the expert's even if the intuitions held are the same (assuming that the expert's belief forming process isn't chancy in the same way). This, or other arguments concerning the role of luck (Pritchard 2015), could identify the physician's belief as less justified. It could also be argued that even if luck isn't involved in any robust sense,**Error! Bookmark not defined.** hastily formed beliefs are otherwise less justified.

As it stands, there might be three possible ways of resolving this tension and solving the puzzle:

(1) Accept that the clinician's ignorance is still culpable, as though they may have accessed the same intuition, there are crucial parts of the expert process missing – a deliberation on the basis for intuitions being justificatory, a consideration of opposing evidence (counterarguments from conciliatory views, for instance), etc. Since the clinician has accessed just the intuition and not the rest, they would still be culpable, if the externalist criteria we've specified is an acceptable one (let's say if social epistemology was taught in medical school).

(2) Bite the bullet and accept that the clinician's ignorance is not culpable, because it is not truly ignorance – the clinician has accessed and reasoned using the same data (the content of the intuition) and in the same way (finding intuitions justificatory) as the some of the 'experts'. This might make us re-evaluate whether social epistemology, formulated for such a case (peer disagreement), admits of expertise at all, or

(3) Reject such an externalist criteria (or further refine it before accepting it) as one that is appropriate for the assessment of whether the epistemic condition for moral responsibility has been met.

This puzzle concerning the epistemic conditions for moral responsibility deserves further attention. As it is currently formulated, it is essentially a question about the extent to which physicians satisfy the requirements of rational reasoning in clinical practice. However, it is not a puzzle that arises just in cases of physicians using MLCs – if it were, we could ensure that physicians have the relevant familiarity with arguments in social epistemology so that they're judgments are no longer implicit (hasty, uncritical, intuitive, possibly chancy). Rather, it is a problem that arises for moral responsibility more generally when the actions which are to be evaluated for blameworthiness are motivated by beliefs that are (1) held implicitly but intuitively,

in the above manner, by the agent, (2) also held justifiably by competent practitioners, and (3) are justified on grounds of intuitive appeal. If action based on such actions leads to harm, we face the same puzzle. We can imagine that any actions based on beliefs from domains where such beliefs are justified on grounds of intuitive appeal, where there is disagreement about such beliefs due to differing intuitions, and where such beliefs can also be luckily and uncritically held by laypersons will face this issue. Such a description might be said to apply to very many beliefs about what morality or rationality demand.

In this section then, I have attempted to lay out the main contours of the debate on the epistemic conditions for moral responsibility, highlighting where the consensus lies and how we might map the situation of physicians using MLCs appropriately to these discussions. I have also flagged some specific sub-discussions in the literature which might be of particular interest in our case, such as the discussion around 'reasonable foreseeability' as well as discussions around the difficulty of correcting ignorance and how that bears on blameworthiness. Finally, I have raised a puzzle that arises when we attempt to analyze the use of classifiers in clinical settings by looking at the epistemic conditions of moral responsibility.

03.5 Next Steps

The aim of this chapter has been to sketch an approach for thinking about the responsible use of MLCs by physicians in clinical settings. One necessary condition for physicians to be held responsible for any consequences resulting from their use of MLCs is that they are either aware of the facts relevant to the use of such classifiers, or that their ignorance is culpable. In section 3.3, I have outlined what these sorts of facts may be and how we might determine them, and in section 3.4, I have outlined what are the conditions under which ignorance of them would be culpable, as well as the questions we have to solve even to determine whether ignorance obtains when the facts are about what it is rational to do when it comes to higher-order evidence.

In the stitching together of these various literatures, there have been several questions flagged for future resolution, and I shall briefly list the mains ones out here:

i. How do explanations offered by models for their recommendations count as higher-order evidence to physicians using such models?

ii. How can the physician best measure his epistemic status, vis-à-vis his accuracy but also his rationality, with the models that he uses?

iii. What are the norms according to which the physician should use the model's outputs (and characteristics) as higher-order evidence about the decisions he makes in clinical settings?

    a. Which of the traditionally considered views on whether disagreement, for instance, constitutes higher-order evidence (e.g. steadfast vs. conciliatory views) are most applicable in clinical settings, either by virtue of being the correct view or in virtue of other desiderata?

    b. Whichever view is most applicable, how can we operationalize this for clinical use by the physician?

iv. Our discussion of epistemic rationality relies on how we understand justification, as does our discussion on moral responsibility on how we understand culpable ignorance. There seem to be concepts, such as the notion of 'reasonable' competence or foreseeability, which serve a similar function in evaluation of moral responsibility in practical settings.

How can we understand and situate such concepts in traditional philosophical discussions about rationality and responsibility?

v. How difficult is it for physicians to inform themselves about the most rational ways in which to use such models, what kind of difficulty is it, and to what extent should this difficulty mitigate their blameworthiness if they fail to do so?

vi. How can we better characterize the implicit beliefs, about what it is rational to do, relied on by physicians when making clinical judgments? To what extent are these plausible candidates for *justified* beliefs about what it is rational to do, especially about higher-order evidence?

These are at least some of the questions which we need to tackle to better understand what epistemic conditions physicians need to satisfy to be held morally responsible for their use of MLCs in clinical settings.

However, even apart from these questions, there are further uncertainties if we want to consider the larger problem of physician responsibility beyond the epistemic condition. First, there is the question of the conditions under which physicians may fail the control condition, even if they satisfy the epistemic condition. Second, we've so far been concerned with responsibility for using these systems to make more accurate clinical decisions. However, accuracy is not the only desideratum in clinical decision-making. Recently, the locus of clinical decision-making has shifted towards the patient such that the desiderata of being in line with patient values has come to be seen as increasingly important (Herring *et al.* 2017). The use of MLCs then has to be considered in light of this desideratum of clinical decision-making – what are the epistemic and control conditions that need to be met for responsible use of MLCs to ensure practice in line with patient values? Third, we've so far been considering the conditions for moral responsibility, understood as blameworthiness. This has generally been understood as the conditions needing to be satisfied by an agent such that he is deserving of certain sorts of reactive attitudes such as resentment, indignation, etc. However, if we wish to translate these conditions for moral responsibility into more practical settings, for instance into conditions for legal culpability in cases of medical malpractice, then we will have to adapt our analysis. This analysis will have to be adapted from determining the conditions that are to be satisfied to be deserving of the relevant reactive attitudes, to determining the conditions that are to be satisfied for legal culpability. Legal culpability may then be constituted not just by whether one is deserving of these reactive attitudes, but also by other purposes such as whether one is deserving of punishment, whether being found culpable can serve as a deterrent to others, etc. It may seem that the conditions for responsibility will not have to be modified greatly when this transition is made, but this further analysis has to be done nonetheless.

The work in this chapter thus represents just a first step in a much larger enterprise. Subsequent chapters will focus on many of these questions, and put forth the argument for why explanations (either through interpretable models, or even with exogenously 'explainable' ones) are not necessary. The next chapter will begin this illustrative approach by showing how a doctor might still rely epistemically on deep-learning models, in a justifiable way, without access to any sort of explanation from the model.

04. Clinical AI and Epistemic Disagreement

04.1 Introduction

Given the success of deep-learning models in medicine, their capabilities can have an unprecedented impact not just in resource-poor settings with lower access to trained medical practitioners, but also as clinical decision aids in relatively resource-abundant settings. However, in these settings where they are to be used by a clinician, the following question is raised: how should a clinician use these models? More specifically, how should the clinician take a model's outputs as evidence in cases of disagreement between the model and the clinician, given the model's reliability?

This question gains further importance once we consider that traditional approaches to judging appropriate reliance on clinical decision-support systems (Bussone et al. 2015), or even other agents, are predicated on our ability to inspect the reasoning that leads to their recommendations / decisions. However, as has been discussed so far, such an option isn't natively available with deep-learning models, which are 'black-box' in nature. In the absence of such an 'explanation' for the model's recommendation, are there alternative approaches to appropriate epistemic reliance on such models? If there are, then the force of worry from the opacity of these models compromising their trustworthy use (Kaur et al. 2023, Ribiero et al. 2016) is lessened.

There is a rich body of work in human factors psychology dealing with failures of reliance on automation. There have been many cases of humans over-relying (a phenomenon known as automation bias) and under-relying on automated systems, with adverse outcomes (Parasuraman and Riley 1997). Such inappropriate reliance occurs due to different reasons ranging from inappropriate trust in the system, high mental workloads, and inability to appropriately deal with risk. As has been noted, "[t]he more advanced a control system is, so the more crucial may be the contribution of the human operator" (Bainbridge 1983). The field of human factors aims to understand, quantify, and reduce this inappropriate reliance primarily through the improved design of information systems, automation, and the environment that such systems are embedded in. However, there is an alternative (and complementary) perspective that can also be taken – reducing inappropriate reliance through improving the heuristics and epistemic procedures used by the human operators, who in our case are the clinicians.

In this chapter then, I will aim primarily to explore how this alternative approach can be pursued. I will do this by focusing on discussions of epistemic disagreement and higher-order evidence in social epistemology, and using them to model the problem of appropriate reliance on AI as a problem of epistemic disagreement. The relevance of these discussions to determining appropriate reliance on AI models has been suggested before, though not discussed in detail (Mishra 2019, Krishnan 2019, Crote and Berens 2020). Through the discussion here, I will aim to

establish certain conditions and desiderata that if met, will allow for epistemic approaches to acting under disagreement to bear productively on the question of AI-reliance. Further, structuring the AI-reliance problem as one of epistemic disagreement will allow us a new way of considering the empirical evidence from human factors, such that approaches to the former will bear on the problems raised by the latter.

Reducing inappropriate reliance through identifying better heuristics and epistemic procedures for the user is important for two reasons. First, it might improve outcomes from using AI models beyond what can be achieved by human factors approaches of *designing* the systems for better reliance. Second, it also sheds light on what accountable *use* (as opposed to accountable design) of such systems will look like, by establishing a standard of care for clinicians (more on this in Chapter 6). Furthermore, if insights from epistemic disagreement and higher-order evidence can indeed be harnessed, this might lead to the interesting result of articulating a conception of accountable use without requiring that AI models be explainable.

This chapter is the first of three chapters that takes an existing articulated problem with the use of black-box models in medicine – namely, how can doctors rely on, and indeed trust, such models without access to their reasoning – and shows how the problem might be addressed without reliance on explanations.

The structure of the rest of the chapter will be as follows. In section 4.2, I will briefly present and evaluate the evidence so far on how well medical models perform relative to human practitioners. As we shall see, this evidence is fairly promising. In section 4.3, which will be the main part of the chapter, I shall introduce the discussions of epistemic disagreement and higher-order evidence, and consider how well they apply to the question of AI-reliance. I will further articulate some conditions and desiderata that would allow for a more productive application of insights from these discussions. I will then end with some closing comments.

04.2 Comparing Medical-AI and Human Performance

Over the last few years, deep learning systems have garnered a lot of attention for outperforming previous state-of-the-art models across many domains. This step-change in deep learning capabilities was highlighted best by the development of a deep convolutional neural network (CNN) for image classification that significantly outperformed existing benchmarks (Krizhevsky *et al.* 2017). Image classification tasks in particular have seen a gold rush, given that previous AI architectures were especially ill-suited for such tasks, and deep learning approaches through CNNs are particularly suited to it. This has led to a lot of claims of human-level, and sometimes even superhuman-level, performance by ML models in domains that have traditionally been dominated by human experts.

On what basis are such claims made, and how reliable are they? Claims about model and expert performance in such studies are assessed by the accuracy of the predicted outcomes – for instance, with what accuracy can a model or an expert diagnose skin cancer by looking at an image of a potential skin lesion (Esteva et al. 2017)? The accuracy is usually measured either by (a) looking at the level of true positive, true negative, false positive, and false negative counts, or (b) combining these counts at the level of sensitivity (proportion of all positive cases captured by the model) and specificity (proportion of all negative cases captured by the model), or (c) combining *these* measures at the higher level of the area under the receiver operator characteristic (AUROC) curve, or other related measures (F1 score, AUPRC, Brier score, etc.). Typically, these measures of accuracy are calculated (in increasing order of reliability) on retrospective data that has already been collected prior to the study, though in some cases they're calculated based on

prospectively collected data, and sometimes even for data from real-world clinical use. The 'ground truth' for what the correct diagnosis is, against which model and expert performance is judged, varies across studies – for instance in the skin cancer case above, the ground-truth labeling for the data is done by other dermatologists. As can be imagined, the more reliable such labeling is, the more reliable the associated performance measures are.

So how reliable are such measures? There have been a couple of recent meta-analyses done to understand (1) how reliable the studies that establish such performance are, and (2) how AI model performance in medical imaging has generally fared compared to human practitioners.

There are a few things to note about the reliability of most studies using ML for medical imaging (Nagendran et al. 2020). First, the vast majority of such studies are conducted on retrospective data, with only a few using prospective data (let alone data from the clinical deployment of these models). Second, there is a lot of variation in how such studies are validated, with only a few validating through out-of-sample datasets (where performance is tested on data from new populations not represented in the training data, though this can still be retrospective data) from another geographical region or a different time period, let alone both. Third, most studies use a combination of expert and non-expert human practitioners to compare model performance against, and the average overall human comparison group is generally small. Finally, independent evaluation of the risk of bias for these models yielded a classification of 'high risk' for most of the models assessed in these meta-analyses – this leaves open the possibility that once these models are tuned to reduce such risk, model performance might fall.

That being said, the results about model performance are promising. In a meta-analysis of 25 studies where the performance is measured through an out-of-sample external validation, average pooled model sensitivity and specificity values (88.6% and 93.9%) are significantly higher than for human practitioners (79.4% and 88.1% respectively) (Liu et al. 2019). However, once we only consider studies where this out of sample validation is done by using the same sample for both the model and human practitioners, the difference in average pooled performance falls (model sensitivity is 85.7%, specificity is 93.5%; human sensitivity is 79.4%, specificity is 87.5%). When we further compare only the best model and human accuracy scores produced in each of these studies, the difference falls further (model sensitivity is 87%, specificity is 92.5%; human sensitivity is 86.4%, specificity is 90.5%) such that the model performance seems approximately equal to human performance.

What these numbers suggest (through an analysis that has an admittedly small sample size) is that currently, medical imaging models are seemingly performing on-par with the best expert human practitioners. In the coming years as new datasets become accessible, computational power increases, and there are further advances in AI theory, it is extremely plausible that this performance will improve further. Arguably, this will happen not just for medical imaging models but also for models classifying based on more than just one piece of medical evidence, such as those trained on full EHR data. What the medical AI research is comparatively silent about, is how these very capable models *should* (in the epistemic sense) be used by clinicians as decision-aids. Indeed, almost all studies don't consider the question of what the accuracy of such a hybrid human-AI team would be, let alone how the accuracy would vary based on different strategies a clinician could use to rely on a model. In the absence of such empirical evidence, our task here will be to generate theoretical hypotheses of how a clinician might rely on a model, by considering the disparate fields of social epistemology and human factors.

04.3 Disagreement and AI-reliance

The field of social epistemology is particularly well-suited to contribute to the question of how such well-performing models should be used, in its discussion of epistemic disagreement. In this section, we shall look at (a) a brief overview of the field of epistemic disagreement, and (b) how it might be relevant to reliance on machine learning models in healthcare. Given the state of active philosophical disagreement over these topics, our goal here will not be to distill conclusive and actionable insights on how clinicians should interact with models of various accuracy. Rather, it will be to outline the parts of the disagreement discussion so far that are relevant to our question of medical AI reliance, identify which sub-questions in particular need further answering, and what are the conditions that a theory that specifies responses to disagreement would need to satisfy for it to be useful. Given the many open questions we shall encounter, as well as our practical purposes, one might very well ask the question of why we should turn to a debate that remains so unresolved. To this, my response would be that we ought to do so for two reasons. First, these fields concern themselves with a question that lies at the very heart of our concerns about how clinicians ought to rely on AI models: how should any agent revise his beliefs in light of information about his own reliability? While the less-than-conclusive resolutions in the field are not ideal, it still merits a closer look given it's extremely relevant topicality. Second, by identifying conditions that theories of disagreement need to satisfy to be relevant to the AI-reliance problem, we might narrow the disagreement on how disagreement should be managed, at least for our purposes.

04.3.1 Epistemic Disagreement

The epistemology of disagreement focuses on the following question: how (if at all) should we revise our beliefs when we find ourselves disagreeing with others? The other agents that we find ourselves disagreeing with can fall into three categories – our epistemic inferiors, our epistemic peers, and our epistemic superiors. Traditionally, the field of disagreement has concerned itself with the circumstances of epistemic *peer* disagreement, as disagreement with epistemic inferiors and superiors is seen to be relatively unproblematic – epistemic inferiors should not make us revise our beliefs while superiors should be trusted and deferred to. Peer disagreement, on the other hand, has been much more divisive.

First though, how might we identify epistemic peers, inferiors, or superiors? While initial formulations of epistemic peerhood required the twin conditions of a peer having an equal access to the relevant evidence, and an equally capable disposition to respond to said evidence (Kelly 2005, Lougheed 2020),' more recent accounts have distinguished between *rationality* peers and *accuracy* peers (Christensen 2016). Rationality peers are those who're peers in the aforementioned sense (of the second condition), such that they're equally likely to be rational in their response to a set of evidence. Accuracy peers, on the other hand, are those who're equally likely to form accurate (true) beliefs about the particular set of evidence – an agent might thus be your accuracy peer without being your rationality peer, as they may commit more errors of rationality than you but end up having the same accuracy, or vice versa. One can extend these two conceptions of epistemic peerhood to similarly define epistemic superiors (and inferiors), as those who're either more (or less) rational in their response to a body of evidence, or those who're more (or less) likely to produce accurate beliefs given a body of evidence.

With these tentative conceptions of relative epistemic competence, we can now consider how we ought to act under circumstances of disagreement. For the circumstance of epistemic peer disagreement, there have been three broad responses provided in the literature. One response has been that peer disagreement requires *conciliation* – that is, it requires us to revise our beliefs in some way. A strong version of the conciliatory approach is the equal weight view (EWV), which states that when disagreeing with a peer, we are rationally required to split the difference, giving

equal weight to both our own view as well as our disagreement partner's (Elga 2007). Thus, if one clinician considers a patient and all associated evidence and forms a credence of 0.3 that the patient has disease X, and her peer considers the same evidence and forms a credence of 0.7, then both clinicians operating under EWV ought to revise their credence to 0.5. More moderate versions of the conciliatory approach require *some* belief revision, though might not require it to the point of giving one's peer's view weight equal to one's own. On the other hand, *steadfast* views on peer disagreement claim that the mere fact of disagreement doesn't require a belief revision on rational grounds – one can continue to be steadfast in one's initial belief without failing to be rational (Kelly 2005).**Error! Bookmark not defined.**

Beyond conciliatory and steadfast responses, there is also a third response – that response to peer disagreement should neither be strictly conciliatory nor steadfast. It should instead follow an overarching principle that sometimes requires conciliatory behaviour, and at other times requires one to be steadfast. The *total evidence view* (TEV) is one approach that takes this form, arguing that in any given situation the evidence of disagreement needs to be considered alongside the original evidence used to arrive at one's pre-disagreement view (Kelly 2010). Depending on whether this original evidence or the disagreement evidence is more compelling, TEV would require the response to resemble either a steadfast one or a conciliatory one (respectively). Another approach of this sort is a Bayesian one, where the appropriate response to disagreement is to just use standard Bayesian updating procedures (or some variant that is less cognitively demanding) to revise one's beliefs (Mulligan 2019, Easwaran et al. 2016).' Such updating procedures will then arguably provide steadfast norms and conciliatory norms in cases where each would be intuitively suitable. One additional benefit of a Bayesian approach is that it is also argued to be generalizable to epistemic superiors and inferiors (Mulligan 2019, Goldman 2001).**Error! Bookmark not defined.**' Another benefit to the Bayesian approach has been that in contrast to the total evidence view, its proponents have outlined actionable and mathematically precise updating procedures rather than more ambiguous, high-level principles (as we shall see in the next sub-section).[17]

In general, epistemic disagreement is taken to be a particular form of *higher-order evidence*, that is, evidence about one's own reliability in making judgments about any particular subject-matter (Christensen 2016). In comparison, first-order evidence is evidence that we might make an initial judgment based on (without assessing our own reliability), such as a clinician assessing a patient to judge whether they have disease X.[18] When we disagree with someone, the question is whether that disagreement itself is higher-order evidence about our own reliability in responding to first-order evidence. In the case of peer disagreement, steadfast views say no (or that the evidence is weak or misleading), conciliatory views say yes (to differing extents), and total evidence and Bayesian views say that it sometimes is and it sometimes isn't.

Distilling the above discussion, the procedure for managing disagreement has three stages. In the first stage, the agent arrives at a belief based purely on the first-order evidence, in isolation to any higher-order evidence. In the second stage, the agent identifies the exact circumstances of the disagreement or higher-order evidence she is exposed to. For instance, the agent identifies whether there is disagreement (or agreement), and what the reliability of her disagreement partner is vis-à-vis herself – in other words, whether she is disagreeing with a peer, a superior, or

[17] One might argue that given the level of generality that TEV takes, the Bayesian approach might actually be a specific instantiation of the TEV approach. As we shall see in the next section, this isn't quite right, at least in terms of how the Bayesian approach has been articulated so far in the literature.

[18] It has been argued that there may not be a very precise and clean way of assessing which evidence is first-order and which is higher-order, even if we may for the most part use the terms fruitfully. For reasons of scope, I will not pursue this further here.

an inferior. Finally, in the third stage, the agent applies the relevant disagreement norm or epistemic approach for belief revision, depending on the circumstances of the disagreement, to her initial belief, so as to arrive at her final belief. The different approaches to disagreement outlined above vary based on the requirements they specify for the second and third stage.

04.3.2 Reliance on AI Models

As might have been guessed from our discussion so far, the consideration of epistemic disagreement and higher-order evidence is uniquely suited to the question of appropriate epistemic reliance on AI models. The latter question can be posed as follows:

> How should a clinician with epistemic competence *c* revise their beliefs about a patient when they find themselves agreeing or disagreeing with the outputs of a model with competence *m*, in cases where the model and the clinician have access to the same information?

Let us unpack one part of this further before we proceed. It is unlikely that a clinician will always have exactly the same information about the patient that the model has access to – in most cases, models are prebuilt to take in certain categories of information, and a clinician will usually have access to more. For instance, a model that screens for lung cancer through computed tomography (CT) scans will do so on the same CT scans that a doctor has access to,**Error! Bookmark not defined.** but the doctor will also have access to a patient's family history of cancer, lifestyle habits, etc. The difference between information that the doctor has access to and information that the model has access to might be lower when we consider models that are trained to classify diseases based on extensive EHR charts rather than a single radiological image, but that difference will still persist. However, this does not make the question articulated above inapplicable to the realities of medical practice, as we might employ heuristics to overcome this. For instance, a clinician could only compare her beliefs about a patient with the model's when those beliefs have been derived from the same information that the model possesses. Additional information can then be factored in subsequently.

Further, there are various clinically-relevant considerations that would also feature in live, medical decision-making. For instance, clinicians are interested not just in the presence of the disease, but also its severity, and its rate of progression. There are also other decision-theoretic factors that are considered in a clinical encounter – whether avoiding a false positive or a false negative is of greater importance, whether there are treatment options that can treat for multiple different potential conditions, and whether the treatment of certain diseases for a patient, while less likely to be present than other diseases, should still be prioritized given the relative severity and aggressive progression. Additionally, in difficult/borderline cases, doctors don't always rely on their unilateral judgments, but often make a decision as part of a multidisciplinary team – perhaps a team that has been assembled comprising expert colleagues. These factors complicate the analysis beyond the relatively straightforward formulation in this chapter of what the appropriate credence for proposition Y about disease Z given evidence X is, but are still crucial for actual medical decision-making. However, for the purposes of this chapter and the goals of making progress on the theoretical problem, we will abstract away these complications here.

That being said, we may now explore how discussions of epistemic disagreement might shed light on the above question. In what follows, we will reflect on the nature of decision-making and reasoning that clinical reliance on AI requires, and propose some conditions that would allow us to narrow the set of approaches that have philosophically been proposed to answer to the fact of epistemic disagreement. We might start with a very simple condition:

C0: The epistemic approach needs to be 'correct', or at least 'plausible'.

This is a condition that we have implicitly already embraced in this chapter, as the approaches to managing disagreement outlined in the discussion in section III.1 that we have limited our examination to, and the ones we will discuss moving on, are already 'live contenders' in the field. This is so that we only take as approaches to reliance on AI and managing disagreement those approaches that have some broad acceptance in the field as potentially epistemically appropriate ways of managing disagreement.

Moving on and borrowing the earlier example, we can imagine a scenario where a clinician is examining a patient's CT scan to screen for lung cancer, and arrives at the conclusion that the patient is not likely to have lung cancer. On the other hand, the model the clinician has access to considers lung cancer likely. Depending on their relative competences c and m, a clinician disagreement with the model's outputs might take the form of disagreement with a peer or a superior. This leads us to our first condition that any approach to disagreement or higher-order evidence needs to satisfy, to bear on our question of appropriate AI reliance:

> **C1:** The epistemic approach needs to specify appropriate procedures for belief-revision in interaction not just with epistemic peers, but also epistemic superiors, and perhaps epistemic inferiors.[19]

This can be done in two ways. One way is to provide separate and non-continuous guidelines for engagement with peers and superiors (and perhaps inferiors). One representative approach could be with peer disagreement simply being dealt with using Equal Weight View (or more moderate conciliatory views, or steadfast views, or TEV), while disagreement with superiors is dealt with simply by deferring to the superior's recommendation fully. Alternatively, a second way is to provide one disagreement norm that can handle cases of disagreement with all manner of epistemic agents – peers, superiors, and even inferiors. We might imagine an updated version of TEV, TEV^{+}, that applies to disagreement with epistemic agents of varying competency, by comparing in each case the weight of both the first-order evidence and higher-order evidence. For instance, the weight of higher-order evidence will be highest for disagreement with superiors, followed by peers, followed by inferiors. Bayesian approaches would also work in this way, by allowing for the specification of different likelihoods of competence depending on the competency of the disagreement partner (more on this later). While it would seem at this point that a more parsimonious approach, like the latter one of one disagreement norm for all cases, would be preferable to a more complex and discontinuous one, we need not make a judgment on this now – we can return to this after considering all other conditions.

The second condition comes from the fact that ultimately, the question of how to rely on AI models is a fundamentally practical one. Any epistemic approach to disagreement or higher-order evidence, then, needs to be sufficiently action-guiding:

> **C2:** The epistemic approach needs to be sufficiently action-guiding in clinical settings.

The epistemic approach needs to be action-guiding in both of the relevant stages outlined in the previous sub-section: in determining the circumstances of epistemic disagreement or higher-order evidence that a clinician might find themselves in, and in identifying how the clinician

[19] The question of whether an approach for the use of AI systems which are epistemic inferiors is relevant depends on whether relying on epistemically inferior models (for instance, as measured by accuracy) can lead to better epistemic outcomes, not just theoretically but also practically. I will not address this question here beyond a few preliminary remarks throughout the chapter.

should revise his belief based on those circumstances. We shall look at how previously articulated approaches fare when it comes to being action-guiding for each of these stages in turn.

First, we need to be able to determine the circumstances of epistemic disagreement or higher-order evidence in a way that is action-guiding. This would require us to be able to assess the reliability of our disagreement partner vis-à-vis our own. As we've discussed, the competence of disagreement partners (as peers, superiors, or inferiors) can be established in terms of their rationality or their accuracy (Christensen 2016).**Error! Bookmark not defined.** Given the somewhat opaque nature of deep learning models used, it is difficult to identify how *rationally* they're responding to the evidence compared to human practitioners, not to mention that assessing this would also first require us to gain consensus on what rationality requires when an agent responds to their first-order evidence (and not just generally, but in the medical field in question). Fortunately, identifying reliability through accuracy is much easier in our case, given that we have access to the model's past performance in terms of its specificity, sensitivity, and AUROC scores.[20] We also have such scores for the human practitioners that these models were tested against, which we can take as a proxy for the reliability of the clinician. While this might be an imperfect proxy, we can make it more reliable by categorizing human scores by level of seniority, as some studies have done (Liang et al. 2019).**Error! Bookmark not defined.**

Settling on establishing relative competence through the relative accuracy of the model rather than its relative rationality is a step forward, but not enough if we want to establish peerhood, inferiority, or superiority. For that, we need to be able to specify thresholds for sensitivity, specificity, AUROC scores or other accuracy measures to demarcate what performance is considered inferior, superior, or at the level of a peer. Consider for instance the most accurate scores mentioned in section II above, with the pooled average for most accurate models being a sensitivity of 87% and a specificity of 92.5%, compared to the pooled average for the most accurate human practitioners with a sensitivity of 86.4% and specificity of 90.5%. Would this indicate that the models are epistemically superior to humans and should be practically treated as such, even though the difference in the sensitivity and specificity values are just 0.6% and 2% respectively? Or should they be treated as peers? Approaches to disagreement would require a clean delineation of inferiors, peers, and superiors if they advocate distinct strategies for belief-revision for each type of disagreement partner. It would seem much too strict, and perhaps even arbitrary, to insist that such accuracy values be exactly identical (or identical to the nearest whole percentage point) to qualify for peerhood, with any deviation landing the model either as an inferior or a peer. How then would such thresholds be established? It would be preferable if the accuracy measures could just stand on their own, without needing to indicate different classes of relative competencies. Such a requirement would favour approaches which do not explicitly invoke the categories of epistemic peerhood or superiority or inferiority, such as TEV^{+} or the Bayesian approach (which we shall see in further detail shortly).

We can now consider how well the various approaches to disagreement satisfy C2 for the second stage of managing disagreement – providing rules for belief-revision. EVW, for instance, satisfies this condition sometimes. If we consider clinicians as dealing with graded beliefs, such that their judgments express credences (for instance, a 0.7 credence in the patient having disease X), EWV just requires that in cases of peer disagreement with a model a clinician average their own

[20] Going the accuracy route also allows us to potentially sidestep the issue of *uniqueness* in cases of rational peer disagreement – whether there is rationally only a single belief state that agents can have to a given body of evidence, or whether rationality permits for multiple states to co-exist. For more on this, see endnote 14.

credence with the model's (which would be the model's quantified uncertainty, for instance).[21] Depending on whether the resulting credence passes the diagnostic threshold the clinician favours, they can then proceed to act as if the disease is present or not. When dealing with full beliefs however, such that a clinician either believes a disease to be present or not (credence of 1 or 0), EVW requires suspension of belief in cases of peer disagreement (Ferrari et al. 2020). If a clinician thus disagrees with a model, with one agent saying disease X is present and the other disagreeing, it is unclear how suspension of belief would be practically helpful – how is a clinician to proceed having suspended his belief on the patient having disease X? Perhaps the thing to do would be to look for more information as a tie-breaker, but in time-sensitive settings (such as critical care), this may not be sufficiently action-guiding.[22]

TEV$^+$, on the other hand, does not satisfy C2. TEV$^+$ merely says that in cases of disagreement the agent should weigh his first-order evidence against his higher-order evidence to determine whether a conciliatory or steadfast response is merited – however, it provides no account of how such a weighing is to be done, and under what circumstances do either type of evidence win out over the other. There are some intuitive examples provided, but no comprehensive and actionable account is available (Kelly 2010).**Error! Bookmark not defined.** Similarly, more general conciliatory approaches also need not satisfy C2. One approach proposed is the 'simple thermometer model', which states that "in cases where the agent has reached an initial credence in C, and then gets some higher-order evidence, her final credence in C should match her independent hypothetical credence in C" (Christensen 2016).**Error! Bookmark not defined.** While this captures the epistemic concepts that need to come into play for conciliation, it doesn't provide an account of how the independent hypothetical credence should be arrived at. Similarly for steadfast approaches such as the 'right reasons' view, which states that under disagreement, the party who in fact reasoned correctly from the evidence can rationally maintain her belief (Ferrari et al. 2020)**Error! Bookmark not defined.** – while this sounds plausible, it is hard for a clinician to know whether or not she did in fact reason correctly when she disagrees with the model.

One approach that seems to fare quite well when it comes to C2 is the Bayesian approach. The Bayesian approach to disagreement has been argued to be captured by Bayesian conditionalization (BC) in the following way (Mulligan 2019, Easwaran et al. 2016):**Error! Bookmark not defined.** **Error! Bookmark not defined.**

$$c_1' = \Pr(X|c_2) = \frac{\Pr(c_2|X)\Pr(X)}{\Pr(c_2|X)\Pr(X) + \Pr(c_2|{\sim}X)\Pr({\sim}X)}$$

Here, $\Pr(X)$ is the clinician's initial credence (the prior) that the patient has disease X (where $\Pr(X) = c_1$), and c_2 is the model's confidence in the same. c_1' is thus the clinician's final belief (the posterior) about X, after conditionalizing on the model's output c_2. Traditionally, one of the biggest impediments to using Bayesian conditionalization for human-human disagreement has been that $\Pr(c_2|X)$ and $\Pr(c_2|{\sim}X)$, which amount to the values for the sensitivity and (1-specificity) of the disagreement partner for the specific value taken on by c_2, are not accessible for other humans that we disagree with. It is difficult to estimate what the sensitivity and

[21] The analysis here would be complicated if the model can't express its own recommendations in similarly graded terms, though we do know that medical AI models can be built to express their confidence in their recommendation as a percentage.

[22] Of course, there's no strict reason why our epistemic disagreement approach to model reliance needs to be applicable in *all* cases of model reliance, beyond such a comprehensive approach being theoretically satisfying. Given our pragmatic commitment, an approach to reliance that works in all non-time-sensitive settings would still be very useful.

specificity of the people we disagree with are. However, what makes the Bayesian approach suited for our purposes is that these values are quite easily available for any model (as seen in section II), and thus can be represented in BC.

BC is thus eminently action-guiding. All that is needed is the clinician's prior, which is their assessment of the first-order evidence, and the model's sensitivity and specificity, which are readily available. Furthermore, this approach also allows for reliance on models of varying accuracy. Models which have lower accuracy will have a lower value for $\Pr(c_2|X)$ and a higher value for $\Pr(c_2|{\sim}X)$ than models with higher accuracy, and this will be reflected in the total magnitude of the update (the difference between c_1' and $c_1/\Pr(X)$).

However, this comes at a cost – BC violates C1. To see why, consider that C1 requires epistemic procedures for belief-revision in cases of epistemic peers, superiors, and inferiors. What BC provides is a procedure for belief-revision in the case of models of varying epistemic capability, but not varying epistemic capability *relative* to the clinician. If the clinician were to rely on BC, she would revise her belief taking into account only the model's reliability but not her own. For the latter, BC1 would have to include information about the clinician's sensitivity and specificity in some way, which it does not. Essentially, BC mistakenly treats evidence of disagreement as first-order evidence, where the clinician's (or any general operator's) own reliability isn't considered, rather than higher-order evidence.

To see why this is problematic, consider three clinicians of varying competence, as measured by their sensitivity and specificity when diagnosing for disease X. Say clinician 1 has sensitivity and specificity of 0.7, clinician 2 has sensitivity and specificity of 0.8, and clinician 3 has sensitivity and specificity of 0.9. Assume further that all of them have the same credence ($c_1/\Pr(X)$) for a patient having disease X based on the medical facts of the matter (first-order evidence). Imagine them relying on a model with a sensitivity and specificity of 0.75. If they were to follow BC, given that BC will only take into account the sensitivity and specificity of the model and the clinician's original credence, all three clinicians will end up with the same value for c_1'. This is clearly extremely unintuitive and difficult to accept, as it advocates for instance that a very competent clinician should treat disagreement with a relatively inferior model as seriously as a much less competent clinician should treat disagreement with a relatively superior model (to the extent that such categories of epistemic capabilities are determined through agents' sensitivity and specificity).

Unfortunately, there isn't a quick fix to this. We can try to update Bayesian conditionalization as follows (to yield BC*):

$$c_1' = \Pr(X|c_1, c_2) = \frac{\Pr(c_1, c_2|X)\Pr(X)}{\Pr(c_1, c_2|X)\Pr(X) + \Pr\left(c_1, c_{2_2}\middle|{\sim}X\right)\Pr({\sim}X)}$$

Here, the posterior credence c_1' is conditionalized on both c_1 and c_2, allowing us to factor in the reliability (through sensitivity and specificity) of both the clinician and the model through $\Pr(c_1, c_2|X)$ and $\Pr\left(c_1, c_{2_2}\middle|{\sim}X\right)$. However, this raises a new problem now – what value do we assign now for $\Pr(X)$ and $\Pr({\sim}X)$? Initially in BC, the values we would have assigned to it were c_1 and $(1 - c_1)$, but those are now being invoked for sensitivity and specificity. We essentially have no prior to plug into BC*, which makes it not at all action-guiding and so in violation of C2. There are some avenues that we might explore for a fix to this. For instance, we might plug in a value for $\Pr(X)$ by relying on some form of principle of indifference (Eva 2019) - distributing credences equally over possibilities in the absence of evidence – to distribute our

credences over the possibility of disease X and all other possibilities. Alternatively, we might try to run various simulations with different values for $\Pr(X)$ to see if there is convergence on a particular value that bears out our disagreement intuitions. Either way, further work needs to be done to ensure that a Bayesian approach is compatible with both C1 and C2.

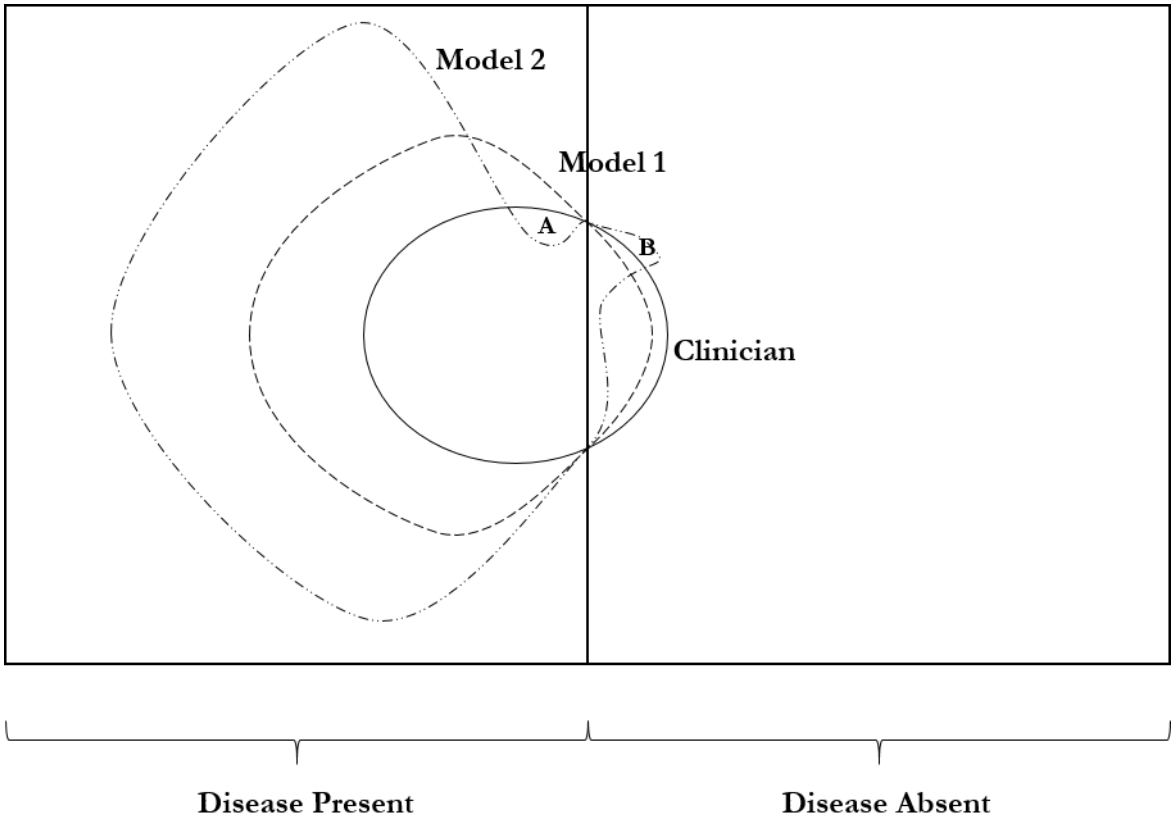
There is a further problem that is raised if we are to add a third condition for epistemic approaches to managing disagreement:

> **C3:** The epistemic approach needs to apply to graded beliefs rather than just all-or-nothing beliefs.

Here, by graded beliefs, we mean beliefs about any proposition P (such as presence of disease X in a patient) that aren't just of the form P or ~P, but admit of degrees between the two. Epistemic approaches to managing disagreement should thus be able to deal with clinicians not just having a belief that the disease is present or not, but having beliefs of the sort 'it is likely that the disease is present', or even more granularly, 'my credence in the disease being present is 0.7'. Considering that doctors are encouraged to use probability to quantify uncertainty and use Bayes theorem as a way of forming beliefs about first-order evidence (for instance tests with a certain sensitivity and specificity) (Sox et al. 2013) – thus making it likely that clinicians have graded rather than all-or-nothing beliefs – C3 is a necessary requirement.

C3 doesn't seem like a particularly difficult condition to meet if one meets C1 and C2, considering that the language of graded beliefs is ubiquitous in the disagreement literature. However, as we shall see, it poses a specific problem to disagreement strategies that rely solely on values of false positives, false negatives, true positives and true negatives (and all their derivations, including sensitivity and specificity) to estimate competence. This is a problem that has been so far underexplored in the literature. Considering that almost all studies developing clinical AI models assess their competence using these measures, the following problem becomes a salient one.

Let us imagine a clinician having the option to rely on two models, model 1 and model 2. The capabilities of the clinician and the two models, in terms of their track record, are displayed in the image below, with each encircled area being the patients that are selected as possessing the disease. The left part of the image includes those patients who *actually* have the disease, while the right part includes those who *actually* don't. As can be seen, both models as well as the clinician incur false positives (instances from the 'disease absent' section that are included inside the decision boundary) as well as false negatives (instances from the 'disease present' section that are not included inside the decision boundary) when they pick out patients as having the disease.



What is obvious from the image is that both models have superior accuracy to the clinician, as they yield comparatively fewer false positive and false negatives. However, we might describe model 1 as *strictly superior* to the clinician, while model 2 is *non-strictly superior*. What that means is that model 1 picks out all the true positives and true negatives that the clinician picks out, as well as more on top of that. If the clinician were to find herself disagreeing with model 1, their track record doesn't pick out any instance where the model has been wrong and the clinician right. In comparison, there are cases where model 2 has been wrong and the clinician right (regions A and B in the image), even though model 2 is superior in accuracy to the clinician. Intuitively, it would then seem that when the clinician disagrees with model 1, she should defer completely to model 1's output given the completely one-sided track record of disagreement between them. However, if the clinician disagrees with model 2, the deference shouldn't be complete – after all, there have been cases where model 2 has been wrong and the clinician right. What makes this scenario especially interesting is that model 2 itself is superior to model 1 in terms of accuracy (judged from true and false positives and true and false negatives, and thus sensitivity and specificity). Thus, while model 2 is superior to model 1, the clinician seems to be intuitively required to defer only partially to model 2 but completely to model 1. [23]

Now, if the clinician were only operating based on all-or-nothing beliefs, this peculiarity wouldn't really matter. Given that both models are still superior to the clinician, past track records would have to indicate that when it comes to disagreement either model is more likely to be right than the clinician. Circumstances of disagreement would thus require the clinician to switch her diagnosis from disease absent to disease present, or vice versa. However, if the clinician were

[23] It's important to note here that neither of these cases are that of *peer* disagreement, such that the steadfast approach competes with the conciliatory approach advocated here. Models 1 and 2 are both epistemically superior to the clinician, measured in terms of true and false positives and true and false negatives. Intuitively, it seems that the clinician has to defer to both – the comments here just outline that this deference has to be complete in the first case and partial in the second.

operating using graded beliefs, which is likely, disagreement with model 1 and 2 would produce different responses. Disagreement with model 1 would require the clinician to update her credence in the disease being present to match model 1's credence. Disagreement with model 2 would require the clinician to update her credence in the direction of model 2's credence, probably more than half-way given that model 2 is epistemically superior, but not all the way like for model 1. Assuming model 1 and 2 output the same credence in the disease being present, the clinician updating for disagreement would update more for model 1 than model 2, even though model 2 is superior to model 1. Depending on the confidence threshold for clinical action, this might lead to different actions being pursued by the clinician.[24]

C3 thus imposes a requirement that cannot be fulfilled if the disagreement norm assesses competence solely through true and false positives, and true and false negatives (and their derivatives like sensitivity, specificity, AUROC, etc.). This disadvantages, for instance, the Bayesian approach outlined above even if BC* can be resolved adequately, since it relies purely on sensitivity and specificity to represent reliability of epistemic agents. For BC* to be successful, it will have to quantify and represent further the extent to which epistemic agents are *strictly* and *non-strictly* superior to each other, beyond just mere superiority.

Importantly, C3 also imposes a condition not just on epistemic approaches to disagreement, but also on model developers. It is no longer sufficient to measure model (and comparative human) performance purely through sensitivity, specificity, and their derivatives, as has overwhelmingly been done so far. While those measures might suffice to represent model and human reliability in isolation, epistemically appropriate human-AI interaction will require the measurement of the disagreement track-record between the model and the human practitioners (when both have indicated diverging opinions on cases). This track record can then be assessed alongside differences in sensitivity and specificity to update for the disagreement. Designing studies to be sensitive to clinical beliefs being graded rather than all or nothing could also be furthered by ensuring that studies calculate sensitivity and specificity values for both models and humans not just for a classification of a disease being present or not, but for varying levels of confidence in the disease being present, perhaps at the level of confidence intervals (Brown et al. 2003) (although this might be harder to do for humans for logistical reasons).

Two final points on disagreement norms are worth mentioning. First, an implicit assumption in all our analysis so far is that we should treat track records of sensitivity and specificity as equal evidence for reliability of models and human practitioners. That is, if a model's track record shows a sensitivity and specificity of 0.8, and a human's track record shows the same, they can be treated as equally reliable. However, this assumption might be problematic as model accuracy might be less stable than that of human practitioners. The meta-analysis discussed previously also showed that when we compare the pooled average of medical AI performance that was just validated internally to performance that was validated out-of-sample, sensitivity and specificity drop by 3.1% and 0.8% respectively (Liu et al. 2019).**Error! Bookmark not defined.** This suggested instability though, might just be an argument for ensuring that only accuracy values that come from models extensively validated out-of-sample are accepted for our purposes.

[24] Here, we've assumed that as long as any model is strictly superior to an agent, deference to it should be greater than any other model that is non-strictly superior, even if the overall accuracy of say Model 2 is greater than that of Model 1. However, this is a point that can be debated. Consider if Model 1 is strictly superior to the clinician, but barely. Whereas Model 2 is non-strictly superior, but massively (it perhaps picks out almost the entire space of 'Disease Present' in its decision boundary. Under such conditions, it would be difficult to maintain that in such a case of disagreement, one should update more for Model 1 than Model 2. Perhaps then the point around featuring the 'strictness' of superiority can be softened a little bit – these considerations are influential, but not decisive, and an all-things-considered approach might give weight to both the level of strictness of the superiority, and the extent of the non-strict superiority.

Despite this, the general tendency of deep learning models to 'overfit' to their training data is troubling, where such overfitting leads to seemingly high model performance that is brittle because it comes from memorizing the training data rather than learning from it (Ravi et al. 2017). Whether or not such overfitting should be considered problematic depends on (1) whether the model has undergone extensive external validation prior to use as a clinical decision-aid, and (2) whether the performance of human practitioners is similarly brittle out-of-sample. If the answer to these questions is 'no', this raises concerns for whether model sensitivity and specificity is equal evidence of reliability when compared with human sensitivity and specificity.

Second, while we've primarily concerned ourselves with the phenomenon of disagreement, some argue that norms for *agreement* should also feature when considering belief-revision for higher-order evidence. As Easwaran and colleagues argue, we might consider (at least peer) agreement as yielding 'synergy', where agreement should raise an individual's credences higher than any of his agreement partners, and a failure for approaches to belief-revision to yield synergy is a strike against them (effectively making synergy an additional condition for such epistemic approaches) (Easwaran et al. 2016).**Error! Bookmark not defined.** Claims along these lines have also borne out empirically, where 'extremizing' (increasing credence when faced with agreement/consensus) has led to forecasters performing better (Tetlock 2015). While I will not pursue this further here, this is an important feature of belief-revision that needs to be factored in.

04.4 Conclusion

In this chapter, we have considered the issue of how clinicians might respond to clinical AI decision-support systems in cases where epistemic judgments on medical matters of fact (such as diagnoses, or treatment recommendations) might be conflicting between the clinician and the AI. We have seen that such an issue is not merely a hypothetical one, given the advances in medical AI that have yielded comparative levels of accuracy between humans and models. We have also seen that the field of epistemic disagreement is a way of framing this problem that can allow us to fruitfully bring existing philosophical tools to bear on resolving it.

In doing so, I have argued that for approaches to epistemic disagreement to be relevant to the problem of AI-reliance, they will need to meet certain conditions. To re-iterate, these conditions are

> **C0:** The epistemic approach needs to be 'correct', or at least 'plausible'.
>
> **C1:** The epistemic approach needs to specify appropriate procedures for belief-revision in interaction not just with epistemic peers, but also epistemic superiors, and perhaps epistemic inferiors.
>
> **C2:** The epistemic approach needs to be sufficiently action-guiding in clinical settings.
>
> **C3:** The epistemic approach needs to apply to graded beliefs rather than just all-or-nothing beliefs.
>
> **(C5:** Synergy, possibly)[25]

[25] There have also been other conditions outlined in the literature which we haven't discussed as much here, such the disagreement norm being extendable to disagreement with multiple agents and the disagreement norm accommodating the degree of dependence between epistemic agents. For a discussion of these see endnote 22. Beyond reasons of space, we haven't discussed these two conditions here as it seems simpler to consider reliance on just one AI model to start with, and also some of the disagreement norms discussed (E.g. BC and BC*) already satisfy the independence condition.

While discussions on epistemic disagreement feature multiple competing approaches and positions, once we accommodate for these conditions the competition will narrow significantly. We've also raised desiderata for model developers to satisfy if the AI-reliance problem is to be modeled as one of epistemic disagreement – namely, that specifying model and human sensitivity and specificity in isolation is insufficient, and the disagreement track record between the two needs to also be articulated.[26] Through this discussion we can now start to think productively about modeling AI-reliance as an epistemic disagreement problem. The end-state to this project, if I were to imagine it, would reflect a simple way of managing disagreement between models and clinicians that would be similar to how clinicians are taught to use Bayes rule to update for test results (first-order evidence). It would also give model developers a standardized set of performance metrics to report if their tool is meant to be used in clinical settings. Of course, this approach can extend easily to other domains where AI-reliance fits the same patterns – perhaps intelligence, finance, etc (other ethical issues not withstanding, as we shall see in the next chapter).

Interestingly, this allows us to evaluate appropriate reliance depending just on the accuracy of the model's outputs, without invoking it's mechanisms, something which can yield interesting insights about their accountable use. The approach painted in this section is a start towards doctors using deep-learning models in medicine justifiably despite their opacity. In the next chapter (Chapter 5), we shall see how the problematisation of shared decision-making with black-box systems can similarly be defused.

[26] This need not require studies to actually have human practitioners interact with the model. All that might be needed is to determine for any case (e.g. a particular medical image), what classification did the model make, what did the human practitioner made, and who was right.

05. Does Reliance On Clinical AI Compromise Shared Decision Making?

05.1 Introduction

While AI models are being used across a wide variety of medical tasks, importantly, some of these tasks involve recommending interventions for patients who are facing preference sensitive choices where the choice involves two or more options that feature trade-offs and are equally reasonable, or where evidence of effectiveness is unclear, and/or where different patient values can impact the choice significantly (Entwistle et al. 2016, NHS 2021, Whitney et al. 2004).

The development of AI systems for such tasks have raised questions about whether their use in clinical settings might compromise the ideal of shared decision making (SDM). This has been argued especially due to the black-box nature of deep learning models, where explanations for model recommendations are not humanly understandable, and thus inaccessible to clinicians, patients, and even model developers. More specifically, two arguments have been made that the use of black-box AI systems compromises SDM:

(1) Black-box AI models do not provide us with medical explanations for their recommendations, which are critical for SDM

(2) Black-box AI models risk excluding important patient values and preferences from important clinical decisions, which diminishes patient autonomy and undermines SDM.

In this chapter, we will consider each of these questions in turn. I will ultimately argue for the following positions: The inability to access medical explanations when relying on black-box AI models in clinical settings doesn't compromise patient autonomy in ways traditionally problematized. However, it may compromise SDM through values other than patient autonomy – to get a definite answer on this, we need further empirical work. There are very special cases where black-box models definitely will compromise SDM, but the burden of proof to demonstrate that is significant.

Given the broad argument of this thesis, this chapter will be the second chapter to take a specific problem that has been posited to arise from the use of black-box, deep-learning models in medicine, and provide an argument for how this problem can be addressed – this will serve as an illustration of explanations (through interpretable models) not being necessary for the ethical use of such models. It will thus continue the thrust of the broader argument from the previous chapter, which similarly showed that explanations were unnecessary for appropriate epistemic reliance on such models by healthcare practitioners.

For the sake of fruitfully limiting our focus, we are also only concerning ourselves with clinical AI models that are decision-aids for clinicians rather than systems which are autonomously deployed. This will allow us to simplify the discussion for what we might find deployed currently and in the short-term. No doubt, AI models will outperform clinicians in the future, and this might perhaps cause them to be deployed autonomously even for preference-sensitive decision making. Regardless, we will not consider them here apart from noting that many of the conclusions reached here for AI decision-aids will no longer apply.

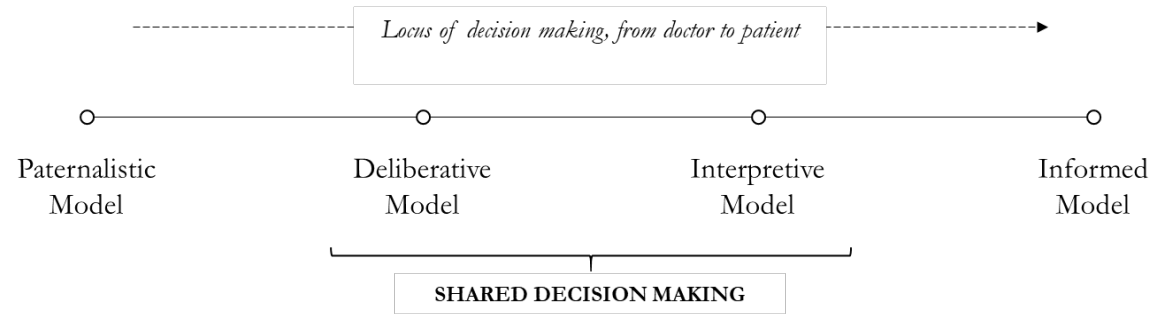
The structure of the chapter will be as follows. Chapter 1.1.2 and 1.1.3 already cover an explanation of the different kinds of models and how their black-box nature might arise – readers might refer back to those sections for a refresher. In section 5.2, I will summarize the literature on SDM, canvassing the history of thinking on the topic as well as the various different conceptions of SDM that have been raised. I will also cover how these conceptions of SDM have been operationalized through the use of SDM measures in clinical settings. Section 5.3 will contain the meat of the philosophical analysis – I will entertain each of the two problematizations of black-box AI models for SDM, and argue for the positions I have mentioned earlier. In section 5.4, I will briefly consider some more speculative and long-term impacts of opaque processes on SDM, and I will end with a summary of the discussion and next steps in section 6.

With this, we can now move to examining how deep-learning models have been argued to undermine shared-decision making.

05.2 Origins and Elements of Shared Decision Making

The last few decades have seen a shift in the locus of decision making in clinical encounters. The traditional paternalistic model of the doctor-patient relationship, where the doctor dictates diagnostic procedure and prescribes treatments unilaterally to the patient (with an expectation of patient obedience) has given way to a shared model of decision making.

The term 'shared decision making' was first popularized in a report by the President's Commission on the ethical and legal implications of informed consent (President's Commission 1982), in response to perceived limitations in the existing legal practice of informed consent. Existing practice merely incentivized the disclosure of relevant medical information by the doctor instead of incentivizing the facilitation of patient understanding and participation. The practice of informed consent was thus argued to "fall short of the law's professed commitment to the value of self-determination".**Error! Bookmark not defined.** In response to this failing of legal practice, several ethical models of the doctor-patient relationship have been defined and discussed over the subsequent years (President's Commission 1982, Veatch 1972, Emanuel and Emanuel 1992, Charles et al. 1982, Charles et al. 1999).**Error! Bookmark not defined.** [Error! Bookmark not defined.]



These models are primarily characterized based on the direction and type of information exchange that occurs, the parties involved in deliberation, and the parties that decide on the ultimate decision on diagnostic procedure or treatment (Charles et al. 1999). At one end of the spectrum lies the paternalistic model (also known as the 'priestly' model), where medical information is conveyed to the patient, no values or preferences are solicited from the patient, and the clinician deliberates and decides on the best course for the patient both medically and in terms of the values that guide the decision. Medical information here refers solely to medical matters of fact – such as the clinical presentation of the patient, any test results, as well as what that might suggest about the underlying biological cause and potential treatment approaches. In contrast, the values or preference of the patient concerned refer to any of the patient's commitments that would make, for instance, certain treatment options more preferable to the patient than others for reasons separate from their efficacy. As an example, a patient might prefer to have a slightly diminished lifespan instead of a longer one that required more regular medical interventions, or required carrying the burden of certain conditions for longer.

The paternalistic model seemed to have persisted despite the legal practice of informed consent, and was eschewed due to the persistent power asymmetry undermining patient self-determination. At the other end of the spectrum lies the informed model (also called the 'engineering' or 'informative' model), where the information flow is still one-way in that the relevant medical information is conveyed by the doctor to the patient. However, the patient then deliberates on the final clinical decision based on both the medical information conveyed as well as his own values and preferences, and then decides on the final option on his own. Under such a model, the clinician is merely the conduit for relevant medical information, and the tasks of deliberation and final decision-making rest *solely* with the patient.

Despite seemingly maximizing patient autonomy, the informed model of decision making is seen as problematic for two reasons. First, it assumes a clear distinction between facts and values in practice, and assumes that the patient's values and preferences are fixed and easily accessible to them, waiting to be straightforwardly utilized for clinical decision-making.**Error! Bookmark not defined.** However, this is too idealized a model of the clinical experience of patients – the patient's values and preferences are not always fixed and known to the patient, and have to be reconstructed (or indeed even constructed for the first time) given the clinical facts of the matter. Second, the informed model is argued to diminish the role of the clinician beyond what is ethically appropriate.**Error! Bookmark not defined.**, **Error! Bookmark not defined.** Not only does the informed model overcorrect for paternalism by allowing for complete moral abdication by the clinician for the clinical encounter, it can sometimes do so in practice when the patient requires decision-making assistance (beyond needing informational clarity) the most.

For these reasons, neither the paternalistic model nor the informed model are seen as providing an appropriate ethical model for clinical practice. What is necessary is a model that falls somewhere in between, that (1) allows for a two-way information flow from the doctor to the patient regarding all the relevant medical facts and from the patient to the doctor regarding the patient's relevant values and preferences, (2) does not impractically treat the patient's values and preferences as fixed and accessible, but as sometimes requiring reconstruction in clinical settings, and (3) allows for maximal self-determination for the patient by requiring informed, value-aligned consent while not leaving them without deliberative or decisional support in a moment of need. A shared decision making (SDM) model is thus any model which maps roughly onto criteria outlined above. While there isn't a strict definition of SDM, it can be understood as a family of models of clinical operation that resemble each other in these ways. Both the deliberative model and the interpretive model can be seen as SDM models,**Error! Bookmark not defined.** with the only difference being that while the interpretive model merely requires the clinician to assist the patient in reconstructing

their values and preferences, the deliberative model provides license for the clinician to advocate for "health-related values". For instance, the deliberative model would allow, as part of SDM, for a clinician to urge patients with high cholesterol levels who smoke to change their dietary habits and/or to quit smoking.

Of course, it is not ethically appropriate to advocate that all clinical practice should conform to any one particular type of SDM model, or even an SDM model at all. For instance, it would be difficult to ensure that unhurried and shared decision making, or even informed consent, takes place in some situations of emergency care. Similarly, while it might be argued to be ethically appropriate for the doctor to practice value-advocacy in the case of the high-cholesterol patient above, value-advocacy would be less ethically appropriate in the case of a patient choosing between treatment options that either allow a longer, less comfortable life or a shorter, more comfortable one – under such circumstances then, an SDM approach might require constraints on what values clinicians can advocate for. Thus, while SDM has been accepted as the new standard of care in clinical settings, there are still exceptions and variations. Further, an exact account of the circumstances under which distinct specific models of SDM are appropriate has still not been worked out.

Despite this, more recent analyses of the concept have characterized two distinct conceptions of SDM – a narrow conception and a broad one (Entwistle et al. 2016). Narrow conceptions focus on the decision parameters that have been discussed so far: the content and direction of information exchange (including information about medical facts as well as patient values and preferences) and the participation of the doctor and the patient in both the deliberation as well as the final decision. Further, narrow conceptions also restrict SDM to clinical decision situations that are 'preference sensitive': where the choice involves two or more options that feature trade-offs and are equally reasonable, or where evidence of effectiveness is unclear, and/or where different patient values can impact the choice significantly (Entwistle et al. 2016, NHS 2021, Whitney et al. 2004).

On the other hand, broad conceptions of SDM go beyond the above narrow requirements for clinical settings, and instead emphasize attention more to "the ethos of health care consultations, the quality of interpersonal relationships, and the subjective experiences and emotions these reflect and generate." (Entwistle et al. 2016) There is thus an emphasis on being patient-centered beyond just methodically soliciting values and preferences from the patient and providing them with information about risks and benefits of different medical options. Such conceptions re-emphasize and build on the earlier point that patient values and preferences can't simply be taken as fixed, clear, and accessible. This is especially so given that the practice of modern medicine deals in complex situations such as (1) patients having multiple health problems, (2) long-term doctor-patient relationships where medical problems and preferences are revisited, (3) additional agents in the clinical encounter including both other physicians as well as patients' family and friends, and (4) the different structures of care delivery that can affect a patient's experience within a health care system. Broader conceptions of SDM thus argue that simplified narrow approaches, which articulate the elements that need to obtain in a clinical encounter for shared decision making to obtain, are unsuited to the complexities of modern medicine. Shared decision making in more diverse forms of doctor-patient collaboration can therefore only obtain through focusing on the ethos of patient participation and empowerment by cultivating an open and respectful clinical environment.

These conceptual foundations and elements, articulated through the various models and conceptions of SDM, have been operationalized into various measures that can be used to assess the extent of SDM in any particular clinical encounter. These measures can be classified using two

approaches. The first categorizes SDM measures as pertaining to either clinical decision antecedents, the decision process itself, or the decision outcomes (Scholl et al. 2011). Decision antecedents can concern, for instance, the role played by preferences that patients have in advance of beginning the discussion regarding a particular decision, such as the extent of control they want to exercise over the medical decision making. SDM measures focusing on the decision process instead focus, among other things, on the extent to which patients are involved in deliberation and decision making, whether the risks, benefits, and uncertainties of the various options are communicated to them, and the questions asked by the clinician. Finally, measures of decision outcomes focus on the extent to which patients are satisfied or regretful about the choices made in the clinical encounter, the extent of decision conflict they felt, and so on. The second approach to classifying SDM measures categorizes them based on whether they are patient-reported measures, which are surveys filled out by patients, or observer-reported measures, where the clinical encounter is observed by an impartial observer and rated for the extent to which SDM was achieved.[27] A variety of such measures have been deployed in clinical settings over the last few decades. Though there isn't a clear consensus on which specific measure tracks SDM most effectively (in part due to the lack of consensus even at the conceptual level), the deployment and validation of SDM measures, and the insights they yield, has become an integral part of the overall landscape.

05.3 Clinical AI and Shared Decision Making

Having looked at the core elements and concepts of SDM, we can now consider to what extent the use of black-box clinical AI systems, where clinicians and patients are not able to access an explanation of the system's recommendation, undermines SDM. As mentioned, such systems have been argued to undermine SDM in two ways. First, the black-box nature of these systems has been argued to obscure critical information about how the clinical recommendation was arrived at. This information is argued to be necessary for SDM, especially for the patient to be sufficiently informed about the choice they are to make. Let us call this the 'insufficient information' problem. Second, clinical AI systems have been argued to run the risk of excluding important patient values and preferences, especially if they are black-box in nature. We can call this the 'value congruence' problem.[28] We shall consider and address each of these two critiques in turn.

05.3.1 Problem 1: Insufficient information

The basic argument from insufficient information presented in the literature runs as follows:

> **P1:** SDM requires that the patient be informed of all information relevant to the medical decision at hand
>
> **P2:** The use of black-box AI does not allow the patient access to a medical explanation of how the diagnostic or treatment recommendation was reached.
>
> **P3:** A medical explanation of how the diagnostic or treatment recommendation was reached is information relevant to the patient (in the sense of P1).
>
> **-**----
>
> **C:** The use of black-box AI does not meet the requirements of SDM

[27] There also exist clinician-reported measures of SDM, but those are used relatively infrequently.

[28] There is also a third problematization of black-box AI models made in the literature: that it is unclear if even the mere use of such models is to be disclosed to patients, and if not whether the lack of such disclosure constitutes a violation of informed consent. For reasons of space, this will not be discussed in this chapter.

Here, the term 'medical explanation' refers to the type of mechanistic, causal explanation, which invokes medical concepts, that is typically conveyed in non-AI clinical settings. A simple example can be a doctor telling a patient that the most likely cause of persistent anemia is a pre-existing gastrointestinal condition that causes ulcerative bleeding, that has probably once again become active due to the use of anti-platelets prescribed after a cardiac event. Here, there is a clear evidential chain of reasoning starting from the evidence at hand, moving to what that indicates about the particular patient's physiological state, what that physiological state might cause, and how that supports the final differential diagnosis and subsequent treatment recommendation. Machine learning systems, specifically deep learning models with many hidden layers, depart from the norms of existing medical practice precisely because they aren't able to provide such medical explanations.

Different versions of the above argument have been advanced and considered numerous times in the literature. For instance, it has been argued that "when the practitioner cannot make sense of the relevant medical information buried in the deep learning network, then neither can he present the information in a way that enables a patient to comprehend and process it rationally." (Bjerring and Busch 2020) This ability to comprehend the information and process it rationally is seen as crucial for patients to make autonomous decisions. Similarly, it has been argued that "[a]s the patient is not provided with sufficient information concerning the confidence of a given diagnosis or the rationale of a treatment prediction, she might not be well equipped to give her consent to treatment decisions." (Grote and Berens 2020) Elsewhere, commentators note that "for an informed consent process to proceed appropriately, it requires physicians to be sufficiently knowledgeable to explain to patients how an AI device works, which is rendered difficult by the black-box problem." (Schiff and Berenstein 2019) Similar versions of this argument have been considered in other commentaries as well (Schönberger 2019).

Whether such an argument succeeds in showing that the use of black-box AI models in clinical settings undermines SDM relies on various factors, which we shall now look to.

05.3.1.1 Patient's ability to meaningfully understand and utilize the medical explanation

As seen in section 2, the most crucial argument for the importance of shared decision making is that it promotes patient autonomy and self-determination. This is both by allowing patient values and preferences to feature in the clinical deliberation and decision making in the first place (in contrast to the paternalistic model) and by ensuring that patients have sufficient support in the decision scenario and are not left overburdened and alone (as compared to the informed model) (Ubel et al. 2018). Whether a medical explanation of the diagnosis or treatment recommendations, of the sort that deep learning models cannot provide, is important for the patient thus depends on whether it'll allow the patient to promote their own autonomy or to self-determine more effectively.

In most cases when a medical explanation is conveyed to them, patients might be able to understand the explanation in that they would be able to (1) register certain facts about their physiological or biological states, or more generally register medical or pharmacological knowledge, and (2) follow along the chain of inferences as their doctor presents it to them, utilizing some of these facts and arriving at others in turn. In the example used above for instance, the patient would be able to register as facts that their blood test results over a period of time indicate persistent anemia, that other test results show a resurgence of intestinal ulceration. They would further be able to understand that intestinal ulcerations can cause anemia even if blood in stools is not seen as the bleeding occurs in trace amounts, and understand as a general medical rule of

thumb that anti-platelet use can revive previously dormant gastrointestinal inflammatory conditions.

What most patients are not able to do unless they are themselves clinicians in the same sub-domain that they're seeking treatment for is to evaluate that information for its reliability, and subsequently decide whether the diagnosis or treatment recommendations arrived at are medically valid. For instance, the above patient would not be able to know whether no blood in the stools really is consistent with large intestinal ulcerative bleeding severe enough to cause persistent anemia. Put another way, most patients are not able to disagree with their doctor on whether the medical diagnosis or treatment recommendation the doctor arrived at is reasonable, and so are not able to *utilize* the medical explanation that was conveyed to them to adjust their credence in the diagnosis or recommendation. The most they might be able to do is to provide the doctor with data, such as symptoms or family history, that she might not have factored in. The medical explanation provided to them is thus causally superfluous to their decision making, and it is thus difficult to understand why it can be autonomy enhancing. It neither increases the options that they have open to them, nor does it allow them to reason better about the existing options such that they might make a choice for one over the other.

Importantly, this is not the case when patients are provided with information about treatment options or further steps in the diagnostic procedure. When patients are told that they have two or more equally reasonable treatment options with differing risks or benefits, for instance if one option has a higher chance of a longer life but one that requires persistent medical care whereas the other option would likely result in a shorter remaining lifespan but one without protracted medical care, this information is useful for their decision making. Depending on their preferences about the kind of life they'd like to have, they may choose one way or another. Such information, in contrast to medical explanations, actually enhances patient autonomy.

This is why when speaking about the level and type of information necessary for informed decision making, key elements include diagnoses, prognoses, the treatment and its alternatives, the benefits and risks of all available options, and the uncertainty surrounding any of these facts (Braddock 1999, Beauchamp and Childress 2012). What is not included as a requirement for informed decision making is an ability to understand and engage with the medical facts and reasoning used to arrive at these diagnoses, treatment recommendations, and the associated risks, benefits, and uncertainties. Further, this is the case not just for philosophical accounts of informed decision making, but also for SDM measures used in practice. For instance, the popularly used OPTION Scale as an observer-reported measure of the clinical decision process checks for clinician-patient discussion on the different options available to the patient, the pros and cons of these options, and the patient's expectations and fears (Elwyn et al. 2003). It does not, however, check for the patient's understanding of the clinician's reasoning in arriving at the diagnostic or treatment decisions. Similarly, the Decisional Conflict Scale, as a patient-reported measure of clinical experience, inquires about the patient's understanding of the different treatment options, their benefits, risks, and side effects, but not about their understanding of the medical reasoning used (Traditional Decisional Conflict Scale 2021). The focus of other popular SDM measures is similarly towards the options and the associated risks and benefits, rather than the medical reasoning used to arrive at those options that would be presented as part of a medical explanation.

All of these factors point to medical explanations – understood in the previously specified mechanistic, causal sense – as being inessential to autonomy and thus informed *and* shared decision making, and seem to vindicate clinical reliance on black-box AI models. However, as mentioned, this rests on a model of patients as passive recipients of medical expertise from their doctors. Recent characterizations of patients, on the other hand, model them as active, autonomous, and

competent inquirers in the medical setting (Kukla 2007). This has been the result of a proliferation of alternative sources of medical information that patients can access and learn from. What follows from such a model of patients is the possibility that in some clinical situations, patients may actually be able to engage with and evaluate the quality of medical expertise they're receiving. We might imagine patients coming in to clinical situations having informed themselves of possible treatment approaches through online research or speaking with other doctors. Under such conditions, medical explanations actually might have a causal impact on the care that such patients elect to receive, and so access to such explanations can enhance patient autonomy and improve shared decision making. While not all, or even the majority of, clinical decision situations might be characterized by such patient engagement, we face the epistemic problem of not knowing when such situations may arise. Switching over to reliance on black-box AI models thus risks ruling out clinical encounters where patient autonomy might have been enhanced.

The concern here is that by advocating a clinical process that cannot provide medical explanations (when clinicians and patients rely on black-box models) over a clinical process that can provide medical explanations, we leave some autonomy gains on the table. But this need not be the case. Even if black-box AI systems reduce the ability of clinicians to provide medical explanations, there are other alternative explanations that can be provided. For instance, if a clinician is relying upon an AI system that provides ranked treatment recommendations for breast cancer, the clinician can provide a statistical explanation to the patient to justify her trust in the AI model. The clinician can convey the past track-record of the system in recommending effective treatments, and what the associated health outcomes were. The patient can also be informed of the various patient population groups that contributed the data that the AI was trained on, as well as performance rates for each of these various sub-groups to show that the patient's sub-group is well-represented.

A critic might argue that such explanations are different from biological, mechanistic ones that patients can potentially understand, and so would actually frustrate the understanding needed for self-determination. However, patients aren't able to natively understand such medical explanations either – active, autonomous, and epistemically competent patients develop their ability to understand such explanations and critically engage with them because such explanations are the *lingua franca* of the standard of care. If the standard of care requires reasoning for the doctor based on statistical explanations, then such epistemically forward patients would similarly be incentivized to critically engage with these types of statistical explanations as well.

Another approach that the clinician might take is to provide a best-guess account of what the medical explanation could actually be. If the AI system's recommendation and certainty estimate is roughly in line with the clinician's independent determination of what treatment would be appropriate, the clinician could provide a medical explanation based on how she independently arrived at her treatment recommendation. Alternatively, if the black-box deep-learning model can be used with a second interpretable (non-black-box) model that provides best-guess estimations of the medical explanation for the first model's recommendation, these estimations could be conveyed by the clinician to the patient.[29] While these best-guesses of the medical explanation underpinning the black-box model's recommendation would not be guaranteed to be true, under certain circumstances they might provide good probabilistic evidence of what the medical explanation might be.

More importantly, regardless of whether the explanation is a statistical one or a best-guess medical one, it might still be possible for active, autonomous, and competent patients to evaluate it in the same way that they evaluate medical explanations. In the same way that patients can independently

[29] These approaches are used to arrive at *explainable* models, as compared to *interpretable* models which are transparent by design and have the medical explanation underpinning their recommendations be easily accessible.

learn enough to evaluate medical explanations in certain situations, they might also do so to evaluate non-medical explanations. For instance, they might independently educate themselves on model performance metrics such as sensitivity, specificity (and their derivatives – AUROC, F1, etc.) as well as the conditions under which such metrics can be seen as reliable (extensive multi-site validation, RCTs, etc.). They might then use this understanding to ask the doctor about the model's performance history, training and testing data, and check whether for their profile the model has both been adequately tested and performed satisfactorily, to ultimately develop trust in the model's recommendations. Avenues such as online resources, speaking to other doctors, and speaking to fellow patients will all remain open to learn from for patients to equip themselves for clinical encounters where their doctors rely on opaque AI models. Possibilities to enhance patient autonomy and shared decision-making would thus continue to exist if black-box models are used in clinical settings.

This argument thus rebuts the conclusion from the basic argument by doing one of three things. It denies P3 from the basic argument in the case of epistemically passive patients, by arguing that medical explanations don't count as being information relevant to the patient since they don't contribute to the patient's autonomous choice. In the case of clinical encounters where patients aren't passive but instead are active, autonomous, and competent inquirers, it argues that either P1 or P2 are false. If the explanation provided is a best-guess medical one, P2 is false because some level of reconstructed medical explanation is still available to them. If the explanation provided is a statistical one instead of a best-guess medical one, either because of reduced confidence in the doctor's or the model's ability to provide a best-guess explanation or because a statistical explanation is available anyway, P1 is shown as false. P1 states that *all* information relevant to the medical decision at hand should be shared with the patient, but this need not be true – as long as *sufficient* relevant information is provided, that would satisfy SDM. Here, either medical or statistical explanations in isolation are seen as sufficient. SDM can excuse the absence of some information that may be autonomy-enhancing (medical information) if other information is available (statistical information) that would serve to enhance patient autonomy in the same way (by allowing active, autonomous, and competently inquiring patients to evaluate their options by independent learning).

05.3.1.2 Information that doesn't have a causal impact but is material to the patient

So far, we've been operating under the assumption that information can only be relevant to the patient if that information enhances their autonomy, and that information can only enhance their autonomy if it has a causal impact on their decision. What if information is valuable to a patient even if it doesn't have a causal impact on their decision? The discussion in the previous section argued that patient autonomy can only be enhanced by information that either enlarges the set of options open to them, or allows them reason better and choose differently between the options that already exist. However, this account of autonomy might be contested, and it might be argued that information that doesn't enlarge the set of options or allow for a better choice between options can still contribute to enhancing autonomy. Alternatively, one might argue that even if such an account of autonomy is acceptable, autonomy is not the only value grounding SDM. As Faden and Beauchamp note, "information can be material for a person's deliberation even if it makes no causal difference to the outcome," where for something to be material to a person's deliberation it would be viewed by that person as being worthy of consideration regardless of whether it has a causal impact on the outcome of the deliberation (Faden and Beauchamp 1986).

We can sketch a few examples to illustrate this point. As Faden and Beauchamp note, a patient might want to know whether a scar occurs from a procedure even if they would elect for the procedure regardless. Knowledge about the possibility of a scar would thus be material to the

patient. A pregnant woman might want to know about the implications of a C-section delivery in terms of the "length and placement of the incision, the expected level of post-operative pain and discomfort, the increased risk of C-section in subsequent pregnancies" even if she would make the ultimate decision between a vaginal and C-section delivery only on the basis of implications to her baby's life.**Error! Bookmark not defined.** In such cases, patients would be glad to have the information prior to the decision, would value their understanding, and would be upset to discover afterwards if they made the decision without access to that information. It is their "viewing of them as material [that] *makes* them material."

Similarly, perhaps one might argue that medical explanations provide information that is material to the patient, information that patients would be glad to have and be upset to go without even if it doesn't causally impact their final medical choice because of their inability to critically engage with the material. Such an argument would reiterate and flesh out P1 from the basic argument by arguing that information relevant to the patient refers to information *material* to the patient about the decision at hand. While the literature hasn't sufficiently fleshed out what such a criteria for materiality might be, we might imagine at the very least (or perhaps one way of capturing materiality) that for some information to be material is for it to be *viewed* as material by the patient. This is also perhaps one of the things that separates broad conceptions of SDM from narrow conceptions of SDM. Recall that broad conceptions argue for patient-centeredness beyond straightforwardly providing information about risks on the basis of it being autonomy enhancing in complex clinical arrangements. This might be because broad conceptions service a different account of autonomy compared to narrow conceptions, where information and the subjective experiences of patients matter for their autonomy even if they don't enlarge their option-set or allow them to reason about it better. Alternatively, broad conceptions might differ from narrow conceptions in that they service values other than just autonomy, such that all material information is considered relevant. Under such an understanding, medical explanations might be relevant when we go by a broad conception of SDM even if not relevant for a narrow conception.

This is an intuitively compelling argument. Medical explanations might indeed provide the sort of information that patients might be glad to possess – knowledge about how the various physiological events happening within their body that might cause a diagnosis or treatment decision to be the right one for them. Statistical explanations might not be seen as material in quite the same way, since they're not quite as personal, or just might not be *viewed* to be material. Best-guess medical explanations might work, but perhaps the increased uncertainty associated with them compared to *bona fide* medical explanations might make them less material.

We might imagine a thought experiment where the world has so far been using only black-box processes for medical decision making (even with human doctors) – accompanied by statistical and best-guess explanations – and there suddenly is an opportunity to switch over to transparent AI models which supply *bona fide* medical explanations along with their recommendations. Assuming that (1) the performance of these new models, in terms of their accuracy, is comparable to the previous black-box approaches, that (2) clinicians are able to use these new models equally well to achieve the same level of health outcomes, and that (3) mechanisms of accountability and responsibility-attribution in health care apply equally well to the new transparent models, we can ask the following question: should we choose to switch over to the new transparent models? Having bracketed away non-SDM concerns, if we do choose to switch over, there seems to be some value in medical explanations that is not substitutable through the kinds of explanations provided for black-box models.

Ultimately, this seems to be an empirical question. If what makes information material to a patient is whether they view that information as being material, whether or not medical explanations are

material to patients and whether (and the extent to which) they can be substituted for with statistical or best-guess medical explanations is something that can only be determined by empirically surveying patients. However, in advance of such empirical work being carried out, we might note two conceptual points.

First, the examples provided above, noting the materiality of information for patients even if there is no causal impact to their decision-making, concerns information about the nature of the intervention and its possible risks or side-effects. This is information that is available when relying upon black-box models even if no medical explanation is available. A clinician relying upon a black-box AI model to decide on surgical interventions will independently be able to tell the patient what the probability of a resulting scar might be, or whether a C-section would contribute to greater post-operative pain and an increased risk of C-sections in future pregnancies. All a medical explanation does is provide information on why certain test results, symptoms, or medical history indicate what intervention is most likely to be effective. Thus, intuitions and empirical work concerning the nature of the intervention and the associated risks might not in fact be reflective of the materiality of medical explanations for patients.

Second, we might imagine that there are limits to the information required for shared decision making, even if such information is viewed as being material by the patient. For instance, our patient trying to understand why he has persistent anemia might reasonably inquire about how exactly iron absorption works in the body and how certain pharmacological interventions might enhance that process. However, he might also inquire at a very granular level about the exact chemical processes in play and ask for an education on the chemical fundamentals of how gastrointestinal ulcers are formed, or ask about other information which is medically not relevant but yet believed to be so by the patient. While it is the patient's right to not pursue further diagnosis or treatment with a doctor who is not able to answer these latter questions, we would not think that shared decision making had been compromised due to the doctor's inability or unwillingness to answer such questions. There is presumably a limit to the information that can be reasonably understood as required by SDM. It is unclear whether the provision of medical explanations would fall within this limit.

05.3.1.3 New power and informational asymmetries against the patient

Problematizations of the paternalistic model of clinical decision-making have noted the informational and power asymmetries between the doctor and the patient.**Error! Bookmark not defined.** These asymmetries diminished the ability of patients to be autonomous agents as far as their own medical care was concerned, and SDM as a model of clinical decision-making was encouraged as a way of reducing the asymmetries. For the most part, it seems that the clinical use of black-box AI models will not worsen such asymmetries between the clinician and the patient, as neither the clinician nor the patient have access to medical explanations and any statistical information justifying AI-reliance that the clinician has access to can be passed on to the patient. However, there is a possibility that new asymmetries might emerge.

Specifically, it might be argued that a new informational and power asymmetry might exist between the model or model developers on the one hand and patients on the other. Although the medical explanation for an AI's clinical recommendation is not accessible to the clinician or the patient (or even the model developers), it still exists embedded within the various hidden layers of the model. What prevents it from being understood by humans is the mismatch between (1) the mathematical optimization in high dimensionality that is characteristic of deep learning models and (2) the demands of human-scale reasoning and styles of semantic interpretation which cannot capture this high dimensional optimization (Burrell 2016). Similarly, it might be argued that although model

developers can't access such medical explanations either, they understand a lot more about how such models function, how they are structured, and how they are likely to perform for instance in circumstances different from those they were trained in. Put another way, model developers would have a much better statistical understanding of the model, the same understanding that can be used to provide statistical explanations for the model's recommendations.

Would such informational asymmetries translate to power asymmetries in a way that causes SDM in the clinical encounter to regress? When it comes to the information hidden in the layers of a deep learning model, this seems unlikely. There are other aspects of the traditional care-pathway that have relied on epistemically inaccessible methods of medical knowledge creation or medical care delivery, to no detriment to SDM. Early stages of modern drug discovery (or rational drug discovery), for instance, rely on the screening of active agents against biological targets without necessarily an understanding of what the agent-target interaction consists of, and why the therapeutic effects obtain (Sausville 2012). Such combinatorial, brute-force approaches, don't necessarily reveal the explanatory logic of why certain lead compounds are successful when others are not, even to the researchers working on the process, but this is not seen as compromising the autonomy of patients who ultimately benefit from the drugs produced. If these existing epistemically/explanatorily-opaque approaches are not seen as introducing problematic informational asymmetries for SDM, it's unclear why deep learning models which are black-box in nature should be argued to do so.

However, when we consider informational asymmetries between model developers and patients, the situation may be different. Considering that we've argued in the previous section that epistemically active and competent patients can use statistical information to enhance their option set or reason better about the options that already exist within the set, it is hard to deny that statistical explanations can matter for patients. Asymmetries in access to statistical information, or more generally non-medical model explanations (including model training and validation details, for instance) are problematic for SDM. Such asymmetries in access to statistical information may still remain unproblematic when it comes to epistemically passive patients (who might limit their epistemic desires to information about outcomes, risks, etc.), but lack of access to relevant statistical information or understanding of model structure (which is possessed by model developers) may prevent some epistemically active patients from ascertaining if the AI's recommendations are truly in their best interests, and might thus diminish their autonomy. One way to understand this better is by revising the basic argument to refer to statistical explanations, or explanations regarding the structure of the AI model, instead of (or in addition to) medical explanation. Such an argument would be far more successful than the basic argument laid out in section *5.3.1.1.*

There are two responses that can be made here. The first is to claim that if epistemically active and competent patients wish to understand better how deep learning models are structured, how they can be evaluated, how statistical information about model performance should be evaluated and what the common metrics of evaluation are – in other words, if they wish to gain some level of access to information that model developers might possess – then there are a variety of ways in which they can educate themselves independently. There are enough resources that can be perused online, and these patients are certainly free to seek out other users of such AI models to verify their understanding of model behaviour. However, such a response fails on two counts. First, it lays the burden of navigating the new approach to medical decision making (statistical and other forms of evaluations of deep learning models) entirely on the patient. If improving access to learning about how medical decision making enhances the autonomy of some patients, then good SDM requires working towards providing such access. More importantly, the response also fails because even if epistemically competent patients find resources to learn more about how deep

learning models function and how their statistical performance is to be evaluated, this might not allow them access to the relevant details for the model that they encounter in question. Such information about that model might not be accessible for a variety of reasons – it is intellectual property that is protected by the model developers, or perhaps it just hasn't been reported widely even if it isn't.

This leads to the second, and more promising, response. If paternalism of doctors to patients was overcome by looking for SDM approaches that would allow patients to receive more information and participate more in clinical settings, perhaps informational and power asymmetries between model developers and patients can be addressed in a similar way. What epistemically active and competent patients require is the opportunity to engage with experts regarding the reasoning through which medical decisions are made. Prior to the use of black-box AI, these experts were doctors who were able to provide comprehensive medical explanations that patients could engage with. SDM for the new informational asymmetry would require a way to allow patients to engage with the new experts (or at least more competent reasoners) on statistical and structural aspects of the model. This might be done in a few ways. If there is a significant difference in competence between doctors and model developers as far as a statistical and structural understanding of the model is concerned, then doctors might be required to learn a lot more on this topic before they rely on such models in clinical settings. While this has been something that has been encouraged so far to produce better epistemic decision making by clinicians when relying on black-box AI models, it will also promote SDM as patients will be able to ask doctors directly about the non-medical explanations that are relevant to them. Alternatively, we might cut out the middleman and allow patients to engage with model developers more directly. This might happen through patient consultations and engagement for each model developed, and a subsequent publication of the findings of the consultation (and conversations between epistemically competent patients and developers) so that patients not involved in the consultation can still access and use it to make decisions about their medical care. Alternatively, new models of on-going patient consultant and engagement can be pursued, such as pursuing it in a publicly accessible way online.

One additional concern that critics might raise is that such approaches may not fully eliminate the informational asymmetries, or sufficiently eliminate them, as there are still critical insights, and areas of understanding, that model developers might have that are not imparted to the patient (either through the doctor or directly). This might still perpetuate some power asymmetries and compromise SDM. However, this is setting the bar for SDM too high. There is plenty that doctors themselves have access to that can't be imparted onto the patient, regardless of how shared the decision-making is: instincts, intuitions, judgments from years of learning and experience, and a whole host of epistemic apparatus outside of and supporting the relatively neat package of biological/pharmacological concepts and causality that constitute a medical explanation. We don't include the ability to communicate that and, even more stringently, to inform the patient of it sufficiently for it to feature in the patient's decision-making/awareness as a necessary condition of SDM. We expect medical explanations to a reasonable level. Similarly, we can expect statistical explanations, or model-specific explanations, from model developers to a reasonable level. SDM isn't about eliminating the informational and power asymmetries - that would require the patient to become the practitioner. It's about reducing it enough that patient values and decision-making have a path in.

All in all, as we've seen, informational asymmetries between model developers and epistemically active and competent patients can have a foreseeable impact on patient autonomy, and new empirical approaches to SDM are necessary to prevent a regress. The discussion so far has just presented the conceptual foundations for such a regress, and briefly sketched solutions – more work needs to be done.

05.3.2 Problem 2: Value congruence

The second way in which the clinical use of AI is argued to compromise SDM is by reducing the influence of the patient's values and preferences on the clinical decision. This is argued primarily for treatment-focused decision making, but does also apply in some diagnostic situations. There are two main claims made:

A. AI systems that recommend treatments for patients don't necessarily factor in the patient's values and preferences when making the recommendation, as they're built to function for a large number of patients. This is especially so if the treatment choice is 'preference sensitive'.

B. The problem in claim A is further exacerbated if the AI system is opaque: opaque systems do not allow scrutiny of the decision-making process driving the final recommendation, and in doing so might prevent doctors and patients from seeing the AI's recommendation as being value-laden at all.

One example given in the literature is that of Watson for Oncology, an AI decision system developed to make recommendations for cancer treatment (McDougall 2019). This model ranks treatment options based on the value of maximizing lifespan – this is not something that can be adjusted on behalf of a particular patient and is the sole value driving treatment recommendations for all patients. In fact, it is not even explicitly mentioned as the value based on which the treatment rankings are generated. Given that not all patients make treatment decisions based solely on longevity (claim A), and that indeed it is in fact unclear that treatment determination is being driven by the value of longevity (claim B), this is argued to compromise SDM. As noted, "these types of AI systems currently do not encourage doctors and patients to recognize treatment decision making as value-laden at all" (McDougall 2019).**Error! Bookmark not defined.**

The same point has been made by multiple commentators. For instance, one commentator notes the key question of clinical AI use as "if it is possible for the AI systems to take these interests [patient's preferences and interests] adequately into account in their treatment planning", and if not, whether this can ground a right to refuse diagnostics and treatment planning by AI (Ploug and Holm 2020). Elsewhere, the argument is presented instead in the language of decision bias, where it is noted about claims A and B: "[t]he first relates to outcomes for patients; the second to the potential for systematic but undetected bias", where the bias in question is in favour of those patients whose preferences and values happen to line up with the AI's pre-set ones.

One of the proposed solutions for the problem of value-based AI decision making is to have AI systems be *value-flexible* by design (McDougall 2019).**Error! Bookmark not defined.** As McDougall notes, when AI systems are value-flexible by design, they "allow for diversity among the values of individual users and can incorporate different values into decision making based on the specific user." This can be considered a technical design problem, in contrast to the normative problem of which values should be encoded within such AI systems. A system that is not value-flexible in such a way risks, it is argued, that clinical encounters "return to more paternalistic medical care" (McDougall 2019).**Error! Bookmark not defined.**

Let us consider what it would take for this problem to indeed be a serious one, and whether the solution of 'value-flexible' design is an effective one. First, it is important to note that even if we accept the value-flexible design solution as an effective one against AI systems which have the value congruence problem, it is still only applicable for some AI models. It applies to those models

which provide treatment recommendations for preference sensitive decisions, where the impact of different treatment options can be substantial depending on the patient's values and preferences, and when there isn't a clearly superior choice in terms of health outcomes (so equipoise exists between them). AI models which provide treatment recommendations in other settings would not need to be value-flexible. This is an important point to note, because there can be a temptation to say that all clinical AI models should be value-flexible, which doesn't follow from the above critique. For instance, a model that provides recommendations from a class of treatments, where there are minimal side-effects or risks from any of the treatments and the choice is primarily based on details around the patient's clinical presentation, is not concerned with preference-sensitive decisions – such a model would not need to be value-flexible.

Second, it is not clear that a solution to the value congruence problem requires the *AI system itself* to be value-flexible, rather than requiring value-flexibility in the overall clinical decision-making system. Here, by the overall clinical decision-making system, we include the AI model, the clinician relying on the model, the patient conversing with the clinician, as well as any other agents and processes involved in care delivery. The goal of SDM, at least for narrow accounts of SDM, is to ensure that the clinical decisions are made in line with the patient's values and preferences. Given that the final clinical decision is made by the overall clinical decision-making system (as opposed to just the model), ensuring that every sub-component of that system, such as the AI model, is value-flexible by design is only one way in which this can be achieved. It is not the only way – we might allow AI systems to not be value-flexible, but for their ultimate ranking of treatment options to be corrected for by the clinician once they have solicited and factored in the patient's values and preferences. Taking the example of the cancer treatment AI model, even if the model's recommendations are based on maximizing longevity, the clinician knowing this can check with the patient if that is the value he'd like to optimize for. If the patient would rather have the treatment recommendations be optimized for an alternate value, such as minimal morbidity, he can raise this with his clinician and the clinician can reorder the treatment rankings based on her own medical expertise.

McDougall does consider such an approach, but ultimately argues that "such an approach diminishes the patient's role and represents a backwards step in respecting patient autonomy" (McDougall 2019).**Error! Bookmark not defined.** However, it is unclear why this is so. Once again considering the narrow conception of SDM, imagine two scenarios. In the first scenario, the clinician uses a value-flexible AI that takes in the patient's preference for minimizing morbidity rather than maximizing longevity, and outputs a ranked list of treatment recommendations. In the second scenario, the clinician uses an AI system that isn't value-flexible, and receives a ranked list of treatments based on longevity. The clinician then reorders the list based on her experience to yield a resultant list optimized for minimizing morbidity. If the final list arrived at in both scenarios is the same, it is unclear why the first approach is still preferable from the point of view of SDM.

One response to this might be that there are certain scenarios where the clinician might not be able to compensate for the model's value-ladenness. When decisions need to be made quickly, such as in critical and emergency care situations, there might not be enough time for the clinician to provide new recommendations if they're relying on a value-inflexible AI system. However, in such an emergency and time-sensitive situation, SDM is usually suspended anyway since there may not be enough time for patients to provide a preference. Of course, there might be certain situations where there is enough time for patients to convey their values and preferences but not enough time for clinicians to factor that in to the model's list of recommendations to reorder them. For clinical decision-making in such a 'goldilocks' zone, value-flexible AI would indeed be useful, though it is hard to think of such scenarios.

It might also be hard for clinicians to reorder AI recommendations if the AI system is being used for computationally complex tasks that exceed human capabilities, and if it isn't possible to decouple the ethical decision-making from the computational decision-making for that task. For instance, we might consider resource allocation AI models, like risk prediction to prioritize patients for high-risk care (Obermeyer 2019), or kidney exchange algorithms that match prospective recipients for kidney transplants who have willing but incompatible donors with other such pairs (Roth et al. 2004, Freedman et al. 2020).[9] In such situations, a subsequent modification of the recommendation by the clinician to factor in patient values would not be possible without reworking the entire complex computation, which clinicians would not be able to do without a drop in performance – the entire point of the model is that it can handle such complex computation better than humans can. Under these conditions, value-flexible AI would be necessary to ensure that patient values and preferences drive treatment decision making and the requirements of SDM are met. However, it is important to note that both examples of this mentioned in the literature are public health tasks rather than medical decisions made for a patient. Public health tasks cannot be deliberated on based on the values of particular patients, and so SDM doesn't apply as a normative ideal there.

That being said, this by no means precludes a scenario where models performing complex computations in a value-laden way could be developed and used for clinical decision making that is personalized to a particular patient. We might imagine this happening for instance when deep learning models significantly outperform clinicians for treatment recommendations in preference-sensitive choice situations. Currently, we might feel comfortable saying that even if AI models aren't value-flexible, that is acceptable given that clinicians still outperform them and so can adjust the model's recommendations accordingly. However, once deep learning models significantly outperform humans, they are essentially producing solutions to computational tasks that are too complex for humans to replicate. When that happens, value-flexibility would be a very important desideratum for the model in question.

However, value-flexibility does *not* require that AI models be interpretable and not black-box. We might imagine developing a deep learning model for preference-sensitive situations in such a way that each piece of training data (such as a diagnosis or set of medical evidence) receives as many labels as there are operationalized values that could vary amongst patients. For instance, if a cancer treatment recommendation model is being trained on patient cases, each patient case it is being trained on could be labeled twice: once to indicate the right treatment if the patient wants to maximize their longevity, the other to indicate the right treatment if the patient want to minimize their morbidity. Opaque models can thus still be value-flexible. Further, if black-box models are designed to be value-flexible, this attenuates the danger presented in claim B as well – if clinicians receive black-box AI models that are designed to be value-flexible and are aware of this, they will if anything be more aware that the decision-making in that instance can be value-laden.

05.4 Conclusion

In this chapter, we have considered two main challenges that the 'black-box' nature of deep learning AI models might pose to the achievement of shared decision-making in the clinical context. First, their inability to provide doctors and patients with medical explanations can be seen to keep information from patients that is relevant to their participation and decision-making. Second, their opacity risks worsening their exclusion of the values and preferences of individual patients that they make recommendations for. We have seen that these two broad arguments are very commonly made in the literature, their potency has been overstated.

I've argued that black-box models *won't* undermine SDM for the following cases: (1) for epistemically passive patients, since black-box models don't occlude what these patients are concerned with – the outcomes, risks, benefits, and uncertainties around clinical decisions; (2) for epistemically active patients, since patients can still access best-guess explanations from secondary explanatory mechanisms (the doctor's best-guess, or a secondary system explaining the first) as well as statistical explanations, which they can similarly learn to critically engage with and evaluate; and (3) cases where the model is value-inflexible, but the overall clinical decision-making system can compensate for that by clinician's subsequently adjusting the model's outputs for the patient's particular values and preferences.

I've also argued that black-box models *may* undermine SDM if medical explanations (in contrast to statistical explanations) are seen as *material* by the patients to the decision at hand, even if such explanations wouldn't otherwise change the ultimate decision that the patient would make. However, this ultimately bottoms out in empirical considerations, around what sorts of information patients do ultimately consider 'material'. Further, I've argued that it is unlikely that in this case, what materiality would consist in are the facts that constitute the causal reasoning that led to a certain decision. It is more likely that materiality consists in the outcomes of that reasoning process – the risks, the benefits, the uncertainties.

Finally, I have argued that black-box models *will* undermine SDM in a very specific case. This case obtains if the following conditions are met: (1) the model is making a clinical recommendation for specific patients (as opposed to for populations), (2) the recommendations concern decisions that are preference-sensitive, (3) the model's performance is significantly superior to the clinician's such that the clinician can't simply adjust the model's recommendation post-hoc in light of the patient's values, and most importantly (4) the model's value-inflexibility is *necessarily* tied to it's black-box nature, such that the only way to achieve flexibility would be for the model to be interpretable. If such conditions obtain, then SDM will be undermined.

This chapter has thus put forward a strong argument for why the opacity of deep-learning models doesn't undermine SDM. If this argument is successful, it becomes an additional illustration of explanations not being necessary for the ethical use of such models in medicine – alongside the illustration of their non-necessity for epistemically appropriate use in the previous chapter. The next chapter (Chapter 6) will now pull both these strings together to make a third argument for why explanations need not be necessary for the *accountable* use of such models as well.

06. Medical AI and the Many Faces of Accountability

06.1 Introduction

One of the key concerns that have been raised with the increasing use of black-box AI models – specifically of the deep learning variety – is that their 'black-box' nature disturbs or entirely prevents accountability in their use. This concern has been raised across all domains of AI use, from medicine, to financial decision-making, to lethal autonomous weapon systems. However, although this observation has been made repeatedly by commentators coming from many disciplines (including ethics, computer science, law, and sociology to name a few), the exact nature of what is meant by 'accountability' is usually left unspecified. Further, arguments outlining the exact way in which accountability – whatever it may mean – is undermined are usually not reconstructed. This has led to accountability taking on many 'faces', and very little structured discussion of what their accountable use requires and how the 'black-box' nature of AI systems impedes that.

The goal of this chapter then is two-fold. First, it is to understand the different ways in which the concept of accountability is used, and whether there are any key elements that recur throughout these different uses. Second, it is to (re)construct and assess the argument that the black-box nature of deep learning models disturb or prevent their accountable use, so that we may critically engage with it. Ultimately, this chapter will argue that there are three key senses in which accountability is used – as a means of adjudicating responsibility, as a way of promoting ethical or desirable behaviour, and as a means of providing redress for injury resulting from the use of AI systems – and that arguments for the opacity of deep learning systems undermining any of these senses of accountability mostly fails.

To do so, this chapter will proceed as follows. In section 6.2, I will survey the different ways in which accountability is used, and distill three main 'faces' of accountability. In section 6.3, I will use the problematization of the black-box nature of deep-learning models to reconstruct the argument from opacity undermining accountability. I will argue that this argument fails, because although current standards of care sometimes (though not always) may be rendered inapplicable through the use of opaque systems, which thus disturb accountability, there are good reasons to think that updated, alternative standards of care can co-exist with the use of such systems. In section 6.4, I will consider the additional problem of responsibility gaps as discussed in the literature, and briefly consider whether such gaps may be worsened by updating standards of care for the use of opaque systems – I will propose a way of structuring existing discussions of responsibility gaps to accommodate this shift in the standard of care and the problems it might pose, while finally arguing that a potential widening of responsibility gaps need not jeopardise certain forms of accountable use of AI systems, such as the ability to provide redress to injured patients. I will then end with some parting comments.

Consistent with the previous two chapters, this chapter aims to provide additional reason to doubt the necessity of explanations for the ethical and epistemic use of deep-learning models in medicine (alongside the arguments put forward in the last two chapters). As chapter 4 defused the argument from inappropriate epistemic reliance, and chapter 5 defused the argument from compromised shared decision-making, this chapter aims to defuse the argument from unaccountable use of AI models. In doing so, it addresses the third of the key objections to the use of deep-learning models in clinical practice, with the goal of showing the possibility of appropriate epistemic and ethical use of black-box models.

06.2 The Many Faces of Accountability

Across discussions of accountability, both for the domain of AI-use and outside of it, the term is used to refer to a multitude of different phenomena. Accountability can (and has been used to) refer both to a virtue and to a mechanism (Bovens 2010), to varying accounts of responsibility, liability and other means of redress, responsiveness, or ethically or professionally desirable conduct, and to the general notion of transparency. Within technological ethics, algorithmic accountability (Wieringa 2020) refers to any and all manner of approaches to ensuring that algorithms and AI models are ethically robust in their design and use.

For our particular discussion, we care about accountability because we want to know the conditions for the 'accountable' development and use of medical AI, so that we may judge the extent to which the 'black-box' nature of modern deep learning AI systems compromises their accountable use. There seems to be something ethically desirable about the development and use of such models, like that of any other system, being accountable. For this, we need to know what the term 'accountability' actually picks out, or the number of things it might pick out. Using accountability as a conceptual catch-all, as seen above, impedes our ability to apply it fruitfully to topics such as medical AI. As noted, "if accountability is everything, it may be nothing."**Error! Bookmark not defined.** Such a 'broad' sense of accountability is sometimes contrasted with a narrower conception – "the obligation to explain and justify conduct" (Bovens 2007). In this section, then, we will survey, delineate, and elaborate on the key concepts that are invoked when 'accountability' is used in the literature. As we shall see, accountability and related concepts have been scrutinized and interrogated across disciplinary boundaries – in fields ranging from moral philosophy, bioethics (and other professional ethics), political science, and law. Mapping the many faces of this concept will be useful as a first step, prior to assessing whether the use of non-interpretable and potentially opaque AI systems in medicine truly undermines accountability.

In general, we can take conceptual use of the term 'accountability' to fall into three primary categories. First, accountability is used to refer to mechanisms by which we might adjudicate responsibility for an adverse outcome. Second, it is used in general reference to ethically (or otherwise) desirable behaviour from all relevant actors. These first two uses roughly map onto what Bovens calls accountability as a mechanism, and accountability as a virtue.**Error! Bookmark not defined.** Third, accountability is seen a way of providing redress to parties that suffer the adverse outcomes. As we shall see in how accountability has been invoked in the literature on AI ethics, one or more of these different uses are invoked at different points by different authors – indeed, all three are interlinked, and draw on similar theoretical apparatus.

06.2.1 Accountability as Responsibility

The first use of accountability in discussions of AI is tightly tied to the idea of responsibility. In the philosophical literature, such a conception of accountability is similar to what is known as *negative responsibility* (or alternatively, a backward-looking account of responsibility, contrasted with positive or prospective responsibility later on) (Gotterbarn 2001, Ladd 1989).[7] Negative responsibility, or the malpractice model of responsibility, concerns how we might assess the conditions under which blame or punishment might be appropriately directed if a harm has occurred. It concerns itself mainly with two questions which seem to tug at the essence of the concept of responsibility: 'Who is responsible? How can we hold them responsible?'

We can see such an understanding of accountability in play for the AI use-case, for instance, when Raji et al. state

> We use accountability to mean the state of being responsible or answerable for a system, its behavior and its potential impacts (Raji et al. 2020).

Similarly, when Edwards and Veale state,

> At its most general, accountability is about individuals who are responsible for a set of activities and for explaining or answering for their actions. Accountability therefore entails procedures and processes by which one party provides a justification and is held responsible for its actions by another party that has an interest in the actions (Edwards and Veale 2017).

Finally, as Kaur et al. note,

> Algorithmic accountability includes assessing the algorithms based on various parameters and assigning responsibilities of harm to different stakeholders involved in developing the algorithms (Kaur et al. 2022)).

This concept of negative responsibility has received the bulk of the philosophical treatment of the concept of moral responsibility in general. More contemporary discussions have approached this topic from two angles.

First, commentators have focused on picking out the key elements of the concept of 'responsibility', when we speak of someone being responsible or being held to be morally responsible. Shoemaker and others, for instance have argued that there are multiple distinct 'faces' to the concept of responsibility – attributability, answerability, and accountability (the last of which, to avoid confusion, we shall refer to as R-accountability) (Shoemaker 2011, Shoemaker 2015).[8] When an agent satisfies the condition of attributability, their actions can be said to be reflective of, and to originate from, their 'real' or 'deep' self (Wolf 1990). They reflect the agent's true values and volition. When an agent satisfies the condition of answerability, they owe a justification of their actions based on their own evaluative judgments and reasons for actions (Shoemaker 2011).**Error! Bookmark not defined.** When an agent satisfies the condition of R-accountability, it is permissible for them to be held accountable for their actions by the use of blame and punishment based, for instance, on their lack of regard for others (Watson 1996, Shoemaker 2011).[9] **Error! Bookmark not defined.**

While these different conceptions of responsibility were primarily developed to explain the different ways in which the concept of moral responsibility is used (thus, to capture the *descriptive* sense of moral responsibility), we can also fruitfully use them to show the three distinct stages of assessing whether an agent is responsible for a particular adverse outcome (the *normative* sense of responsibility). First, as per attributability, we determine whether the actions that led to the outcome are indeed attributable to, and reflective of, the agent in question. This step is looked over in the above two conceptions of accountability (as a mechanism and as a virtue) – usually

for accountability mechanisms in healthcare (or more generally for other applied fields), the depths of the core valuational systems of agents are not plumbed. Second, as per answerability, the agent is asked to provide justification (or excuse) for their actions – why certain actions were chosen over other alternatives, what the relevant beliefs in play were, etc. Third, as per r-accountability, the agent's actions and justification are compared to an independent normative standard of how the agent ought to have acted – if the agents actions diverge from those recommended by the standard and the justification is not exculpatory, the agent can be blamed, warned, and/or punished for those actions (depending on the specific accountability mechanism in question).

Importantly, this way of picking out the key *descriptive* concepts of responsibility don't *directly* provide the conditions for judging whether an agent is morally responsible (in the negative sense) for something – they are merely providing a description of our use of the concept of responsibility. This is where the second angle of consideration comes in – going beyond a cataloguing of our concept of responsibility to lay out relevant conditions for someone to be held morally responsible for some act or attitude, such that they may be justifiably praised or blamed. Here, we encounter proposed conditions for the allocation of moral responsibility such as (among others) the *control* condition (whether the agent possesses the relevant control or freedom in performing the action – tied closely to the principle of alternative possibilities) (Frankfurt 1969), the *epistemic* condition (if an agent acts wrongly out of ignorance about some fact, when does that ignorance exculpate) (Wieland 2017), and the *moral competence* condition (when does an agent's moral incompetence, e.g. as a result of their upbringing, discount the responsibility they bear for something). These conditions can be seen as domain-agnostic philosophical conditions that hold for someone to be held responsible for some act or wrong-doing – they would apply regardless of if we're speaking about medical AI development and use, or political decision-making, or choices in everyday life.

However, responsibility cannot be judged based merely on domain-agnostic conditions such as the ones mentioned above. A key feature of accountability as negative responsibility is the necessity of an independent, domain-specific, normative standard for the evaluation of the conduct of the agent under assessment, given their alleged contribution to the harm that has occurred. For the use of AI in medical care, multiple standards exist (or will exist) for multiple domains, for multiple loci (or types of actors) (Emanuel and Emanuel 1996). For instance, standards relevant to a harmful clinical outcome could include professional standards such as the standard of care for a clinician when they're not relying on the AI model, or the standard of care on how a clinician ought to use a model's recommendations as evidence. They might also include ethical standards, such as the standard of care for patient-centered medicine. There might also be relevant standards for other actors such as model developers, such as how they might design AI systems to minimize the real-world error rate, or to best achieve a non-discriminatory or fair model, or to ensure that the appropriate values guide the reasoning of the model. Importantly, these standards exist whether or not they're actually enforced through professional hearings or legal rulings. Whether or not the standards are enforced merely influences what the actual punishment might be, not whether the standards exist in the first place.

This understanding of accountability as negative responsibility carries well outside of discussions of moral philosophy, in tracking how the concept of accountability is used for instance in bioethics, as per Emanuel & Emanuel,

> At its most general, accountability is about individuals who are responsible for a set of activities and for explaining or answering for their actions. Accountability therefore entails procedures and processes by which one party provides a justification and is held

> responsible for its actions by another party that has an interest in the actions (Emanuel and Emanuel 1996).**Error! Bookmark not defined.**

It also mirrors a widely accepted conception of accountability provided by Bovens, originally for political science and governance,

> Accountability is a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgment, and the actor may face consequences (Bovens 2010).**Error! Bookmark not defined.**

In general, we might understand 'accountability as responsibility' as featuring a few key components – an *act* (with an associated effect) carried out by an *actor*, which may or may not be in breach of either certain *domain-agnostic conditions* or *domain-specific conditions* (or both), such that the actor may be held responsible for the act and the outcome, and there may be *appropriate consequences* levied against the actor. To illustrate, a clinician (actor) might fail to defer to an AI model's diagnostic recommendation (act) despite guidance that such deference is the right strategy given the model in question (breach of domain-specific conditions) which leads to a costly misdiagnosis (outcome), and so the clinician is held to be clinically negligent by expert witnesses in court with damages to be paid (appropriate consequences). This indicates a legal means of adjudicating responsibility, but there might similarly be non-legal mechanisms through which responsibility is adjudicated and consequences imposed.

Ensuring that AI systems are accountable, in the 'accountability as responsibility' sense, then requires that a process for adjudicating responsibility be present, and that such a process not be encumbered in any way.

06.2.2 Accountability as Ethical / Desirable Conduct

Apart from being a stand-in for negative responsibility, another key way in which accountability is used in the case of AI systems in health is by being a stand-in for any ethically or otherwise desirable feature of the AI system or the workflow which it is assisting. It is therefore, to echo the language in the previous section, a reference to the standard against which the design and use of these systems is measured. It serves the second function of accountability, as a way of pre-emptively promoting desirable behaviour from all relevant actors. This is what Bovens refers to as accountability as a virtue.**Error! Bookmark not defined.** It is quoted in the AI ethics literature, for instance, as:

> As algorithms are increasingly applied within a rapidly expanding variety of fields and institutions affecting our society in crucial ways, new ways to discern and track bias, presuppositions, and prejudices built into, or resulting from algorithms are crucial. The assessment of algorithms in this matter has come to be known as 'algorithmic accountability'.**Error! Bookmark not defined.**

Also, in the case of the design of complex algorithms,

> … in order for a computer system to function in an accountable way – either while operating an important civic process or merely engaging in routine commerce – accountability must be part of the system's design from the start (Kroll et al. 2017).

Finally,

> The question of how to hold AI systems accountable is important and subtle: poor choices may result in regulation that not only fails to truly improve accountability, but

> also stifles the many beneficial applications of AI systems (Doshi-Velez and Kortz 2017).

One way to understand this conception of accountability is to see it as more connected to the philosophical concept of *positive* responsibility, in contrast to negative responsibility. Positive (or prospective) responsibility concerns "the virtue of having or being obliged to have regard for the consequences that his or her actions have on others", so that one strives to minimize adverse outcomes resulting from one's actions in a forward-looking manner.**Error! Bookmark not defined.** It thus focuses on "what ought to be done rather than on blaming or punishing others for irresponsible behaviour" (Noorman 2020).

Importantly, this requires the independent, domain-specific normative standard for evaluation that we also required for accountability as negative responsibility. It is this same domain-specific normative standard that is necessary both to adjudicate responsibility for harms, and to provide guidance for actors to satisfy forward-looking positive responsibility – it is necessary thus to 'be held responsible' and also to 'act responsibly'. The normative standard that would guide, say, a clinician using an AI model or a developer designing and deploying such a model to act in a responsible way, would be the same standard that would then allow us to judge whether these actors were morally responsible for any adverse outcomes that resulted from their actions. In the case of the clinician, for instance, this would just be the relevant standard of care that they would have to act in accordance with.

Occasionally, there is an overlap between domain-specific requirements across multiple domains, with 'transparency' for instance being a very common requirement (with others including responsiveness, controllability, liability) (Koppell 2005, Bovens 2010).**Error! Bookmark not defined.** These requirements are then sometimes understood as just being constitutive of accountability as a virtue, or as ethically desirable conduct.**Error! Bookmark not defined.** Transparency in particular is sometimes seen as exhausting the concept of accountability in AI ethics. As a requirement, transparency's importance is then extended across the board to model design (interpretable design), model use (clear guidelines), model development (transparent auditing) and more. We shall discuss this more in section 3, but it's important to distinguish when this is an intrinsically desirable requirement (transparency as an end in itself), and when it is instrumentally desirable (transparency as a means to achieving some other desirable value).

06.2.3 Accountability as (or in service of) Redress/Liability

Finally, the third sense in which accountability is invoked is as a means of allocating duty/liability to provide redress to the victims of the adverse outcome. This allows it to serve its third function as mentioned above. For instance, accountability for the use of AI in healthcare is seen as the obligation to provide an account for the purposes of "redress including compensation and allocation of blame", which is dealt largely through matters of liability (Schönberger 2019). Further, the link between accountability and redress/liability is stated more explicitly:

> Over the last few decades, physician accountability to individual patients has become more prominent. This responsibility is mediated by the standards of the profession. The prototypical but not exclusive mechanisms by which patients express their dissatisfaction with their physician or hospital is malpractice litigation… (Bovens 2007)**Error! Bookmark not defined.**

At first glance, this seems to follow straightforwardly from accountability as negative responsibility – after all, matters of tort liability, for instance, are merely the adjudication of negative responsibility that is pursued and enforced through legal rulings. However, while many

approaches to redress for the injured party are mediated through first adjudicating negative responsibility, there are some approaches to redress that do not require it.

Fault-based liabilities, such as medical negligence or certain aspects of product liability, do require the prior judgment of whether the actor is responsible (in the negative sense) for the adverse outcome. In such cases, they are directly continuous with the judgment of negative responsibility. For instance, a successful case of medical negligence in the UK usually requires the injured party show that a clinician has breached their duty of care, which first requires the establishment of the standard of care – here, the standard of care would be the independent normative standard required for assessing negative responsibility, established based on the current state of knowledge and the respected body of opinion (Herring 2016). Similarly, product liability in the case of medical device suppliers requires that the design of the product, and any warnings associated with the product that are communicated to users, be in accordance with the appropriate standard of care.

However, where accountability in service of redress is no longer continuous with negative responsibility is when the action falls under the purview of strict liabilities, which are not fault-based. Here, unlike in the case of fault-based liabilities (and negative responsibility more generally), comporting yourself as a *reasonable* person or professional does not defeat the blame or punishment (Coleman et al. 2022). Strict liability approaches to redress thus operate under an entirely different approach to redress and compensation, and accountability in such a case would require not the adjudication of either negative or positive responsibility and the associated independent standards of conduct, but merely the presence of harm or injury. Such no-fault liability approaches are being favoured for certain cases of adverse outcomes due to AI and emerging technologies, such as for the presence of 'defect' in products (European Commission 2019). Similar no-fault systems are also operated in place of medical negligence, in countries including New Zealand and Sweden. These allow for accountability (understood as redress or compensation) in medical decision-making without necessarily relying on domain-specific (or domain-agnostic) standards which actor behaviour needs to be checked against. However, a detailed technical legal discussion on the merits of strict or product liability approaches is beyond the scope of the chapter – as such we shall leave that discussion there.

06.3 Black-Box AI and Accountability

Having now looked at the different ways in which 'accountability' has been used in the AI ethics literature, we can turn to our main question: Does the use of 'black-box' AI models compromise accountability? So far, we've only considered the different types of accountability that the use of AI systems is said to compromise, not *how* their use is argued to compromise them. We can now consider the main ways in which opacity can be, and has been, argued to compromise the accountable development and use of AI systems.

Opacity is argued to compromise accountability because deep learning models make many of the existing, domain-specific standards (in this case for medical decision-making) inapplicable, as these standards rely on being able to inspect the decision-making logic of operators. For instance, standards governing how doctors should treat patients for the best medical outcomes, or standards governing how patients should be involved in clinical decision-making to best realise shared decision-making in the clinical encounter, or around how models should be built to not be discriminatory or biased against any demographic. These are the same domain-specific standards that are used to adjudicate negative responsibility (as discussed in section 2.1), to provide guidance regarding positive responsibility (as discussed in section 2.2), and to provide some avenues for redress through the legal adjudication of negative responsibility (for instance, through

determination of medical negligence, as discussed in section 2.3). If the use of deep learning models in medicine *does* make *all* standards guiding these determinations inapplicable, then that would indeed compromise their accountable use.

Essentially, the argument for this has been made as follows:

> **P1:** The use of deep-learning models bars epistemic access to the decision-making logic of the model
>
> **P2:** For current standards guiding domain-specific behaviour to be applied, they need epistemic access to the decision-making logic of the model.
>
> **P3:** If current standards guiding domain-specific behaviour are not applicable, accountability (in the use of deep-learning models) is disturbed.
>
> ---
>
> **C:** The use of deep-learning models disturbs accountability

P1 and P3 are relatively straightforward premises, given what we've discussed so far. P1 essentially points to the non-interpretable nature of deep-learning models, with the multitude of hidden layers that contribute to a high-dimensional reasoning process that is not comprehensible for us. We are thus barred from knowing what the specific decision-making logic was that was used by the model to generate the recommendation – in our case diagnostic or treatment-related. P3 similarly seems to follow straightforwardly from our discussion in section 2. If domain-specific standards are not applicable or available, we struggle to adjudicate negative responsibility, to promote positive responsibility, and in some cases to provide redress – all crucial aspects of accountability. It is seemingly therefore P2 where most of the action is.

As an example, if a clinician is using a deep-learning model to assist in a diagnosis of acute stroke by analysing CT brain scans, and the model provides a high-probability classification that there is indeed a risk of stroke (NICE 2022), how is the clinician supposed to take as evidence this recommendation? Current guidance for what the standard of care might be, either as taught in medical schools or communicated by experts, might be to look for specific kinds of hypoattenuation, or effacement to the cerebral sulci. Importantly, when consulting a colleague with a different perspective on why they believe the risk of stroke is low, why the same features in the scan are non-conclusive, the clinician would be required to understand their colleague's reasoning about these features and the extent of their instantiation in the scan, and how that connects to the biological mechanisms that could be causing them. Whatever the standard of care is, the domain-specific standard in this case, judging whether or not the clinician has acted appropriately would require an understanding of their reasoning process, or in the case of a second opinion, the reasoning process of their colleague and how the original clinician took that as evidence. Current standards of how to diagnose and treat strokes require such epistemic access, and in its absence, they would be hard-pressed to have a point of view on whether the action taken was appropriate or not.

This holds regardless of whether the decision in question is purely a medical one, as above, or even if it is an ethical one. To the extent that one of the ethical goals for a clinical encounter, or more generally the doctor-patient relationship, is to make space for shared decision-making, AI has been argued to disrupt that as well due to its opacity (Bjerring and Busch 2020). The argument here similarly is that current guidelines for promoting shared decision-making require the doctor to understand and communicate to the patient the reason why certain diagnostic or treatment recommendations are being provided, and with deep learning models, this is not possible. The domain-specific standard in this case, what it takes for shared decision-making to

obtain, is similarly not being met by the 'black-box' nature of the model, undermining accountability. Similar arguments can be made about the ability of model developers to build models to be fair, if we cannot tell whether 'protected' features about patients are being considered by the model in its reasoning. This is the essence of P2.

Does the overall argument succeed on the merits of P1, P2, and P3? It depends on the domain-specific standard in question. As the argument is posed right now, it doesn't operate at a sufficient level of granularity. The argument has to be considered for each of the most important domain-specific standards in question. We've made reference to two already – the standard of care which comprises medical decision-making for different tasks such as stroke diagnosis, and the standard of care around what satisfies the requirements of shared decision-making. There are probably many others as well, concerning for instance discriminatory or biased decision-making in clinical settings, protection of patient privacy and what that requires, etc.

Let us use the two we've already covered as an initial starting point to evaluate the argument. As argued above, it is sometimes true that for the standard of medical care to apply, an understanding of the model's decision-making logic in arriving at the recommendation is necessary. This standard of care requires medical explanations and reasoning, which invoke medical and physiological concepts, connecting them through mechanistic, causal explanations to arrive at a final judgment about a diagnosis or treatment recommendation for the patient. To draw on a previous example, a doctor could look at a brain CT scan, identify areas of hypoattenuation or sulcal effacement, reason about the specific features by connecting them with a best explanation for what are the biological processes that could be causing them to show up in a scan, and identify the physiological mechanisms which may be driving these biological events (stroke or no stroke). Using such a causal chain operating at the level of medical and biological concepts, the doctor can thus make a diagnosis.

To understand why such standard of care is important for accountability, we only need consider how it features in adjudications of medical negligence. In legal settings, a necessary step to showing that a clinician has been negligent in the care that they have provided is to show that they have breached their duty of care, and not performed up to the expected standard of care. The standard of care is generally understood as the conduct of a medical professional exercising reasonable skill and care, and a test for breach would be if "the error was one which would be made by a professional exercising reasonable skill and care" (Herring 2018). However, in most judgments concerned with clinical work, this more general requirement has been replaced by a more specific test known as the *Bolam* test, which states that,

> … [a] doctor is not guilty of negligence if he has acted in accordance with a practice accepted as proper by a responsible body of medical men skilled in that particular art (*Bolam v Friern* 1957).

Essentially, a medical professional is assessed (in most cases) on the charge of having breached their duty of care by expert witnesses, which are other medical professionals who're deemed competent. Furthermore, these experts have to be competent in making the particular judgment that the medical professional has been allegedly negligent about. If there are experts representing a responsible body of medical opinion that find the judgment of the professional to be one that they themselves would have made, the *Bolam* test would find the professional not guilty of medical negligence on the grounds that they did not breach their duty of care.

If the standard of care is thus one that operates on the mechanistic, medical reasoning outlined above, as it does right now, it is this same standard of care that will be expressed by experts in

adjudications of medical negligence – the legal determination of negative responsibility. However, if there is an element of the overall process where this access to medical reasoning and explanations is blocked, through the use of deep learning models, it is unclear how existing standards might apply – or so the argument goes. And if these existing standards can't apply, then this might disrupt the adjudication of negative responsibility, as experts might have insufficient grounds to assess whether the doctor in question acted reasonably in using the deep learning model. P2 thus seems plausible, as does P3, in that deep learning models disrupt accountability for the domain of medical decision-making of this sort.

However, this is only true if there isn't an alternative standard of care that is compatible with pockets of opacity in the system. While deep learning models may not provide medical explanations of the sort above, they do provide other information that may be used to understand their reliability (and the reliability of their recommendations) – for instance, they might provide a confidence level associated with their recommendation, and developers might also provide summary statistics about their performance in the past on similar tasks (NICE 2022).**Error! Bookmark not defined.** Their performance relative to other doctors, of varying levels of seniority, might also be provided (Liu et al. 2019). These statistical explanations, rather than medical explanations, can still perhaps be used to reason about the extent to which the model's recommendation should be relied upon. If we can, for instance, arrive at a consensus for how disagreement with a model should be resolved – perhaps by updating for the model's recommendations through Bayesian conditionalization in a way similar to how medical test results are updated for – and if this way of reasoning over the model's recommendations was empirically shown to be produce patient health outcomes that were at least comparable, then this would ground an alternative standard of care that would now be applicable (Chapter 4, this dissertation). This standard of care could then be used even for something like adjudicating medical negligence, as expert witnesses would now be able to judge whether a doctor acted reasonably by the lights of this standard of care, without being impeded by lack of access to the model's decision-making logic. P2 might thus be true in that current standards are frustrated by opacity, but P3 would be false because we might create or discover new standards of care that need not be, and wouldn't disturb accountability. However, if we can't discover such alternative standards that do produce at least comparable results when it comes to patient outcomes, then the argument that opacity disturbs accountability does succeed, at least for this domain.

What about the domain of standards for promoting shared decision-making in the clinical encounter? Here, there are strong reasons to think that P2 itself is false for most (if not all cases) where shared decision-making is necessary (Chapter 5, this dissertation). In most cases, black-box models don't occlude what patients are actually concerned with – the outcomes, risks, benefits, and uncertainties around clinical decisions are available to the doctor to communicate to the patients even if the doctor can't communicate the exact decision-making logic the model used to arrive at the recommendation. For patients who are more epistemically active and want to reason alongside the doctor, they can similarly utilize either the statistical explanations, or the 'best-guess' post-hoc explanations (described in section 3.1). The argument that opaque, deep learning models thus impede our ability to promote shared decision-making in clinical settings, is thus just based on a misunderstanding of what the standards for shared decision-making require of us. The disturbance to accountability is thus similarly minimal.

Perhaps similar arguments can be made for other domain-specific standards involved in the use of AI in clinical practice – that either these standards are compatible with opaque models, or where current standards are not, updated standards might be. We might even argue that P1 also doesn't stand up to scrutiny, because access to explanations is still possible (though not of the highest fidelity) with deep learning models through post-hoc explainability techniques which

might be fit for this purpose. After all, it is unclear how representative the communicated reasoning of human doctors is (for adjudications of negative responsibility) of the actual reasoning employed in making the medical decision. With these arguments, the strength of the argument of opacity disturbing accountability is blunted. It is of course possible that alternative standards of care might not be realized, as alternative methods of reasoning over the outputs of black-box systems might not perform at the same level as traditional methods of reasoning over the outputs of interpretable systems, but this is ultimately an empirical question, and needs to be decided as such. In advance of that, the force of the argument seems overstated.

06.4 Black-box AI and Responsibility Gaps

There is one further issue that has been discussed in the literature that might arise afresh even with these arguments against the use of deep learning systems disturbing accountability – that of responsibility gaps. Responsibility gaps have been discussed in the literature regarding the use of neural network-type AI systems (the fundamental architecture of 'black-box' deep learning models) at least as early as 2004 (Mathias 2004), possibly even earlier. They refer to a gap in the actor-landscape for agents that can reasonably be held responsible for injurious or other adverse outcomes from the use of such systems, due to the inability of any actors to control or predict the outcomes resulting from the use of such systems – essentially the inability to satisfy the domain-agnostic control and epistemic conditions discussed above. While these arguments have traditionally been raised for the use of lethal autonomous weapon systems (LAWS), they may carry over to the use of such AI even in medical decision-making, as we shall see (Sparrow 2007, Sparrow 2016).

While a full rehearsal of all the arguments for and existence of responsibility gaps is beyond the scope of this chapter, I will briefly characterize and consider the issue, and offer some thoughts based on the analysis done so far. Traditionally, discussion on responsibility gaps has focused on the adjudication of negative responsibility. A responsibility gap can therefore be characterized as a deficit between installed and accepted mechanisms and frameworks to adjudicate and attribute responsibility for adverse events (along with appropriate penalties), and the total possible space of events and actions leading to harmful conduct.

It's useful to sharpen this further by considering two ways of framing such gaps. First, we might contrast a *weak* v. *strong* framing of responsibility gaps – a weak framing is one where such gaps exist if no actor has departed from any of the domain-agnostic or domain-specific standards discussed above, and harm is still done. We might imagine this as patient injuries occurring even with the doctors using AI systems as the accepted standard of care currently requires, the model developers building these systems according to currently accepted norms of development, etc. A strong framing on the other hand requires such gaps to exist not just with existing standards but with *ideal* standards – standards which are not just as a matter of fact accepted by the communities, but standards which *should* be accepted as they incorporate all of the latest empirical results that would substantially change the domain-specific guidelines on how doctors and model developers should behave. We might imagine a difference in current and ideal standards if, for instance, empirical discoveries affecting the existing standard of care have been made (for instance, how strokes should be treated), but have not yet been disseminated or accepted at large within the community.

The other way of framing such gaps is by subjectively perceiving them as *intuitively acceptable* v. *intuitively unacceptable* responsibility gaps. Intuitively acceptable gaps might be those which are seen as ones where no agent need be held responsible, and the harm is seen as largely unavoidable – or at the very least, as one where it's not obvious that someone objectionably caused it.

Intuitively unacceptable gaps, on the other hand, might be seen as those where someone needs to be held responsible.

These two approaches to framing can produce a 2x2 matrix to better structure our thoughts around responsibility gaps:

	Intuitively Acceptable RGs	**Intuitively Unacceptable RGs**
Weak RGs	RG results from a mistaken belief that current domain-specific standards are up-to-date, and nothing further could have been done, no one can be held responsible.	RG results from a justified belief that someone *can* be held responsible, that current domain-specific standards are not up-to-date, and they should be updated to feature the best new approaches, or best approaches given changing circumstances (e.g. use of deep learning models)
Strong RGs	RGs are not ethically problematic, as far as adjudication of negative responsibility, as nothing *truly* could have been done by anyone to avoid the harm (e.g. patient injury because no better techniques exist, and there are risks even for the best available treatment option for a particular condition)	RGs because the harm could not have been avoided, but someone still needs to be held responsible. As an example, Sparrow's claim that someone must be held responsible for all actions taken in a military conflict, following from the tenets of *jus in bello***Error! Bookmark not defined.**

With such a characterization we might situate the discussion of responsibility gaps relevant to our discussion in this chapter as follows. If legacy medical reasoning, through the use of medical explanations and medical/biological concepts, is inapplicable to the use of opaque AI systems, it creates weak responsibility gaps – either ones where we may not realise that the use of traditional medical reasoning, or of opaque systems, needs to be changed ('intuitively acceptable'), or ones where we realise that they do ('intuitively unacceptable).

However, if the standard of care for the medical use of AI systems is updated, perhaps as outlined in section 3.2, this might yield strong responsibility gaps. Whether or not these gaps are deemed acceptable or not, for instance whether there are principles like *jus in bello* for medical care that require someone to be held responsible even if there truly was nothing that could have been done differently, is what there has been relatively little discussion on, and what we might need to consider further given the analysis in this chapter. The resulting responsibility gaps from shifting domain-specific standards, which take us from weak to strong gaps, might for instance enlarge the total set of outcomes that responsibility cannot be attributed for (however this enlargement of such outcomes is measured).[30] Alternatively, while the total set of outcomes of the strong gap may not be larger than of the weak gap (from before the updated standards), the

[30] Importantly, this is different from saying that the total set of adverse outcomes is enlarged. If it were so, the standard of care wouldn't have shifted. The claim here is merely that although the total set of adverse outcomes might be smaller, the total absolute number of outcomes where no one can be held responsible might be larger.

strong responsibility gap might systematically affect certain decisions, or certain demographics, more. In both cases, there seems to be something ethically objectionable about the new gap, even if nothing could have been done to avoid the outcome.

Addressing such ethically objectionable shifts in responsibility gaps would thus become important. However, there may be multiple ways of addressing these shifts. One option might be to roll back the domain-specific standards to what they were pre-update – this might mean continuing with medical reasoning that invokes medical and biological concepts, even if such concepts may not be accessible through the AI's decision-making logic. However, as discussed earlier, any such shift needs to be an empirical one, based on the impact on patient health outcomes. To insist that bridging a responsibility gap is more important than better outcomes overall for patients seems backwards. There have also been other approaches discussed, such as 'blank check' responsibility, where an actor of sufficiently high standing can accept responsibility for the outcome even if they are not causally linked to them (Champagne et al. 2015). Such approaches deserve further consideration.

In general, we might understand responsibility gaps as generally problematic for three reasons. First, their presence might incentivize negligent or overtly harmful behaviour, as mentioned by Sparrow.**Error! Bookmark not defined.** Second, there seems to be something ethically problematic about no one being held responsible, if someone actually can be. Third, in many cases, attributing responsibility (and thus bridging the gap) is the only way that injured parties can be compensated for their injuries and obtain redress – medical negligence in most countries, including the UK, requires that a clinician be found negligent by breaching their duty of care. Responsibility gaps can therefore restrict damages paid to injured patients or their families. A solution for responsibility gaps, at least in that they might impede redress, would be to consider no-fault compensation system in the use of AI systems in healthcare. This, as mentioned above, is already being done in jurisdictions including Sweden and New Zealand, and would at least blunt some of the impact of any responsibility gaps that do obtain with the use of AI, either with existing domain-specific standards or updated ones.

06.5 Conclusion

The goal of the discussion in this chapter has been to offer an alternative to the majority view, in discussions of AI ethics, that the use of opaque deep learning models fundamentally compromises accountability. It does so by (1) disentangling the different ways in which accountability is used in the AI literature, (2) situating these different conceptions or 'faces' of accountability in existing literature across different fields, (3) making explicit the actual argument for how black-box systems might undermine these different conceptions of accountability, and (4) showing why that argument fails.

The key insight that grounds the blunting of the argument from opacity disturbing accountability really is as follows: the standards that govern the behaviour of clinicians when using such systems either do not require the decision-making logic of models to be transparent (in the case of shared decision-making), or if they do they can be replaced by alternative, updated standards that are compatible with the decision-making logic being incomprehensible. In such cases, the proof is in the pudding, and I have shown how both these claims can be true in previous chapters (Chapters 4 and 5).**Error! Bookmark not defined.Error! Bookmark not defined.** Further work is required to validate both of these claims, but their initial plausibility at least points to opacity not being the death-knell for accountability that it is so often made out to be.

07. Conclusion

The previous three chapters – chapters 4, 5, and 6 – stand as three case-studies. In each case study, a single thematic objection is presented, to the justifiable use of black-box, deep-learning models in clinical settings. This objection is fleshed out in its strongest argument, and then followed by one of two strategies – it is either shown that

(1) such an objection relies on a fundamental mischaracterisation of the key principle it invokes – for instance, in the case of the need for medical explanations for shared decision-making (chapter 5) – in a way such that it actually poses no problems for opaque decision models, or that

(2) such an objection initially poses a legitimate concern for the use of opaque models – such as in the case of how such models might epistemically appropriately be relied upon (chapter 4), or how they might still be accountably used (chapter 6), in the absence of our ability to inspect their reasoning – but that alternative methodologies can be developed (for instance an updated Bayesian approach based on historical relative accuracy of models) to mitigate this concern.

The broad goal of the dissertation is to show that if an adequate standard of care for the use of such models can be illustrated, through each of these case studies, and it can be shown that such models can be used in a way that is ethical as well as leads to good clinical outcomes, we do not need to accept the necessity of explanations (primarily through the use solely of interpretable models in medicine). In this way, the interpretability – performance trade-off doesn't necessarily create a justifiability – performance trade-off, where 'justifiability' refers to ethically and epistemically defensible use of these models.

To be clear, this work merely illustrates the possibility of such a standard of care obtaining, at some point in the future, with suggestions presented in the various chapters as to how that might happen. It does not definitively argue that we are currently ready to adopt non-interpretable, opaque systems in an ethically and epistemically defensible way – much work will need to be done to develop and test appropriate reliance strategies as mentioned in chapter 4, for instance. Many of the ethical concerns raised in the survey within chapter 2 may also need to be individually tackled as the preceding chapters tackle a single thematic objection systematically. Objections from bias and unfairness, for instance, need to be mitigated by showing that black-box models can still be inspected to detect such bias, and that we can implement measures (either at the model-level or the larger AI-human interaction level) to mitigate such bias. Finally, there might be some cases where there is no substitute for interpretability – for instance, in the value that model developers might get for building better and better models. In such a case, the benefits of interpretability will need to be weighed against the human and health costs of not deploying more performant systems earlier.

However, even with these outstanding issues and analysis needed, I believe that this thesis still makes an important contribution to structuring and presenting a framework to make the case how the value of black-box systems might be preserved in clinical settings, and making the first few arguments within that framework. I hope that future work continues carefully further utilising this (and complementary frameworks), given the value that such complex systems might bring.

08. Bibliography

42 CFR § 121.4 - OPTN policies: Secretarial review and appeals.

Angwin J, Larson J, Mattu S, Kirchner L. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica* 23.

Ardila, D., Kiraly, A.P., Bharadwaj, S. *et al.* 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med* **25,** 954–961. https://doi.org/10.1038/s41591-019-0447-x

Arterys -FDA approved deep learning application: https://www.arterys.com/solutions

Awad A, Bader-El-Den M, McNicholas J, Briggs J. 2017. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. Int J Med Inform. 108:185–195. doi: 10.1016/j.ijmedinf.2017.10.002.

Awad, E., Dsouza, S., Kim, R. *et al.* 2018. The Moral Machine experiment. *Nature* **563,** 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775–779.

Barry, M. J., & Edgman-Levitan, S. (2012). Shared decision making—The pinnacle patient-centred care. *N Engl J Med* 2012; 366:780-781. DOI: 10.1056/NEJMp1109283

Beauchamp, Tom L., 2010. Autonomy and consent. In *The Ethics of Consent*, F. G. Miller and A. Wertheimer (eds.), New York: Oxford University Press.

Beauchamp, Tom, and Childress, James. Principles of Biomedical Ethics. 7th edition (2012). Oxford University Press.

Bengio, Y. et al. 2019. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. arXiv:1901.10912 [cs.LG]

Bjerring, J.C., Busch, J. 2020. Artificial Intelligence and Patient-Centred Decision-Making. *Philos. Technol.*. https://doi.org/10.1007/s13347-019-00391-6

Bojikian KD, Lee CS, Lee AY. 2019. Finding Glaucoma in Color Fundus Photographs Using Deep Learning. *JAMA Ophthalmol.* 137(12):1361–1362. doi:10.1001/jamaophthalmol.2019.3512

Bolam v Friern Hispotal Management Committee [1957]

Bovens, M. (2007), Analysing and Assessing Accountability: A Conceptual Framework. European Law Journal, 13: 447-468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Bovens, Mark (2010) Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism, West European Politics, 33:5, 946-967, DOI: 10.1080/01402382.2010.486119

Braddock, C H 3rd et al. "Informed decision making in outpatient practice: time to get back to basics." *JAMA* vol. 282,24 (1999): 2313-20. doi:10.1001/jama.282.24.2313

Bradford, G. – "Hard to Know", *Responsibility: The Epistemic Condition* (eds. Robichaud, Wieland), Oxford University Press (2017)

Brown M.D., M.J. Reeves. 2003. Interval likelihood ratios: Another advantage for the evidence-based diagnostician. *Annals of Emergency Medicine*, 42 pp. 292-297, 10.1067/mem.2003.274

Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data Soc. 2016;3(1):1–12.

Bussone A., S. Stumpf and D. O'Sullivan. 2015. "The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems," *2015 International Conference on Healthcare Informatics*, Dallas, TX, pp. 160-169.

C the Signs, a multiplatform tool for early identification of patients at risk of different types of cancer. This has model been adopted by the NHS: https://www.england.nhs.uk/cancer/case-studies/c-the-signs-how-artificial-intelligence-ai-is-supporting-referrals/ (2019)

Castelvecchi, D. – "Can we open the black box of AI?", *Nature*, 538, 20-23 (2016)

Champagne, M., Tonkens, R. Bridging the Responsibility Gap in Automated Warfare. Philos. Technol. 28, 125–137 (2015). https://doi.org/10.1007/s13347-013-0138-3

Charles, C et al. "Decision-making in the physician-patient encounter: revisiting the shared treatment decision-making model." *Social science & medicine* vol. 49,5 (1999): 651-61. doi:10.1016/s0277-9536(99)00145-8

Charles, C et al. "Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango)." *Social science & medicine (1982)* vol. 44,5 (1997): 681-92. doi:10.1016/s0277-9536(96)00221-3

Christensen, D. – "Higher-Order Evidence", *Philosophy and Phenomenological Research*, Vol. LXXXI No. 1 (2010)

Christensen, D. (2016). Disagreement, drugs, etc.: from accuracy to akrasia. *Episteme, 13*(4), 397-422. doi:10.1017/epi.2016.20

Christensen, David. 2016. Conciliation, Uniqueness and Rational Toxicity. *Nous* 50 (3): 584–693.

Coleman, Jules, Scott Hershovitz, and Gabriel Mendlow, "Theories of the Common Law of Torts", *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/tort-theories/>.

Corbett-Davies S., Goel S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

Dastin, Jeffrey. (2018) Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* Oct. 10. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (accessed 22 April 2020)

David Lyell, Enrico Coiera. 2017. Automation bias and verification complexity: a systematic review, Journal of the American Medical Informatics Association, Volume 24, Issue 2, March 2017, Pages 423–431, https://doi.org/10.1093/jamia/ocw105

De Andrade et al. 2018. Ethics and Artificial Intelligence: Suicide Prevention on Facebook. Philosophy & Technology, 31, 669-684 (2018)

De Fauw, J., Ledsam, J.R., Romera-Paredes, B. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* **24,** 1342–1350 (2018). https://doi.org/10.1038/s41591-018-0107-6

de Miguel Beriain, I. 2020. Should we have a right to refuse diagnostics and treatment planning by artificial intelligence?. *Med Health Care and Philos*. https://doi.org/10.1007/s11019-020-09939-2

DeepMind, referencing the 'DeFauw et al. (2018)' paper: https://deepmind.com/applied/deepmind-health/working-partners/health-research-tomorrow/moorfields-eye-hospital-nhs-foundation-trust/

Dickerson, J. P., and Sandholm, T. 2015. FutureMatch: Combining human value judgments and machine learning to match in dynamic environments. In AAAI, 622–628.

Díez-Sanmartín, C.; Sarasa Cabezuelo, A. 2020. Application of Artificial Intelligence Techniques to Predict Survival in Kidney Transplantation: A Review. *J. Clin. Med. 9*, 572.

Doshi-Velez, Finale, and Mason Kortz. 2017. Accountability of AI Under the Law: The Role of Explanation. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper.

Easwaran, K., Fenton-Glynn, L., Hitchcock, C., & Velasco, J.D. (2016). Updating on the Credences of Others: Disagreement, Agreement, and Synergy. *Philosopher's Imprint, 16*.

Edwards, Lilian; Veale, Michael. Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For, 16 *Duke Law & Technology Review* 18-84 (2017)

Elga, A. 2007. 'Reflection and Disagreement.' Noûs, 41: 478–502.

Elwyn, G et al. "Shared decision making: developing the OPTION scale for measuring patient involvement." *Quality & safety in health care* vol. 12,2 (2003): 93-9. doi:10.1136/qhc.12.2.93

Emanuel, E J, and L L Emanuel. "Four models of the physician-patient relationship." *JAMA* vol. 267,16 (1992): 2221-6.

Entwistle, V et al. "Broad versus narrow shared decision making: Patients' involvement in real world contexts" in *Shared Decision Making in Health Care: Achieving evidence-based patient choice* (eds. Elwyn et al.) (2016) Oxford Scholarship Online

Eshleman, A. – "Moral Responsibility", *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), E. N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/>.

Esteva, A., Kuprel, B., Novoa, R. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542,** 115–118 (2017). https://doi.org/10.1038/nature21056

European Commission, Directorate-General for Justice and Consumers, *Liability for artificial intelligence and other emerging digital technologies*, Publications Office, 2019, **https://data.europa.eu/doi/10.2838/25362**

Eva, Benjamin. 2019. The Principle of Indifference. *The Journal of Philosophy* Vol. 116, Iss. 7, July 2019

Ezekiel J. Emanuel, Linda L. Emanuel. What Is Accountability in Health Care?. Ann Intern Med.1996;124:229-239. doi:10.7326/0003-4819-124-2-199601150-00007

Faden, Ruth, and Beauchamp, Tom. A History and Theory of Informed Consent. (1986) Oxford University Press, New York.

Ferrari, Filippo and Nikolaj J. L. L. Pedersen , "Epistemic Peer Disagreement", in *The Routledge Handbook of Social Epistemology* ed. Miranda Fricker , Peter J. Graham , David Henderson and Nikolaj J. L. L. Pedersen (Abingdon: Routledge, 07 Aug 2019), accessed 09 May 2020 , Routledge Handbooks Online.

Fischer, J.M.; Tognazzini, N. – "The Truth about Tracing", *Nous*, 43:3 531-556 (2009)

FitzPatrick, W.J. – "Moral Responsibility and Normative Ignorance: Answering a New Skeptical Challenge", *Ethics* 118: 589-613 (2008)

Flechet M, Falini S, Bonetti C, et al. 2019. Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKI predictor. Crit Care. 23:282. doi: 10.1186/s13054-019-2563-x

Frances, Bryan – *Disagreement*, Cambridge , UK: Polity Press (2014)

Frankfurt, Harry G., 1969, "Alternate Possibilities and Moral Responsibility", The Journal of Philosophy, 66(23): 829–839. Reprinted in Fischer 1986, pp. 143–52; in Frankfurt 1988, pp. 1–10; and in Widerker and McKenna 2003, pp. 17–25. doi:10.2307/2023833

Freedman, Rachel ; Borg, Jana Schaich ; Sinnott-Armstrong, Walter ; Dickerson, John P. & Conitzer, Vincent (2020). Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence* 283:103261.

Froomkin AM, Michael Froomkin A, Kerr IR, Pineau J (2018) When AIs outperform doctors: the dangers of a tort-induced over-reliance on machine learning and what (not) to do about it. SSRN Electron J. https://doi.org/10.2139/ssrn.3114347

Gerke S, Yeung S, Cohen IG. 2020. Ethical and Legal Aspects of Ambient Intelligence in Hospitals. *JAMA.* 323(7):601–602. doi:10.1001/jama.2019.21699

Goldman, A. (2001). Experts: Which Ones Should You Trust? *Philosophy and Phenomenological Research, 63*(1), 85-110. doi:10.2307/3071090

Gotterbarn D., 2001. "Informatics and professional responsibility," *Science and Engineering Ethics*, 7(2): 221–230.

Grote, Thomas, and Philipp Berens. "On the ethics of algorithmic decision-making in healthcare." *Journal of medical ethics* vol. 46,3 (2020): 205-211. doi:10.1136/medethics-2019-105586

Guerrero, A.A. – "Intellectual Difficulty and Moral Responsibility", *Responsibility: The Epistemic Condition* (eds. Robichaud, Wieland), Oxford University Press (2017)

Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA.* 2016;316(22):2402–2410. doi:10.1001/jama.2016.17216

Gutierrez, G. 2020. Artificial Intelligence in the Intensive Care Unit. *Crit Care* **24,** 101. https://doi.org/10.1186/s13054-020-2785-y

Haque A, Guo M, Alahi A, et al. 2017. Towards vision-based smart hospitals: a system for tracking and monitoring hand hygiene compliance. *Proc Mach Learn Res.* 68:75-87.

Herring, J et al. – "Elbow Room for Best Practice? Montgomery, Patient Values, and Balanced Decision-making in Person-centred Clinical Care", *Medical Law Review*, Vol. 25, No. 4, pp. 582-603 (2017)

Herring, J. – *Medical Law and Ethics 7th ed.*, Oxford: Oxford University Press (2018)

Hoff, Timothy. 2011. Deskilling and adaptation among primary care physicians using two work innovations, Health Care Management Review: October-December 2011 - Volume 36 - Issue 4 - p 338-348 doi: 10.1097/HMR.0b013e31821826a1

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873

Jonas, H. – *The Imperattive of Responsibility. In search of an Ethics for the Technological Age*, Chicago: The Chicago University Press (1984)

Joshi, I. 2019. Waiting for deep medicine. The Lancet, Vol. 292, Iss. 10177, March 2019, pp. 1193-1194. DOI: https://doi.org/10.1016/S0140-6736(19)30579-3

Joshua A. Kroll , Joanna Huey , Solon Barocas , Edward W. Felten , Joel R. Reidenberg , David G. Robinson & Harlan Yu *Accountable Algorithms*, 165 U. Pa. L. Rev. 633 (2017). Available at: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3

Kaplan, Jared, et al. *Scaling Laws for Neural Language Models.* arXiv, 22 Jan. 2020. *arXiv.org*, https://doi.org/10.48550/arXiv.2001.08361.

Kaur Davider, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. Trustworthy Artificial Intelligence: A Review. ACM Comput. Surv. 55, 2, Article 39 (February 2023), 38 pages. https://doi.org/10.1145/3491209

Kelly, T. (2005) "The Epistemic Significance of Disagreement," *Oxford Studies in Epistemology* 1: 167–96.

Kelly, T. 2010. 'Peer Disagreement and Higher-Order Evidence.' In R. Feldman and T. Warfield (eds), Disagreement, pp. 111–74. Oxford: Oxford University Press.

Khalid S, Goldenberg M, Grantcharov T, Taati B, Rudzicz F. 2020. Evaluation of Deep Learning Models for Identifying Surgical Actions and Measuring Performance. *JAMA Netw Open.* 3(3):e201664. doi:10.1001/jamanetworkopen.2020.1664

Kleinberg J, Mullainathan S, Raghavan M. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:160905807 [cs, stat] [Internet]. 2016 Sep 19; Available from: http://arxiv.org/abs/1609.05807

Koppell, Jonathan G.S. (2005). 'Pathologies of Accountability: ICANN and the Challenge of "Multiple Accountabilities Disorder"', Public Administration Review, 65:1. 94–107.

Krishnan, M. Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning. *Philos. Technol.* (2019). https://doi.org/10.1007/s13347-019-00372-9

Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017; **60:** 84–90.

Kukla, Rebecca. "How do patients know?." *The Hastings Center report* vol. 37,5 (2007): 27-35. doi:10.1353/hcr.2007.0074

Ladd. J., 1989. "Computers and Moral Responsibility. A Framework for an Ethical Analysis," in C.C. Gould (ed.), *The Information Web. Ethical and Social Implications of Computer Networking*, Boulder, Colorado: Westview Press, pp. 207–228.

Liang, H., Tsui, B.Y., Ni, H. *et al.* Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med* **25,** 433–438 (2019). https://doi.org/10.1038/s41591-018-0335-9

Liu et al. – "Detecting Cancer Metastases on Gigapixel Pathology Images", arXiv: 1703.02442 (2017)

Liu, L. Faes, A.U. Kale, S.K. Wagner, D.J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit. Health, 1 (6) (2019), pp. e271-e297

Lobo et al. – "AUC: a misleading measure of the performance of predictive distribution models" *Glob. Ecol. Biogeogr.* 17, 145-151 (2007)

London, Alex John. 2019. "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability," Hastings Center Report 49, no. 1: 15-21. DOI: 10.1002/hast.973

Long, R. 2020. Fairness in machine learning: Against false positive rate equality as a measure of fairness. (ms)

Lougheed K. (2020) An Analysis of Epistemic Peerhood. In: The Epistemic Benefits of Disagreement. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 51. Springer, Cham

Maddox TM, Rumsfeld JS, Payne PRO. 2019. Questions for Artificial Intelligence in Health Care. JAMA ;321(1):31–32. doi:10.1001/jama.2018.18932

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology,* 6(3), 175–183.

Maya Sherman & Dov Greenbaum (2019) Ethics of AI in Transplant Matching: Is It Better or Just More of the Same?, The American Journal of Bioethics, 19:11, 45-47, DOI: 10.1080/15265161.2019.1665734

McDougall, Rosalind J. "Computer knows best? The need for value-flexibility in medical AI." *Journal of medical ethics* vol. 45,3 (2019): 156-160. doi:10.1136/medethics-2018-105118

Mirnezami, R; Ahmed, A. 2018. "Surgery 3.0, artificial intelligence and the next-generation surgeon," British Journal of Surgery, vol. 105, no. 5, pp. 463–465.

Mishra, Abhishek. 2019. Responsible Usage of Machine Learning Classifiers in Clinical Practice. *Journal of Law and Medicine* Vol. 27, Pt. 1.

Montgomery v Lanarkshire Health Board [2015] UKSC 11

Muller v Kings College Hospital [2017] EWHC 128 (QB)

Mulligan, T. (2019). The Epistemology of Disagreement: Why Not Bayesianism? *Episteme,* 1-16. doi:10.1017/epi.2019.28

Nagendran Myura, Chen Yang, Lovejoy Christopher A, Gordon Anthony C, Komorowski Matthieu, Harvey Hugh et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies BMJ 2020; 368 :m689 doi: https://doi.org/10.1136/bmj.m689

NHS. "When and where is shared decision making appropriate?" Accessed: 16 May 2021. Link: https://www.england.nhs.uk/shared-decision-making/when-and-where-is-shared-decision-making-appropriate/

NICE 2022: https://www.nice.org.uk/guidance/gid-dg10044/documents/final-scope

Noorman, Merel, "Computing and Moral Responsibility", *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/>.

Obermeyer Z. et al. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* Vol. 366 Iss. 6464: 447-453. DOI: 10.1126/science.aax2342

Panch, T., *et al.* 2019. Artificial intelligence: opportunities and risks for public health. *The Lancet Digital Health*, 1(1): PE13-E14. doi: https://doi.org/10.1016/S2589-7500(19)30002-0

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Ploug, Thomas, and Søren Holm. "The right to refuse diagnostics and treatment planning by artificial intelligence." *Medicine, health care, and philosophy* vol. 23,1 (2020): 107-114. doi:10.1007/s11019-019-09912-8

Poplin, R., Varadarajan, A.V., Blumer, K. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* **2,** 158–164 (2018). https://doi.org/10.1038/s41551-018-0195-0

Povyakalo, A. A., Alberdi, E., Strigini, L., & Ayton, P. (2013). How to Discriminate between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography. *Medical Decision Making*, *33*(1), 98–107. https://doi.org/10.1177/0272989X12465490

President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research. (1982). *Making health care decisions: A report on the ethical and legal implications of informed consent in the patient-practitioner relationship*. Washington, DC: U.S. Government Press

Price WN, Gerke S, Cohen IG. 2019. Potential Liability for Physicians Using Artificial Intelligence. JAMA. 322(18):1765–1766. doi:10.1001/jama.2019.15064

Pritchard, D. – "Anti-Luck Epistemology and the Gettier Problem", *Philosophical Studies* 172: 93-111 (2015)

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of internal medicine*, *169*(12), 866–872. https://doi.org/10.7326/M18-1990

Ravì, D. et al., "Deep Learning for Health Informatics," in IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 4-21, Jan. 2017, doi: 10.1109/JBHI.2016.2636665.

Reese, P. P., N. Boudville, and A. X. Garg. 2015. Living kidney donation: Outcomes, ethics, and uncertainty. *Lancet (London, England)* 385(9981): 2003–13. doi: 10.1016/S0140-6736(14)62484-3.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).

Rogers v Whitaker [1992] 11 WLUK 288

Rosen, G. – "Skepticism about Moral Responsibility", *Phiosophical Perspectives* 18: 295-313 (2004)

Ross et al. – "IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close", Boston: STAT, September 5 2017: https://www.statnews.com/2017/09/05/watson-ibm-cancer/

Roth, A. E.; Sonmez, T.; and Unver,M. U. 2004. Kidney exchange. *Quarterly Journal of Economics* 119(2):457–488.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead Nature Machine Intelligence. 2019:1.

S. Cave, R. Nyrup, K. Vold and A. Weller, 2019. "Motivations and Risks of Machine Ethics," in *Proceedings of the IEEE*, vol. 107, no. 3, pp. 562-574.

Saito et al. – "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 10, e0118432 (2015)

Sausville, Edward A. "Drug Discovery". *Principles of Clinical Pharmacology (Third Edition)* (2012) eds: Atkinson et al. Academic Press

Savulescu, J., Kahane, G. & Gyngell, C. 2019. From public preferences to ethical policy. *Nat Hum Behav* **3,** 1241–1243. https://doi.org/10.1038/s41562-019-0711-6

Schiff, Daniel, and Jason Borenstein. "How Should Clinicians Communicate With Patients About the Roles of Artificially Intelligent Team Members?." *AMA journal of ethics* vol. 21,2 E138-145. 1 Feb. 2019, doi:10.1001/amajethics.2019.138

Scholl, Isabelle et al. "Measurement of shared decision making - a review of instruments." *Zeitschrift fur Evidenz, Fortbildung und Qualitat im Gesundheitswesen* vol. 105,4 (2011): 313-24. doi:10.1016/j.zefq.2011.04.012

Schönberger, Daniel. 2019. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications, *International Journal of Law and Information Technology*, Volume 27, Issue 2, Pages 171–203, https://doi.org/10.1093/ijlit/eaz004

Selbst, Andrew D. 2017. Disparate Impact in Big Data Policing. *52 Georgia Law Review* 109 DOI: http://dx.doi.org/10.2139/ssrn.2819182

Shanafelt, Tait D. *et al.* 2010. Burnout and Medical Errors Among American Surgeons, Annals of Surgery: Volume 251 - Issue 6 - p 995-1000 doi: 10.1097/SLA.0b013e3181bfdab3

Shoemaker, David, 2011, "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility", *Ethics*, 121(3): 602–632. doi:10.1086/659003

Shoemaker, David, 2015, *Responsibility from the Margins*, New York: Oxford University Press. doi:10.1093/acprof:oso/9780198715672.001.0001

Simonite, T. 2018. When It Comes to Gorillas, Google Photos Remains Blind. *Wired.* URL: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/ (accessed 22 April 2020)

Smith, Holly – "Culpable Ignorance", *The Philosophical Review*, XCII, No. 4 (1983)

Sotoodeh, M., & Ho, J. C. (2019). Improving length of stay prediction using a hidden Markov model. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, *2019*, 425–434.

Sox, H et al. 2013. Medical Decision Making, second edition. Wiley Online Library, DOI: 10.1002/9781118341544

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, 24(1), 62–77.

Sparrow, R. (2016). Robots and respect. *Ethics and International Affairs*, 30(1), 93–116.

Strawson, P. F., 1962 [1993], "Freedom and Resentment", in *Proceedings of the British Academy*, 48: 1–25. Reprinted Fischer and Ravizza 1993b: 45–66.

Tetlock, P., Gardner, D. 2015. Superforecasting: The Art and Science of Prediction. Crown Publishers.

Topol, Eric J. – "High-performance medicine: the convergence of human and artificial intelligence", *Nature Medicine* 24, 44-56 (2019)

Traditional Decisional Conflict Scale. Accessed on 17 May 2021. Link: https://decisionaid.ohri.ca/docs/develop/Tools/DCS_English.pdf

Tsai, T. et al. 2003 Computer Decision Support as a Source of Interpretation Error: The Case of Electrocardiograms, *Journal of the American Medical Informatics Association*, Volume 10, Issue 5, Pages 478–483, https://doi.org/10.1197/jamia.M1279

Ubel, Peter A et al. "Autonomy: What's Shared Decision Making Have to Do With It?." *The American journal of bioethics : AJOB* vol. 18,2 (2018): W11-W12. doi:10.1080/15265161.2017.1409844

Vargas, M. – "The Trouble with Tracing", *Midwest Studies in Philosophy*, XXIX (2005)

Veatch, R M. "Models for ethical medicine in a revolutionary age. What physician-patient roles foster the most ethical realtionship?." *The Hastings Center report* vol. 2,3 (1972): 5-7.

Verghese A, Shah NH, Harrington RA. 2018. What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. JAMA. 319(1):19–20. doi:10.1001/jama.2017.19198

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." (2017).

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clinical Psychological Science*, *5*(3), 457–469. https://doi.org/10.1177/2167702617691560

Wartman, Combs. 2019. Reimagining Medical Education in the Age of AI. *AMA Journal of Ethics*, Feb 2019. Vol 21 (2): E146-152.

Watson, Gary, 1996 [2004], "Two Faces of Responsibility", *Philosophical Topics*, 24(2): 227–248. Reprinted in Watson 2004: 260–88. doi:10.5840/philtopics199624222

Whitney, Simon N et al. "A typology of shared decision making, informed consent, and simple consent." *Annals of internal medicine* vol. 140,1 (2004): 54-9. doi:10.7326/0003-4819-140-1-200401060-00012

Wieland, Jan Willem. 2017. "Introduction: The Epistemic Condition", in *Responsibility: The Epistemic Condition* (eds. Robichaud, Wieland), Oxford University Press

Wieringa, Maranke. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 1–18. DOI:https://doi.org/10.1145/3351095.3372833

Winters N, Venkatapuram S, Geniets A*, et al.* 2020. Prioritarian principles for digital health in low resource settings. *Journal of Medical Ethics* **46:**259-264.

Wolf, Susan, 1990, *Freedom Within Reason*, New York: Oxford University Press.

Yeung, S., Rinaldo, F., Jopling, J. *et al.* 2019. A computer vision system for deep learning-based detection of patient mobilization activities in the ICU. *npj Digit. Med.* **2,** 11. https://doi.org/10.1038/s41746-019-0087-z

Zimmerman, M.J. – "Moral Responsibility and Ignorance", *Ethics* 107: 410-26 (1997)

Zimmerman, M.J. – *Living with Uncertainty and the Moral Significance of Ignorance*. Cambridge University Press (2008)

Zou J., Schiebinger,L. 2018. 'AI can be sexist and racist – it's time to make it fair' 559 Nature 324–6.