Philosophy of Medicine
Perspectives

The Allure of Simplicity:
On Interpretable Machine Learning Models in Healthcare

Thomas Grote[1]

[1] Ethics and Philosophy Lab; Cluster of Excellence: Machine Learning—New Perspectives for Science, University of Tübingen, Tübingen, Germany. Email: thomas.grote@uni-tuebingen.de

Abstract

This paper develops an account of the opacity problem in medical machine learning (ML). Guided by pragmatist assumptions, I argue that opacity in ML models is problematic insofar as it potentially undermines the achievement of two key purposes: ensuring *generalizability* and *optimizing clinician–machine decision-making*. Three opacity amelioration strategies are examined, with *explainable artificial intelligence* (XAI) as the predominant approach, challenged by two revisionary strategies in the form of *reliabilism* and *interpretability by design*. Comparing the three strategies, I argue that interpretability by design is most promising to overcome opacity in medical ML. Looking beyond the individual opacity amelioration strategies, the paper also contributes to a deeper understanding of the problem space and the solution space regarding opacity in medical ML.

1. Introduction

The problem of opacity is key to understanding the necessary assurances for implementing machine learning (ML) models into clinical environments. Various studies highlight the potential benefit of ML models in assisting image-based medical diagnosis, predicting acute illness, or treatment planning (Esteva et al. 2017; Tomašev et al. 2019; Yala et al. 2022. However, not knowing what predictors are used by given ML models gives rise to concerns of malfunction, in what can be considered a high-stakes setting. While the problem is well established, there is little agreement on the appropriate opacity mitigation strategy—culminating in different paradigms.

So far, the prevailing strategy is to explain the functioning of ML models post hoc, either through statistical summaries or visualizations of the predictors for single instances (for a review, see Arrieta et al. 2020). Unlike explainable artificial intelligence (XAI), *revisionary strategies* seek to overcome the problem by replacing opacity with a different concept, deemed more crucial. One such candidate, *reliabilism,* is guided by the assumption that accuracy trumps explainability (London 2019; Durán and Jongsma 2021). Its rationale is a consistency argument: ML models should be evaluated according to the same evidential standards as (other) medical interventions. Conversely, this entails that certain kinds of

ISSN 2692-3963 | Date Received 10-11-2022 | Date Accepted 15-08-2023 | Date First Published 27-09-2023
Philosophy of Medicine | DOI 10.5195/pom.2023.139 | Volume 4 | Issue 1 | pp.1–24

opacity are acceptable—whether the underlying physiological mechanisms of a drug or model opacity—if the reliability and clinical benefit of a given intervention has been established.

The second revisionary strand has arisen out of skepticism regarding current XAI methods (for example, saliency maps) since they have been shown to be easily foolable in sanity checks (Adebayo et al. 2018). Hence, they are claimed to be (necessarily) misleading. Rather than deploying XAI methods, the proposed solution is to use models that are *interpretable by design* (Rudin 2019; Babic et al. 2021). This includes classes of models defined through certain structural properties (for example, sparsity and linearity), or those that possess built-in domain knowledge.

This paper's objective is to define the scope and to provide a (conditional) defense of using interpretable ML models within the context of clinical medicine. A secondary objective is to explore the dialectic among the different opacity mitigation strategies in order to gain a better understanding of the what the problem space is and what appropriate solutions are to mitigate opacity in medical ML.

Calls for interpretability are intuitively appealing. All else being equal, philosophers and scientists usually have a preference for simpler theories—deemed more elegant or closer to the truth (Genin 2018). That said, once we move beyond toy examples, there is a need to provide thorough grounding for model interpretability in medical ML. In particular, it is not yet adequately understood what the distinct epistemic properties of interpretable models are, and how their benefits and downsides compare to XAI or reliabilist approaches. To add nuance to the picture, I distinguish between two variants of interpretability by design, *strong* (where the idea is to use simple model architectures) and *moderate* (where the idea typically is to blend deep learning with interpretable model components). While strong interpretability by design may not be feasible, the moderate counterpart fares considerably better with regard to clinical purposes.

To underscore the perks and possible perils of interpretability by design, I draw on a recent study by Alina Jade Barnett et al. (2021), using an interpretable neural network that incorporates case-based reasoning strategies of radiologists to examine mammographic images. This example showcases the core assumptions of interpretable models and facilitates delineating them from XAI methods.

The paper proceeds as follows: Section 2 establishes the opacity problem in medical ML. Here, I rely on a pragmatist strategy, analyzing how the achievement of certain purposes is negatively impacted by the involvement of opaque ML models (Parker 2020). Section 3 discusses XAI (with an emphasis on saliency maps) and reliabilism as opacity mitigation strategies within the context of clinical medicine, with the aim of contrasting them with the interpretability approach. One particular problem for reliabilism is that there are stark discontinuities regarding the transferability of claims about ML models' performance, when compared to causal claims about the effects of medical drugs, established via randomized controlled trials (RCTs). Finally, section 4 addresses interpretability by design. The paper then concludes by defining desiderata for interpretability by design, ensuring that ML models can be implemented in clinical environments in a meaningful way.

2. Opacity in Medical Machine Learning

This section's objective is to establish conceptual common ground by mapping out some key assumptions in the philosophy of ML concerning the opacity problem. The recent advances in medical applications of ML can be characterized as by-products of more general breakthroughs in deep learning—relating to models composed of multiple layers of artificial neurons—which have proved to be particularly powerful at computer vision tasks (LeCun, Bengio, and Hinton 2015; Buckner 2019). ML researchers and philosophers alike widely agree that deep learning models are opaque. However, opacity is not a monolithic concept; different types and causes can be distinguished. Kathleen A. Creel (2020) provides a useful taxonomy, distinguishing between *functional opacity* (not knowing how the model functions as a whole), *structural opacity* (not knowing how the model is realized in code), and *run opacity* (not knowing how single instances were realized).[1] This model-centric perspective on opacity can be expanded by accounting for its political (corporate concealment) or user-centric (data illiteracy) dimensions (Burrell 2016). For the purposes of this paper, *functional* and *run opacity* are particularly relevant—in conjunction with user-centric aspects.

Concerning the causes of opacity, the common denominator is model complexity—defined by the size of parameters. Since deep learning models consist of billions of parameters, it is difficult to infer how and why they arrive at a given prediction/decision (Lipton 2018; Zednik 2021). Besides sheer complexity, other causes are the lack of domain knowledge explicitly encoded into the model, and the nonlinear structure of many models (Rudin 2019).

Furthermore, deep learning models themselves are not theoretically well understood, in particular with regard to how they achieve their generalization properties.[2] On a *behavioral* level, deep learning models have achieved high predictive performance by exploiting various heuristics. As a case in point, they have been shown to be biased toward image textures (for example, the lighting conditions of a medical image) over shape (for example, the structure of a tumor) in image-recognition tasks. This is in stark contrast to human behavior (Geirhos et al. 2018). Taken together, these factors make it challenging to understand model behavior and to provide reliable performance guarantees. As discussed below, the reliance on heuristics, in particular, poses a threat for the reliability of ML models in clinical environments.

2.1 Specifying the Opacity Problem(s) in Medical Machine Learning

ML models are by no means the only thing that is opaque in clinical medicine. The physiology of the human body, the causes of diseases, and the functioning of drugs are often not well understood. Similarly, very few people—and certainly not most clinicians—are able to explain the functioning of medical imaging technologies (such as ultrasound or computed tomography). And yet, despite all skeptical concerns (Ioannidis 2005; Stegenga 2018),

[1] I consider the distinction between functional and run opacity to be tantamount to the global/local distinction in ML research.

[2] Most profoundly, they seem to defy the core assumption of statistical learning theory, assuming that there is a trade-off between the complexity of a model and its performance on data from outside its training sample. The crux of the matter is that the performance of deep learning models seems to continue to improve, no matter how *deep* we allow the model to become, even as the neural nets effectively memorize the training sample. So far, there is no widely accepted mathematical explanation for this phenomenon (see Zhang et al. 2021 for a review).

clinical medicine seems to work reasonably well. So why is model opacity considered problematic, whereas other kinds of opacity are tolerable?

To tackle this question, my approach falls squarely into the *pragmatis*t camp, which can best be characterized by an instrumental stance regarding model opacity (Lipton 2018; Krishnan 2020; Zednik 2021; Nyrup and Robinson 2022). According to this view, opacity is problematic when it obstructs the achievement of certain purposes or goals. A prerequisite for overcoming opacity is therefore to specify what these purposes are, and which stakeholders are involved—in order to get a better grip on the success conditions. Pragmatism is particularly well suited, given that clinical medicine is ultimately driven by instrumental considerations (improving patients' health conditions). Thus understood, there are also parallels with the adequacy-for-purpose view in the modeling literature in philosophy of science (Parker 2020).

My assertion is that, within the context of clinical medicine, there are two key purposes, whose achievement is potentially impacted by opaque ML models: ensuring *generalizability* and *optimizing clinician–machine decision-making*.

2.1.1 Generalizabilty

The goal in training ML models is to find a (parameterized) decision-function, minimizing error when predicting a dependent variable Y over independent variables X within a joint probability distribution D. Their predictive performance is evaluated by showing the model a test set of unseen data, using various classification metrics. However, while many ML models surpass the diagnostic accuracy of expert clinicians under training conditions, distribution shifts are a key concern regarding the transferability of an ML model's performance when implemented into clinical environments. If a hospital labels data differently, applies the ML model to different patient demographics, or uses cameras to make medical images other than those the model was trained with, the result can be malfunctions by the model (see Finlayson et al. 2021 for a review). And while distribution shifts threaten the performance of any statistical model, deep learning models are particularly vulnerable to them in light of the myriad input variables.

While distribution shifts are a general source of performative instability in ML models, a more particular concern is that the ML model's decision-function overfits to the idiosyncrasies of the training data. In turn, it misses relevant variables, hindering the generalizability of the predicted output beyond training conditions. More specifically, deep learning models are prone to achieving high predictive performance by exploiting shortsighted learning strategies (Lapushkin et al. 2019; Geirhos et al. 2020). To give a drastic example, a recent study by Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee (2021), assessing deep learning models used to predict Covid-19 from chest radiographs, found that—despite achieving high diagnostic accuracy—many of the models relied on confounders (for example, textual markers, image edges, or patients' shoulder position), rather than medical pathology. If gone unnoticed, failures in the predictive performance, induced by distribution shifts, are likely to result in incorrect diagnoses and treatment choices. It is therefore crucial to detect potential confounders in the model as a safeguard for ensuring the generalizability of ML models.

Generalizability issues are affected by (functional) opacity in that it impacts model control in the auditing process. However, strictly speaking, transparency is neither necessary nor sufficient to achieve good generalization properties. For example, the model could be trained with even larger and more diverse data, so that it virtually does not have to leave training conditions. Alternatively, we could augment the training data such that we remove all confounding information. However, neither of these strategies is likely to succeed in isolation. First, unlike ML models that are supposed to discern, say, cats from dogs, one usually works with smaller datasets in medicine. This is especially since sharing medical data is impeded by privacy restrictions. Second, there are typically too many variables that can be exploited as confounders to remove all of them during the data processing.

2.1.2 Optimizing Clinician–Machine Decision-Making

ML models are tools, used for specific purposes. In clinical environments, the purpose is typically to provide decision support to clinicians. The model will present its predicted output to the clinician (for example, by way of a binary classification, a probability score, or a posterior distribution), with the clinician making the final diagnosis, and then proceeding to determine the appropriate treatment—along with the patient.

The decision support can come in different forms, distinguishable based on the epistemic relationship between the ML model and the clinician. In a *symmetric* epistemic relationship, the ML model acts as a peer to the clinician, and both are roughly equal in terms of diagnostic accuracy. A pertinent example here is a study by Scott Mayer McKinney et al. (2020), evaluating the accuracy and efficiency of an ML model, used as a second reader in detecting breast cancer from mammograms. By contrast, *asymmetric* epistemic relationships refer to cases where the predictive performance of the ML model is vastly superior to the clinician. Examples are ML models used to predict kidney failure based on patient health records (Tomašev et al. 2019), or the progression of macular degeneration (Yim et al. 2020). This asymmetry is closely tied to the extensive number of relevant predictors, making it challenging for humans to compute reliable estimates.

Opacity in ML models poses several obstacles to clinician–machine decision-making. First, without knowing whether the predictors used by the ML model are coherent with medical knowledge, it is unclear when decisions made by the ML model should be overridden. A large body of empirical studies examines how clinicians are led astray by faulty algorithmic diagnoses (for example, Tschandl et al. 2020; Gaube et al. 2021). This can be particularly problematic in edge cases or rare diseases, in which disagreements between the ML model and the clinician can be difficult to resolve. Moreover, even single mistakes—induced by the ML model—can have substantial consequences for patients' health. Opacity can be morally problematic in cases where a clinician violates due diligence, making treatment decisions based on an ML model's prediction, while being in the dark about the underlying factors. Finally, junior clinicians have been shown to defer to the model output, and opacity can therefore be detrimental for pedagogical purposes, since it hinders advancement to the expert level (Genin and Grote 2021; Grote and Berens 2022). When assessing adequate opacity amelioration strategies, it is important to keep in mind pragmatic factors—such as time constraints and whether a given strategy increases or reduces the clinician's mental load (see also Ehrmann et al. 2022).

While I deem *generalizability* and *optimizing clinician–machine decision-making* to be the key purposes potentially undermined by opacity in ML models, this list is not exhaustive. As Boris Babic et al. (2021) argue, a prerequisite for procedurally fair triage algorithms is to ensure that the ML model registers clinically meaningful variables. Furthermore, it is, of course, important to include the patient's perspective. Especially here, the threat is that patients are unable to engage in shared decision-making concerning treatment decisions, which is critical with regard to the different values entailed in treatment choices—hence the view that the involvement of the ML model reintroduces paternalism in the clinician–patient relationship (McDougall 2019; Bjerring and Busch 2021). Contrary to what has been suggested, I do not think that enabling patients to make informed decisions is best achieved by means of techniques that generate genuine explanations of the model output, tailored to patients (cf. Watson et al. 2019). Rather, higher-order information regarding the functionality of the model should be either conveyed via informed consent forms (see Kiener 2021), or communicated by the clinician, who becomes a mediator between the ML model and the patient. Having to interpret medical information—typically conveyed through visualizations or statistical summaries—puts too much of a burden on patients. Lastly, since the focus here is on clinical applicability, I do not consider how opaque ML models inhibit the process of scientific inquiry (Zednik and Boelsen 2022). Building on this specification of purposes, potentially undermined by opacity in ML models, I now turn to evaluating different opacity amelioration strategies.

3. Overcoming Opacity in Medical Machine Learning

This section discusses two opacity amelioration strategies for the application of ML models in clinical medicine: XAI as the standard paradigm and reliabilism as its fiercest competitor currently. My aim is twofold: First, I evaluate the two strategies in terms of how they enable the reliable achievement of *generalizability* and *optimizing clinician–machine decision-making*. Second, both strategies will be used as contrast classes to the interpretable modeling approach in the subsequent section.

3.1 Explainable Artificial Intelligence

So far, I have argued that, for the reasons noted above, model performance in medical ML cannot be assumed to be generalizable, yet opacity limits assessing generalizability. A popular strategy to overcome opacity is to use various XAI methods. I use XAI as a bundle term, denoting different methods that all seek to overcome opacity in ML models by explaining the predictors for single instances post hoc.[3] Among the most common methods are *Local Interpretable Model-Agnostic Explanations* (LIME) (Ribeiro, Singh, and Guestrin 2016) and *Shapley Additive Explanations* (SHAP) (Lundberg and Lee 2017). Even though the explanations are generated by distinct mathematical techniques (for an accessible overview, see Molnar 2020), both methods have in common that they provide statistical summaries of the key variables in a given prediction. However, their visualization of the explanation may differ. To give a toy example: if a given ML model predicts that a patient in the intensive care unit is at high risk of suffering from kidney failure, the corresponding explanation from LIME might look as follows—25% advanced age, 20% high

[3] Note that this definition of XAI may seem narrow, compared to wider definitions provided in many review papers in computer science. However, my aim in this section is not to provide an extensive taxonomy of XAI approaches but to capture the dialectic of different strategies that all seek to overcome opacity in medical ML.

blood pressure, 15% high creatinine levels, and so on. SHAP, in turn, will provide plots that explain the value/importance of key variables for the predicted output. In contrast, counterfactual explanations try to examine how small changes to the input variables (say, a higher creatinine level by *X*) affect the predicted output (risk of kidney failure). By detecting counterfactual dependencies, the goal is to gain a better understanding of why the model arrives at its predictions (Wachter, Mittelstadt, and Russell 2018).

However, rather than providing a survey of different XAI methods (for reviews, see Gilpin et al. 2018; Arrieta et al. 2020; Molnar 2020), my emphasis is on saliency maps, a popular technique for vision-based medical diagnosis. Indeed, since many medical ML applications revolve around image recognition, one will encounter few research papers in which saliency maps have not been used at some point.[4]

Saliency maps highlight regions in images, based on which deep learning models make their predictions. More precisely, they compute saliency scores for input pixels according to their network outputs (Montavon, Samek, and Müller 2018).[5] As an illustration, if an ML model detects a given eye disease from a medical image, the relevant explanation could take the form of a heatmap, either marking the salient regions in red, or grading the different regions in darker or brighter colors—depending on how relevant they are for the predicted output.

For medical purposes, the appeal of saliency maps is tied to their ability to explain the model behavior in a way that accounts for the background knowledge of medical professionals. Medical diagnosis, in particular, is grounded in perceptual skills that clinicians have acquired over the course of many years of practice, resulting in non-inferential reasoning processes and a distinct visual phenomenology (Stokes 2021).[6] This makes saliency maps a particularly useful heuristic for ensuring that an ML model's diagnostic/prognostic criteria are well aligned with medical background knowledge.

That said, saliency maps have many caveats. These become apparent when considering to what extent they facilitate overcoming the opacity-induced problems of evaluating *generalizability* and *optimizing clinician–machine decision-making*. Let us consider each in turn.

A prerequisite for evaluating *generalizability* is that ML developers have knowledge of the model's functioning of the whole: which variables are used as predictors, how the different variables interact, whether the model is well structured, and so on. And even then, performative instability due to distribution shifts can be challenging to detect, since the assessment of the model needs to account for information external to the model (for example, how patient demographics might have changed over time). While saliency maps can be very useful in spotting glaring cases of shortcut learning (say, the model uses image markers as predictors for Covid-19), they will not be of much help in achieving a holistic understanding of the model's functioning. This is because the visualization of areas of interest is typically coarse-grained, and is confined to explaining the predictors for single instances. If anything, functional opacity, as opposed to run opacity, could be mitigated to some extent by clustering the output of the saliency map across *n* instances. With regard to

[4] However, it has been shown that XAI techniques such as LIME, SHAP, and counterfactual explanations suffer from reliability issues, like those I will discuss with regard to saliency maps (Slack et al. 2020, 2021a).

[5] For an accessible discussion of technicalities in saliency maps, see Zednik (2021).

[6] See also Grote and Berens (2022) and Nyrup and Robinson (2022) on the importance of accounting for clinicians' tacit knowledge for successful algorithmic explanations.

overcoming functional opacity, the role of saliency maps can be best described as a heuristic, indicating which areas/variables play an important role for the model output.

A study by Nenad Tomašev et al. (2019), training a deep learning model to predict acute kidney injury (AKI) in patients within the next 48 hours, is indicative about the possibilities and limitations in obtaining knowledge of the model's functioning of the whole, especially since a nuanced description of the evaluation process is provided by the researchers. They try to ensure the clinical confidence of the ML model by assessing its predictive performance for auxiliary targets (for example, known biomarkers for AKI) and by setting the weights of individual variables to zero—to see how it affects the predicted output (see also Tomašev et al. 2021). These techniques facilitate capturing counterfactual dependencies and thus can be deemed a useful sanity check. However, considering the myriad different variables used as input data, it is safe to say that this does not translate into guarantees for the model's performative stability. The study also illustrates a more conceptual problem about ameliorating functional opacity. Medical knowledge is typically incomplete. For many diseases, no known biomarkers exist and, given the complexity of physiological processes, it can be difficult to assess whether the relationship between variables is causally relevant or spurious.

A different concern arises about the faithfulness of the explanation provided by the saliency map to the original model.[7] Julius Adebayo et al. (2018) conduct a perturbation analysis on popular saliency map techniques. They compare the output of a saliency map on a trained model with the output of the saliency map on a randomly initialized untrained network. They also compare the output of the saliency map for two identical model architectures, where the data labels have been randomly permuted for one model. The upshot is that often the outputs of the saliency maps have been shown to be invariant to changes of the model architecture/data. One interpretation of this finding is that the saliency maps use shortcut learning strategies, like deep learning models—whose opacity they are supposed to overcome. The study has been replicated regarding ML models for medical diagnosis (Arun et al. 2021).

Turning to *optimizing clinician–machine decision-making,* a study by Philipp Tschandl et al. (2020) shows that the involvement of saliency maps could play an important pedagogical function insofar as medical students learned to direct their attention to cues that they otherwise would have missed. However, as Tschandl et al. point out, it is still an open question whether algorithmic explanations indeed lead to a decrease in diagnostic errors (2020, 1232). To the best of my knowledge, no study that tests the benefits of XAI methods for medical professionals in realistic clinical settings exists yet.

At a more conceptual level, saliency maps have two important shortcomings: First, they tell clinicians where to look but not what to see. Thus, when the interpretation of certain areas of interest is ambiguous, they will not resolve the relevant uncertainty. Second, there may be a misalignment between the predictors used by the ML model and the diagnostic criteria/regions of interest that guide clinicians' decision-making. In this case, algorithmic explanations might even drive uncertainty in the clinician: it can be difficult to determine whether the salient regions are mere artifacts, or whether the ML model has discovered meaningful patterns, which have so far eluded humans.[8]

[7] For an interesting proposal of how to establish faithfulness in XAI methods, see Watson (2022).

[8] See Buckner (2021) on the view that deep learning models might perceive objects differently than humans.

While some of these challenges can be met by clinically validating saliency maps (Ayhan et al. 2022), or by using XAI methods that report their associated uncertainty (Slack et al. 2021b)—for the purpose of establishing their faithfulness—it becomes evident that saliency maps in isolation cannot do the heavy lifting required to overcome opacity in medical ML. Granted, an ensemble of different XAI techniques might fare better here. Even so, as has been discussed, the incompleteness of medical knowledge poses a barrier to a holistic internal evaluation of the ML model. As for *optimizing clinician–machine decision-making,* the crucial question is whether clinicians would be able to process and amalgamate the information provided by different kinds of explanations. I tend to be skeptical here, considering that clinicians' decision-making has often been described as boundedly rational (Mullainathan and Obermeyer 2022). Closely related, research in human–computer interactions suggests that people find it difficult to detect glitches even in quite simple interpretable models as a result of information overload (Poursabzi-Sangdeh et al. 2021). Indeed, putting too much emphasis on XAI techniques may increase clinicians' mental load, which then increases their susceptibility to errors.

The upshot is that XAI techniques such as saliency maps are promising tools for ameliorating opacity in medical ML. However, there are structural problems that make it unwise to be overly reliant on them. In the next section, I discuss the opposing strategy to XAI.

3.2 Reliabilism

While the general strategy of XAI is to overcome the opacity problem by understanding how and why an ML model arrives at its predictions—thereby making potential errors *in* the ML model salient—the basic idea of reliabilism is to establish guardrails, minimizing the likelihood of errors occurring in the first place. The crux is that these guardrails are *external* to the model. Reliabilism is an increasingly influential view in the philosophical debate concerned with the implementation of ML in clinical settings, first articulated by Alex John London (2019) and more recently by Juan Manuel Durán and Karin Rolanda Jongsma (2021).

London (2019), in particular, emphasizes the parallels between deep learning and medical science. His main argument can be summarized as follows: Both disciplines are able to develop powerful tools/interventions. And yet, knowledge of the causal structure is lagging behind. Because of the incompleteness of knowledge, justifications of model output by way of causal explanations are often misleading—especially since ML models learn by exploiting associations, instead of discovering causal relationships. Hence, to establish the clinical benefit of ML models, the proposed solution is to evaluate them according to the same standards as other medical interventions—preferably by way of RCTs (2019, 17). Consequently, neither functional opacity, nor run opacity are deemed problematic, if the ML model has been properly validated.

To fully apprehend London's view, we need to consider it against the backdrop of medical epistemology. Here, evidence-based medicine (EBM) has emerged as the leading epistemological paradigm in clinical medicine. It seeks to provide guidance on how clinicians can sort out reliable evidence from unreliable evidence, in order for them to conscientiously decide on the care of individual patients, according to the best evidence available (Sackett and Rosenberg 1995). The gist is that EBM has a very restrictive view on what counts as good evidence; namely, different types of clinical studies, structured hierarchically, based on their evidentiary status. Accordingly, expert opinions rank at the bottom, followed by case-control studies, cohort studies, and finally RCTs—considered the gold standard.[9]

The hierarchy is justified by the fact that the distinct types of studies are less susceptible to researcher biases threatening their internal validity. The reason that RCTs rank at the top is, given certain conditions, they are assumed able to generate statistically unbiased estimates of the average effect of the treatment under study. This is as a result of random assignment facilitating unbiased estimation of average treatment effects by rendering baseline prognostic factors (such as age, race, disease severity, and so on) statistically independent of assignment to treatment. The assumption is that RCTs are the best method for establishing causal claims of treatment effects (Senn 2013; Fuller 2021; see also Genin and Grote 2021).

The evidential hierarchy of EBM entails that those other kinds of evidence are devalued. Aside from expert opinions, this particularly relates to laboratory studies trying to identify physiological mechanisms. These are only considered relevant insofar as they facilitate interpreting clinical observations (Horwick 2011, 21). [10] The skepticism concerning biological, anatomical, or chemical knowledge is tied to the fact that they provide only indirect evidence that a given treatment works (Broadbent 2019, 138–139).

London's (2019) (implicit) presumption is therefore that it is misguided to impose a double standard (Zerilli et al. 2019) to the evaluation of ML models, when compared to traditional medical drugs or devices. The *double standard* argument—with respect to the evaluation of ML models versus humans—has been criticized on the grounds that malfunctions in technical systems result from poor design choices made by humans. To counteract these malfunctions, it needs to be possible to scrutinize the relevant technicalities (Günther and Kasirzadeh 2022).

There are good reasons for taking the double standard argument in the debate about opacity in medical ML seriously, not least because it is a widely held view among medical professionals that ML models should be evaluated like other medical devices. As a case in point, the US Food and Drug Administration (FDA) stipulates that ML models be classified as medical devices in a similar category as ultrasound or X-ray (FDA 2022). Consequently, they must go through the same evaluation process.

Before addressing the limitations of this view, consider another reliabilist position, advocated by Durán and Jongsma (2021). Rather than focusing on medical epistemology, they draw on the literature on the reliability of computer simulations. Building on this, they contend that their account of *computational reliabilism* offers the right epistemic tools to justify a clinicians' belief in an ML model. More precisely, their epistemic conditions are: (1) verification/validation; (2) robustness analysis; (3) a history of (un)successful implementations; and (4) expert knowledge. They also suggest that transparency could act as another epistemic safeguard, even though transparency in itself may not be sufficient to justify trust in ML models (2021, 332). Hence, whereas London (2019) could be considered an *outcome reliabilist,* Durán and Jongsma (2021) are closer to *process reliabilists* (see also Goldman and Beddor 2021).

[9] However, many EBM pyramids will place meta-analyses of RCTs at the top.

[10] However, mechanistic evidence is admitted in cases in which there is no clinical evidence available (Horwick 2011, 21).

Reliabilist positions have many merits. In particular, London (2019) should be commended for pointing out that any evaluation of medical ML models needs to consider clinical endpoints (say quality-adjusted life years) and not merely the predictive performance during training: for various reasons, an increase in diagnostic accuracy may not translate into improved patient outcomes (Genin and Grote 2021). By contrast, one of the merits of Durán and Jongsma (2021) is that they highlight the importance of proper validation and robustness checks for overcoming opacity in ML models.[11] Moreover, on a charitable reading, their conditions (3) and (4) allow us to account for a wider range of socio-technical issues about the implementation of ML models into clinical settings.

With that in mind, both accounts face significant constraints. An issue for London's (2019) view is that there are differences between the transferability of drug RCTs and RCTs involving ML models. Setting statistical issues of extrapolation aside (see also Fuller 2021)—and if members of all relevant groups are well represented among the research participants[12]—effect sizes from drug RCTs are often claimed to be transferrable to nonexperimental settings (see also Post, De Beer, and Guyatt 2013).[13] The same does not apply to RCTs for ML models, given that distribution shifts can severely impact the model's performance, even if it is used in another hospital in the same city. Along these lines, Emma Beede et al. (2020) point out how poor lighting in a hospital in Thailand led to many ungradable images for an ML model, used to detect diabetic retinopathy, while the same ML model surpassed expert ophthalmologists during training.

And while basic research in computer vision suggests that the robustness gap between ML models and human perception is narrowing, this can be mainly attributed to increases in the size of the training datasets (Geirhos et al. 2021). For medical purposes, however, collecting sufficiently large datasets can be exceedingly difficult (as explained in section 2.1). Another weak point in London's (2019) account is that it ignores epistemic and moral issues arising from the interplay of clinicians and ML models. It creates an epistemic environment in which the role of the clinician is basically to be deferential, which jeopardizes the prospects of shared decision-making (again, see section 2.1). Finally, like the problem of deriving optimal treatment for individual patients from average treatment sizes—which is what RCTs capture—the overall predictive performance of an ML model may not clinch its reliability in individual instances.

When judged against the criteria of evaluating *generalizability* and *optimizing clinician–machine decision-making,* the results for computational reliabilism are mixed. On the one hand, conditions (1)–(4) in Durán and Jongsma's' account capture what is at stake. On the other hand, as it stands, computational reliabilism is too vague to be properly informative. The individual epistemic conditions have not been carved out for medical ML models. Moreover, the formal definition of computational reliability, according to which a reliable ML model should have a higher probability of being correct than an unreliable one (Durán and Jongsma 2021, 332), provides a weak justificatory basis: if the predecessor model has a low predictive performance, even a slightly better model would count as reliable. To advance the research program of computational reliabilism, it needs to be

[11] Note that London also underscores the need for empirical testing and regulatory procedures for the aims of ensuring the responsible use of ML models (2019, 20).

[12] This is, of course, a common concern regarding clinical trials.

[13] Although the extent to which causal claims can be extrapolated from RCTs is contested in the philosophy of science literature (Deaton and Cartwright 2018).

spelled out how the individual epistemic conditions should be operationalized by way of validation procedures and what the baselines in terms of accuracy and robustness ought to be for a medical ML model to count as trustworthy. Some progress in this direction has recently been made by Koray Karaca (2021) and Emanuele Ratti and Mark Graves (2022), providing nuanced accounts of epistemic risks across the pipeline of data acquisition, model training, model evaluation, and impact assessment.

Outside of philosophy, the algorithmic auditing literature can be seen as a continuation of the reliabilist project in the spirit of Durán and Jongsma (2021). Xiaoxuan Liu et al. (2022) suggest various approaches for identifying and documenting performance errors in ML models in externally valid settings—including error analysis for patients from underserved subgroups. In that respect, they provide useful criteria for conditions (1) and (2) of computational reliabilism. However, while they should be commended for their detailed analysis of sources of errors and methodological recommendations, their account is confined to the model performance. Thus, it ignores errors that may occur at the level of *clinician–machine decision-making*. Likewise, it is unclear which assurances can realistically be gained by ML models in the evaluation process in light of the plethora of causes that potentially induce performance failures.

To sum up, London's (2019) reliabilist account makes an important contribution by exploring how opacity in ML models can be ameliorated by well-established evaluation processes in clinical medicine. However, my point of contention is that this account does not capture discontinuities between the functionality of ML models and drugs, while also ignoring the specific obstacles that opacity poses for *optimizing clinician–machine decision-making*. Turning to computational reliabilism, the main flaw is that its conditions are underspecified. In that sense, it can be considered a promising research program, rather than a full-fledged theory.

4. Overcoming Opacity through Interpretable Models

So far, I have discussed two strategies for overcoming opacity in medical ML. Both have their merits as well as their shortcomings. This section deals with interpretability by design as an emerging opacity mitigation strategy. To evaluate this approach in a meaningful way, I first lay down its conceptual assumptions and then discuss a case study, presenting an interesting edge case of interpretable models used for clinical purposes.

4.1 Scrutinizing Model Interpretability

Interpretability by design has been popularized by Cynthia Rudin (2019), vigorously arguing for the use of interpretable models over deep learning models for consequential decision-making—even when the latter are supplemented with XAI methods.[14] While Rudin is interested in a broad range of ML applications, including criminal justice and public policy, her main arguments have recently been echoed, with an emphasis on clinical medicine, by Babic et al., who contend that healthcare professionals should be wary of

[14] Note that in addition to epistemic aspects, Rudin (2019) is also concerned with the political dimensions of opacity in algorithmically assisted consequential decision-making. In this paper, however, I confine myself to the epistemic side.

explanations provided via XAI methods and rather focus on the effectiveness and safety of the models (2021, 286).

Rudin's (2019) critical stance toward deep learning models is related to two factors. First, there is a general skepticism of XAI methods, which are often unfaithful to the original model and may not provide a nuanced understanding of the model's functioning of the whole—or even mislead clinicians. These issues have already been addressed in previous sections. Second, Rudin denies that the predictive performance of deep learning models is necessarily more powerful than that of interpretable models. This claim might seem surprising, given that basically all the recent breakthroughs in the medical domain are related to advances in deep learning. In this respect, clinical medicine proves an interesting touchstone for interpretability by design. I will address this issue at a later point. For now, let us turn to the question of what it is that makes ML models interpretable.

Rudin asserts that interpretability is a domain-specific notion, which is why there cannot be a satisfactory all-purpose definition (2019; see also Rudin et al. 2022). However, a common denominator in interpretable models is that they are constrained in some way—for example, by way of monotonicity, additivity, causality, sparsity, or by incorporating domain knowledge (Rudin 2019, 206). Moreover, there is a close link between interpretability and humans' psychological capacities. Provided that a human puts in reasonable effort and possesses the necessary (statistical or medical) background knowledge, she should be in an epistemic position to grasp how the inputs relate to the predicted output in an interpretable model (Babic et al. 2021, 284; see also Erasmus, Brunet, and Fisher 2021).

However, instead of seeking a formal definition, it might be heuristically useful to look at some examples. Paradigm examples for interpretable models are logistic regression models or decision trees.[15] One particular reason why decision trees are considered interpretable is that their structure makes the interaction between the individual variables intelligible. With that in mind, depending on the number of individual variables/nodes, knowledge of the model's holistic functioning can still be hindered (see also Lipton 2018).

This raises some general questions regarding interpretability by design: if knowledge of an interpretable model's functioning can be undermined, given certain empirical conditions, is the distinction between interpretable models and opaque models *gradual* or *categorical*? As will be shown later in this section, the fact that many interpretable modeling approaches within the context of clinical medicine use deep learning models, in conjunction with more interpretable elements, supports the latter assumption. In that respect, a distinction must be made between *strong* and *moderate* model interpretability, where the former refers to, say, regression models or decision trees, and the latter refers to hybrid modeling techniques.

Moreover, if *interpretability* is a domain-specific notion, what are the relevant success conditions for an ML model to count as interpretable? Again, my conjecture is that this question is best dealt with through the lense of a pragmatist approach, analyzing to what extent model interpretability facilitates achieving certain purposes of interest.

Consider *generalizability*. Contrary to XAI or reliabilism, in which the relevant concern is about assessing generalizability, the question initially in the foreground here is whether

[15] A decision tree predicts a target variable *Y* by traveling from a root node to a leaf, with each leaf representing a given variable. Here, the input space will be split at each node on the root-to-leaf path, based on a predefined set of rules (Shalev-Shwartz and Ben-David 2014, 250).

interpretable models can indeed achieve a predictive performance that is roughly as good as state-of-the-art deep learning models. After all, the appeal of deep learning models is that they excel at learning from high-dimensional data, allowing them to greatly surpass other existing ML model architectures at computer vision and natural language-processing benchmark tasks. In the same vein, the breakthrough studies demonstrating that ML models can achieve accuracy in image-based diagnosis at the level of medical experts were all based on deep learning models (Gulshan et al. 2016; Esteva et al. 2017; De Fauw et al. 2018). Thus, when implementing a decision tree, rather than a deep learning model, into clinical settings, there is arguably a risk of trading off accuracy against interpretability.

There are various responses to the accuracy–interpretability trade-off. First, as argued by Rudin (2019), the performance gap between interpretable ML models and deep learning models can be mitigated by way of more considerate modeling choices—for example, ensuring that the model registers the right variables and using higher-quality data. Second, one needs to distinguish between the *discovery* process and the *implementation* process. When ML models are used for scientific discovery, the promise is that they discover novel structures. This can guide the generation of novel hypotheses about meaningful statistical relationships (see also Zednik and Boelsen 2022). However, once some hypotheses have been refuted, we can constrain the solution space. Hence, although ML models initially achieve their highest predictive performance by being trained with the largest possible sets of high-dimensional data, some of the excessive complexity can be reduced afterward.

Along these lines, Geoffrey E. Hinton, Oriol Vinyals, and Jeff Dean (2015) discuss how the knowledge of an ensemble of deep learning models can be transferred into a surrogate model (a simple neural network) by distilling their predictors for the different output layers. Wieland Brendel and Matthias Bethge (2019) develop a model architecture, called *BagNets,* blending features of deep learning models with bag-of-features models.[16] When evaluated in ImageNet (arguably the most prominent benchmark in computer vision), the BagNet model's performance came close to state-of-the-art deep learning models, while the linear structure of the classifier enables understanding how small image patches are integrated for arriving at the predicted output. This makes it easier to detect error cases and to get statistical guarantees in light of performance failures caused by distribution shifts.

Moreover, for BagNets, it is possible to straightforwardly compute saliency maps that are faithful to the model. This is particularly useful for the purpose of *optimizing clinician–machine decision-making*. Since BagNets' predictive performance comes close to deep learning models, while being more interpretable by design, it is claimed to be an interesting compromise for clinical tasks (Ilanchezian et al. 2021). However, it also shows that all these models fall into the category of moderate interpretability by design. Provided that a state-of-the-art ML model achieves an astonishing performance for a clinical task, regression models or decision trees are unlikely to be an equal substitute.

Yet another argument against the accuracy–interpretability trade-off is that many of the seemingly most powerful models achieve their predictive performance by overfitting to the benchmark data. For example, in a critical review of methodological problems in current medical imaging, Gaël Varoquaux and Veronika Cheplygina (2022) point out that particularly the top 10% of medical ML models in one benchmark show high evaluation noise when evaluated in a different benchmark. Thus, they conclude that attempts to

[16] The latter represent clusters of local image features in a vector, used as an input to a linear classifier.

maximize the performance of ML models can have diminishing returns (2022, 3). As an upshot, by constraining the model architecture, or by incorporating domain knowledge, we may end up with ML models whose performance may, prima facie, not be as impressive as that of their unconstrained counterparts on benchmark datasets. However, the interpretable model promises more stable performance across different settings and may lead to better real-world outcomes.

Granted, the argument against the interpretability–accuracy trade-off is an empirical, rather than a principled one. As such, it can be refuted by new model architectures that surpass interpretable models in well-defined benchmark tasks. Provided that interpretable model architectures are unable to perform on par with their state-of-the-art deep learning counterparts, we need to weigh carefully what the epistemic risks are when we trade off accuracy for interpretability. In the end, the balancing may even boil down to different value judgments: should we give preference to technologies in medical environments that perform miraculously well under certain conditions but can go horribly wrong, or should we prioritize ML models where we know *reasonably well* how they function and that their function is stable across different settings? In light of the high-risk setting of clinical medicine, where incorrect diagnoses and prognoses can culminate in suboptimal treatment choices, my preference is for the latter.

4.2 Case Study: Using Interpretable Deep Learning Models in Radiology

In order to better understand the distinct properties of interpretability by design, I want to discuss a recent study by Barnett et al. (2021) on an ML model used to predict malign or benign lesions from mammographic examination. What is particularly interesting about this study is that the researchers use a deep learning model, made interpretable by incorporating expert knowledge. Consequently, the study is also valuable to understanding the boundaries of the interpretable modeling approach.

Mammographic examination is an especially challenging image-recognition task. The evidence in the images is often ambiguous, there are few public datasets available (making external validation of models challenging), and the stakes are extremely high, since it guides the decision on whether to order a biopsy. To establish clinical confidence in the ML model, the objective of Barnett et al. (2021) is to develop an ML model whose reasoning process approximates actual radiologists. This is claimed to be beneficial regarding the *generalizability* of the model beyond training conditions (since it underpins that the model looks at clinically relevant aspects of an image) and *optimizing clinician–machine decision-making* (since it enables detecting incorrect predictions).

Barnett et al. (2021) rely on a case-based reasoning approach to ensure model interpretability, where the model learns a *prototype* of a given disease that it compares unseen instances to. This prototype is implemented by way of an attention mechanism in the last layers of the deep learning model, penalizing the model if it uses confounding information. The prototype has been trained by a subset of pixel-level annotations made by expert radiologists, indicating the mass margin of lesions. The attention mechanism infers prototypical activation patterns from these images, which then constitute the ground truth for the model when examining novel images (Barnett et al. 2021, Appendix, Methods).

When using the attention mechanism to interpret a given diagnosis, it highlights the regions of interest in a mammographic image, while also estimating the probability of the diagnosis by comparing the regions of interest to prototypical images of mass margins that the model has seen before (Barnett et al. 2021, 1061). Borrowing terminology from Catherine Elgin (2007), one might say that the attention mechanism learns *exemplifications,* which are used both to estimate confidence in the model's diagnosis and to foster understanding for the clinician. The attention mechanism learns the epistemic properties that instantiate a given disease. If the symptoms in an image do not correspond to these properties, the attention mechanism will calculate the loss. This enables the attention map to quantify the model's uncertainty. The ML model's performance has been compared to radiologists who had to estimate the probability of a lesion being malignant. While the radiologists had an AUROC of 0.91,[17] the model was roughly seven points under the clinicians (Barnett et al. 2021, 1065). Thus, the model's performance comes close to the radiologists but does not outperform them.

4.3 Challenges for Interpretability by Design

From this point of view, case-based reasoning approaches present a clear advantage over saliency maps since they are more closely aligned to norms of clinical reasoning. Even so, while I find the case-based reasoning approach to model interpretability commendable, it does raise some pragmatic and conceptual concerns.

Beginning with the former, it is unclear how the performance of Barnett et al.'s (2021) model compares to *unconstrained* deep learning models, which is pivotal to understanding whether the increase in interpretability comes with any performance costs.[18] Furthermore, it remains to be assessed whether the case-based reasoning approach indeed culminates in better model control on the part of ML developers. The same goes for a reduction of diagnostic errors in clinician–machine decision-making—especially when compared to existing XAI methods. In addition, how computationally expensive is the training of the attention map once we train the model for a larger array of disease classes? These are all important desiderata that need to be empirically addressed to clinch the superiority of the approach.

More fundamentally, it needs to be asked which general insights can be extrapolated from Barnett et al.'s (2021) study for different ML solutions in medicine. For example, while it is plausible to assume that the approach works well for image-based diagnostic tasks, where the diagnostic criteria are usually well understood, it needs to be considered how

[17] AUROC (area under the receiver operating characteristic curve) is a graphical plot that shows the performance of a classification model at different classification thresholds. Usually, in binary classification problems, a decision threshold must be set. For each possible decision threshold, we can calculate the true positive rate (TPR) and the false positive rate (FPR) of the classifier. If we plot the respective TPR and FPR values, we obtain the receiver operating characteristic or ROC curve, which has a characteristic monotonic shape from the bottom left to the top right. We can now measure the area under this curve, which gives us a value between 0.5 and 1. A high value expresses that there is a threshold for which both a high TPR (also called sensitivity) and a high FPR (also called specificity) can be achieved, stating that our classifier is strong. A value close to 0.5, by contrast, expresses that our classifier is only as good/bad as a random classifier.

[18] Some studies report that their deep learning models perform better than average radiologists (see, for example, McKinney et al. 2020). However, in general, the validity of accuracy claims in ML-based breast cancer studies has been criticized on the grounds of high research bias concerning patient selection and unclear reference standards (for a critical meta-analysis, see Freeman et al. 2021).

applicable case-based reasoning can be in areas where the relevant medical knowledge is lagging behind.[19]

At a conceptual level, the distinction between models that are *inherently* interpretable versus those that are made interpretable/explainable post hoc is blurry. Put differently, when I train a deep learning model to predict a given disease, does it make any differences, in terms of robustness and faithfulness of the explanations provided, whether an attention map is *built into* the last layers of the model or is placed *on top*? This question also entails semantic issues. What exactly does it mean for an explanation to be post hoc? Likewise, what constitutes the boundaries of an ML model? To be sure, none of these concerns is a knockout argument. Rather, they underscore need for further conceptual and empirical grounding—for the purpose of establishing the clinical benefit of interpretability by design.

5. Conclusion

This paper has developed an account of the opacity problem in medical ML. Guided by pragmatist assumptions, I have argued that opacity in ML models is problematic insofar as it potentially undermines the achievement of two key purposes: evaluating *generalizability* and *optimizing clinician–machine decision-making*. Three opacity amelioration strategies were examined in turn, with XAI as the predominant approach, challenged by two revisionary strategies. Concerning XAI, it is apparent that the existing methods are unable to provide the necessary assurances to overcome opacity in medical ML models—at least when used in isolation. Reliabilist strategies, by contrast, highlight the importance of proper external validation of ML models by way of clinical trials and robustness checks. In doing so, they lay the groundwork for an evidentially diverse view on model evaluation (see also Williamson 2021). However, a prerequisite to meaningfully advance the research program of reliabilism is to be explicit about how the individual epistemic conditions can be operationalized. Relatedly, it needs to be clarified what acceptable performance thresholds are for an ML model to be considered trustworthy.

The strengths of interpretability by design, in turn, lie in its emphasis on using simpler model architectures and on ensuring that the model registers medical knowledge. These two factors in particular are why I deem the interpretable modeling approach most promising for tackling the issues of *generalizability* and *optimizing clinician–machine decision-making*. However, here too, there is a need for further empirical studies and conceptual grounding to clarify the distinct features of this approach and to demonstrate its clinical benefit. Looking beyond the individual opacity amelioration strategies, I hope this paper also contributes to a deeper understanding of the problem space and the solution space for overcoming opacity in medical ML.

Acknowledgments

This paper has greatly benefited from comments from the AITE (Artificial Intelligence, Trustworthiness and Explainability) research group. I am also grateful to Emily Sullivan and Timo Freiesleben for comments. I am thankful to the audience at the Center for Advanced Internet Studies (Bochum) for insightful questions. My research is supported by the Deutsche

[19] One touchstone here could be the prognosis of disease progression, often made by combining input data from different modalities (Yim et al. 2020).

Forschungsgemeinschaft (BE5601/4-1; Cluster of Excellence: Machine Learning—New Perspectives for Science, EXC 2064, project number 390727645), as well as the Baden-Württemberg Foundation (Verantwortliche Künstliche Intelligenz) as part of the project AITE. I also acknowledge support from Certification and Foundations of Safe Machine Learning Systems in Healthcare, funded by the Carl Zeiss Foundation.

Disclosure Statement

No competing interest was reported by the author.

References

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. "Sanity Checks for Saliency Maps." *Advances in Neural Information Processing Systems* 31, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. https://proceedings.neurips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Arun, Nishanth, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, et al. 2021."Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging." *Radiology: Artificial Intelligence* 3, no. 6. https://doi.org/10.1148/ryai.2021200267.

Ayhan, Murat Seçkin, Louis Benedikt Kümmerle, Laura Kühlewein, Werner Inhoffen, Gulnar Aliyeva, Focke Ziemssen, and Philipp Berens. 2022. "Clinical Validation of Saliency Maps for Understanding Deep Neural Networks in Ophthalmology." *Medical Image Analysis* 77, art. 102364. https://doi.org/10.1016/j.media.2022.102364.

Babic, Boris, Sara Gerke, Theodoros Evgeniou, and I. Glenn Cohen. 2021. "Beware Explanations from AI in Health Care." *Science* , no. 6552: 284–286. https://doi.org/10.1126/science.abg1834.

Barnett, Alina Jade, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, and Cynthia Rudin. 2021. "A Case-Based Interpretable Deep Learning Model for Classification of Mass Lesions in Digital Mammography." *Nature Machine Intelligence* 3, no. 12: 1061–1070. https://doi.org/10.1038/s42256-021-00423-x.

Beede, Emma, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. "A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy". In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–12. Association for Computing Machinery. https://doi.org/10.1145/3313831.3376718.

Bjerring, Jens Christian and Jacob Busch. 2021. "Artificial Intelligence and Patient-Centered Decision-Making." *Philosophy & Technology* 34, no. 2: 349–371. https://doi.org/10.1007/s13347-019-00391-6.

Brendel, Wieland and Matthias Bethge. 2019. "Approximating CNNs with Bag-of-Local-Features Models Works Surprisingly Well on ImageNet." *arXiv preprint,* 1904.00760. https://doi.org/10.48550/arXiv.1904.00760.

Broadbent, Alex. 2019. *Philosophy of Medicine*. Oxford: Oxford University Press.

Buckner, Cameron. 2019. "Deep Learning: A Philosophical Introduction." *Philosophy Compass* 14, no. 10, art. e12625. https://doi.org/10.1111/phc3.12625.

———. 2021. "Black Boxes, or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour." *British Journal for the Philosophy of Science*. Advance online publication. https://doi.org/10.1086/714960.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3, no. 1. https://doi.org/10.1177/2053951715622512.

Creel, Kathleen A. 2020. "Transparency in Complex Computational Systems." *Philosophy of Science* 87, no. 4: 568–589. https://doi.org/10.1086/709729.

Deaton, Angus and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210: 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005.

De Fauw, Jeffrey, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomašev, Sam Blackwell, Harry Askham, et al. 2018. "Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease." *Nature Medicine* 24, no. 9: 1342–1350. https://doi.org/10.1038/s41591-018-0107-6.

DeGrave, Alex J., Joseph D. Janizek, and Su-In Lee. 2021. "AI for Radiographic COVID-19 Detection Selects Shortcuts over Signal." *Nature Machine Intelligence* 3 no. 7: 610–619. https://doi.org/10.1038/s42256-021-00338-7.

Durán, Juan Manuel and Karin Rolanda Jongsma. 2021. "Who Is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI." *Journal of Medical Ethics* 47, no. 5: 329–335. https://doi.org/10.1136/medethics-2020-106820.

Ehrmann, Daniel E., Sara N. Gallant, Sujay Nagaraj, Sebastian D. Goodfellow, Danny Eytan, Anna Goldenberg, and Mjaye L. Mazwi. 2022. "Evaluating and Reducing Cognitive Load Should Be a Priority for Machine Learning in Healthcare." *Nature Medicine* 28: 1331–1333. https://doi.org/10.1038/s41591-022-01833-z.

Elgin, Catherine. 2007. "Understanding and the Facts." *Philosophical Studies* 132, no. 1: 33–42. https://doi.org/10.1007/s11098-006-9054-z.

Erasmus, Adrian, Tyler D.P. Brunet, and Eyal Fisher. 2021."What Is Interpretability?" *Philosophy & Technology* 34, no. 4: 833–862. https://doi.org/10.1007/s13347-020-00435-2.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542, no. 7639: 115–118. https://doi.org/10.1038/nature21056.

FDA (US Food and Drug Administration). 2022. "Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices." https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

Finlayson, Samuel G., Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. 2021. "The Clinician and Dataset Shift in Artificial Intelligence." *New England Journal of Medicine* 385, no. 3: 283–286. https://doi.org/10.1056/NEJMc2104626.

Freeman, Karoline, Julia Geppert, Chris Stinton, Daniel Todkill, Samantha Johnson, Aileen Clarke, and Sian Taylor-Phillips. 2021. “Use of Artificial Intelligence for Image Analysis in Breast Cancer Screening Programmes: Systematic Review of Test Accuracy.” *British Medical Journal* 374. https://doi.org/10.1136/bmj.n1872.

Fuller, Jonathan. 2021. “The myth and Fallacy of Simple Extrapolation in Medicine.” *Synthese* 198, no. 4: 2919–2939. https://doi.org/10.1007/s11229-019-02255-0.

Gaube, Susanne, Harini Suresh, Martina Raue, Alexander Merritt, Seth J. Berkowitz, Eva Lermer, Joseph F. Coughlin, et al. 2021. “Do as AI Say: Susceptibility in Deployment of Clinical Decision-Aids.” *Npj Digital Medicine* 4, no. 31. https://doi.org/10.1038/s41746-021-00385-9.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. “Shortcut Learning in Deep Neural Networks.” *Nature Machine Intelligence* 2, no. 11: 665–673. https://doi.org/10.1038/s42256-020-00257-z.

Geirhos, Robert, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2021. “Partial Success in Closing the Gap between Human and Machine Vision.” In *Advances in Neural Information Processing Systems* 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, 23885–23899. https://proceedings.neurips.cc/paper/2021/file/c8877cff22082a16395a57e97232bb6f-Paper.pdf.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2018. “ImageNet-Trained CNNs Are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness.” *arXiv* preprint:1811.12231. https://doi.org/10.48550/arXiv.1811.12231.

Genin, Konstantin. 2018. “The Topology of Statistical Inquiry.” PhD diss., Carnegie Mellon University. https://kgenin.github.io/papers/draft4.pdf.

Genin, Konstantin and Thomas Grote. 2021. “Randomized Controlled Trials in Medical AI: A Methodological Critique.” *Philosophy of Medicine* 2, no. 1: 1–15. https://doi.org/10.5195/pom.2021.27.

Gilpin, Leilani. H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lelana Kagal. 2018. “Explaining Explanations: An Overview of Interpretability of Machine Learning.” *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA):* 80–89. https://doi.org/10.1109/DSAA.2018.00018.

Goldman, Alvin and Bob Beddor. 2021. “Reliabilist Epistemology.” In *Stanford Encyclopedia of Philosophy,* edited by Edward N. Zalta. https://plato.stanford.edu/archives/sum2021/entries/reliabilism/.

Grote, Thomas and Philipp Berens. 2022. “How Competitors Become Collaborators: Bridging the Gap(s) between Machine Learning Algorithms and Clinicians.” *Bioethics* 36, no. 2: 134–142. https://doi.org/10.1111/bioe.12957.

Gulshan, Varun, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, et al. 2016. “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs.” *JAMA* 316, no. 22: 2402–2410. https://doi.org/10.1001/jama.2016.17216.

Günther, Mario and Atoosa Kasirzadeh. 2022. “Algorithmic and Human Decision Making: For a Double Standard of Transparency.” *AI & Society* 37: 375–381. https://doi.org/10.1007/s00146-021-01200-5.

Hinton, Geoffrey E., Oriol Vinyals, and Jeff Dean. 2015. "Distilling the Knowledge in a Neural Network." *arXiv* preprint: 1503.02531. https://doi.org/10.48550/arXiv.1503.02531.

Horwick, Jeremy H. 2011. *The Philosophy of Evidence-Based Medicine*. London: John Wiley & Sons.

Ilanchezian, Indu, Dmitry Kobak, Hanna Faber, Focke Ziemssen, Philipp Berens, and Murat Seckin Ayhan. 2021. "Interpretable Gender Classification from Retinal Fundus Images Using BagNets." In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III,* edited by Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, 477–487. Berlin: Springer. https://doi.org/10.1007/978-3-030-87199-4_45.

Ioannidis, John P.A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2, no. 8. https://doi.org/10.1371/journal.pmed.0020124.

Karaca, Koray. 2021. "Values and Inductive Risk in Machine Learning Modelling: The Case of Binary Classification Models." *European Journal for Philosophy of Science* 11, no. 4, art. 102. https://doi.org/10.1007/s13194-021-00405-1.

Kiener, Maximilian. 2021. "Artificial Intelligence in Medicine and the Disclosure of Risks." *AI & Society* 36, no. 3: 705–713. https://doi.org/10.1007/s00146-020-01085-w.

Krishnan, Maya. 2020. "Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning." *Philosophy & Technology* 33, no. 3: 487–502. https://doi.org/10.1007/s13347-019-00372-9.

Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn." *Nature Communications* 10, no. 1, art. 1096. https://doi.org/10.1038/s41467-019-08987-4.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521, no. 7553: 436–444. https://doi.org/10.1038/nature14539.

Lipton, Zachary C. 2018. "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery." *Queue* 16, no. 3: 31–57. https://doi.org/10.1145/3236386.3241340.

Liu, Xiaoxuan, Ben Glocker, Melissa M. McCradden, Marzyeh Ghassemi, Denniston, Alastair K., and Lauren Oakden-Rayner. 2022. "The Medical Algorithmic Audit." *The Lancet Digital Health* 4, no. 5: e384–e397. https://doi.org/10.1016/S2589-7500(22)00003-6.

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

Lundberg, Scott and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems* 30, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4768–4777. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

McDougall, Rosalind J. 2019. "Computer Knows Best? The Need for Value-Flexibility in Medical AI." *Journal of Medical Ethics* 45, no. 3, 156–160. https://doi.org/10.1136/medethics-2018-105118.

McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, et al. 2020. “International Evaluation of an AI System for Breast Cancer Screening.” *Nature* 577, no. 7788: 89–94. https://doi.org/10.1038/s41586-019-1799-6.

Molnar, Christoph. 2020. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu. com.

Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller. 2018. “Methods for Interpreting and Understanding Deep Neural Networks.” *Digital Signal Processing* 73: 1–15. https://doi.org/10.1016/j.dsp.2017.10.011.

Mullainathan, Sendhil and Ziad Obermeyer. 2022. “Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care.” *Quarterly Journal of Economics* 137, no. 2: 679–727. https://econpapers.repec.org/scripts/redir.pf?u=http%3A%2F%2Fhdl.handle.net%2F10.1093%2Fqje%2Fqjab046;h=repec:oup:qjecon:v:137:y:2022:i:2:p:679-727.

Nyrup, Rune and Diana Robinson. 2022. “Explanatory Pragmatism: A Context-Sensitive Framework for Explainable Medical AI.” *Ethics and Information Technology* 24, no. 1, art. 13. https://doi.org/10.1007/s10676-022-09632-3.

Parker, Wendy S. 2020. “Model Evaluation: An Adequacy-for-Purpose View.” *Philosophy of Science* 87, no. 3: 457–477. https://doi.org/10.1086/708691.

Post, Piet N., Hans de Beer, and Gordon H. Guyatt. 2013. “How to Generalize Efficacy Results of Randomized Trials: Recommendations Based on a Systematic Review of Possible Approaches.” *Journal of Evaluation in Clinical Practice* 19, no. 4: 638–643. https://doi.org/10.1111/j.1365-2753.2012.01888.x.

Poursabzi-Sangdeh, Forough, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. “Manipulating and Measuring Model Interpretability.” In *CHI ’21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,* 1–52. New York: Association for Computing Machinery. https://doi.org/10.1145/3411764.3445315.

Ratti, Emanuele and Mark Graves. 2022. “Explainable Machine Learning Practices: Opening Another Black Box for Reliable Medical AI.” *AI and Ethics*. Advance online publication. https://doi.org/10.1007/s43681-022-00141-z.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.” In *KDD ’16, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1135–1144. New York: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

Rudin, Cynthia. 2019. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” *Nature Machine Intelligence* 1, no. 5: 206–215. https://doi.org/10.1038/s42256-019-0048-x.

Rudin, Cynthia, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. “Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges.” *Statistics Surveys* 16: 1–85. https://doi.org.10.1214/21-SS133.

Sackett, David L. William M.C. Rosenberg. 1995. “The Need for Evidence-Based Medicine.” *Journal of the Royal Society of Medicine* 88, no. 11: 620–624. https://doi.org/10.1177/014107689508801105.

Senn, Stephen. 2013. “Seven Myths of Randomisation in Clinical Trials.” *Statistics in Medicine* 32, no. 9: 1439–1450. https://doi.org/10.1002/sim.5713.

Shalev-Shwartz, Shai and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press.

Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. "Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods." In *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 180–186. New York: Association for Computing Machinery. https://doi.org/10.1145/3375627.3375830.

Slack, Dylan, Sophie Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021a. "Counterfactual Explanations Can Be Manipulated." *Advances in Neural Information Processing Systems* 34: 62–75. https://papers.nips.cc/paper/2021/file/009c434cab57de48a31f6b669e7ba266-Paper.pdf.

Slack, Dylan, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2021b. "Reliable Post Hoc Explanations: Modeling Uncertainty in Explainability." *Advances in Neural Information Processing Systems* 34. https://par.nsf.gov/servlets/purl/10381222.

Stegenga, Jacob. 2018. *Medical Nihilism*. Oxford: Oxford University Press.

Stokes, Dustin. 2021. "On Perceptual Expertise." *Mind & Language* 36, no. 2: 241–263. https://doi.org/10.1111/mila.12270.

Tomašev, Nenad, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, et al. 2019. "A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury." *Nature* 572, no. 7767: 116–119. https://doi.org/10.1038/s41586-019-1390-1.

Tomašev, Nenad, Natalie Harris, Sebastien Baur, Anne Mottram, Xavier Glorot, Jack W. Rae, Michal Zielinski, et al. 2021. "Use of Deep Learning to Develop Continuous-Risk Models for Adverse Event Prediction from Electronic Health Records." *Nature Protocols* 16, no. 6: 2765–2787. https://doi.org/10.1038/s41596-021-00513-5.

Tschandl, Philipp, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, et al. 2020. "Human–Computer Collaboration for Skin Cancer Recognition." *Nature Medicine* 26, no. 8: 1229–1234. https://doi.org/10.1038/s41591-020-0942-0.

Varoquaux, Gaël and Veronika Cheplygina. 2022. "Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future." *Npj Digital Medicine* 5, no. 1: 1–8. https://doi.org/10.1038/s41746-022-00592-y.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law & Technology* 31, no. 2: 841–887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf.

Williamson, Jon. 2021. "Evidential Pluralism and Explainable AI." *The Reasoner* 15, no. 6: 55–56. https://blogs.kent.ac.uk/thereasoner/files/2021/11/TheReasoner-156.pdf.

Watson, David S. 2022. "Conceptual Challenges For Interpretable Machine Learning." *Synthese* 200, no. 2, art. 65. https://doi.org/10.1007/s11229-022-03485-5.

Watson, David S., Jenny Krutzinna, Ian N. Bruce, Christopher E.M. Griffiths, Iain B. McInnes, Michael R. Barnes, and Luciano Floridi. 2019. "Clinical Applications of Machine Learning Algorithms: Beyond the Black Box." *British Medical Journal* 364, art. l886. https://doi.org/10.1136/bmj.l886.

Yala, Adam, Peter G. Mikhael, Constance Lehman, Gigin Lin, Fredrik Strand, Yung-Liang Wan, Kevin Hughes, et al. 2022. "Optimizing Risk-Based Breast Cancer Screening Policies with Reinforcement Learning." *Nature Medicine* 28, no. 1: 136–143. https://doi.org/10.1038/s41591-021-01599-w.

Yim, Jason, Reena Chopra, Terry Spitz, Jim Winkens, Annette Obika, Christopher Kelly, Harry Askham, et al. 2020. "Predicting Conversion to Wet Age-Related Macular Degeneration Using Deep Learning." *Nature Medicine* 26, no. 6: 892–899. https://doi.org/10.1038/s41591-020-0867-7.

Zednik, Carlos. 2021. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." *Philosophy & Technology* 34, no. 2: 265–288. https://doi.org/10.1007/s13347-019-00382-7.

Zednik, Carlos and Hannes Boelsen. 2022. "Scientific Exploration and Explainable Artificial Intelligence." *Minds and Machines* 32, no. 1: 219–239. https://doi.org/10.1007/s11023-021-09583-6.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" *Philosophy & Technology* 32, no. 4: 661–683. https://doi.org/10.1007/s13347-018-0330-6.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. "Understanding Deep Learning (Still) Requires Rethinking Generalization." *Communications of the ACM* 64, no. 3: 107–115. https://doi.org/10.1145/3446776.