Philosophy & Technology (2022) 35: 5
https://doi.org/10.1007/s13347-022-00505-7

RESEARCH ARTICLE

Accuracy and Interpretability: Struggling with the Epistemic Foundations of Machine Learning-Generated Medical Information and Their Practical Implications for the Doctor-Patient Relationship

Florian Funer[1]

Received: 4 May 2021 / Accepted: 17 January 2022 / Published online: 29 January 2022

Abstract

The initial successes in recent years in harnessing machine learning (ML) technologies to improve medical practice and benefit patients have attracted attention in a wide range of healthcare fields. Particularly, it should be achieved by providing automated decision recommendations to the treating clinician. Some hopes placed in such ML-based systems for healthcare, however, seem to be unwarranted, at least partially because of their inherent lack of transparency, although their results seem convincing in accuracy and reliability. Skepticism arises when the physician as the agent responsible for the implementation of diagnosis, therapy, and care is unable to access the generation of findings and recommendations. There is widespread agreement that, generally, a complete traceability is preferable to opaque recommendations; however, there are differences about addressing ML-based systems whose functioning seems to remain opaque to some degree—even if so-called explicable or interpretable systems gain increasing amounts of interest. This essay approaches the epistemic foundations of ML-generated information specifically and medical knowledge generally to advocate differentiations of decision-making situations in clinical contexts regarding their necessary depth of insight into the process of information generation. Empirically accurate or reliable outcomes are sufficient for some decision situations in healthcare, whereas other clinical decisions require extensive insight into ML-generated outcomes because of their inherently normative implications.

Keywords Machine learning (ML) · Epistemological challenges · Medical knowledge · Accuracy · Interpretability · Doctor-patient relationship · Responsibility

This article is part of the Topical Collection on *Information in Interactions between Humans and Machines*

✉ Florian Funer
florian.funer@uni-tuebingen.de

[1] Eberhard Karls University Tübingen, Tübingen, Germany

1 Introduction

The period in which the amount of medical knowledge doubles is increasingly becoming shorter (cf. Densen, 2011): it was estimated that it took approximately 50 years in 1950 for our knowledge to double; that time had shortened to only 7 years in 1980, just under 4 years in 2010, and approximately 73 days in 2020. Especially in medicine, the growth of information in recent decades has been overwhelming, not only in quantity but also in potential importance for improving prevention, diagnostics, therapy, and care for the benefit of the patient. The types of information range, for example, from general theoretical approaches about human physiological processes to empirical evidence of pharmacological and other therapeutic modes of action to psychological and social interactions of the unique patient. This increasing complexity has led and continues to lead to fragmentation and differentiation of medical subfields and their actors. Nonetheless, the constantly growing information challenges and even exceeds the processing and knowledge capabilities of physicians. The new hope in medical practices for significant support and at least partial reduction of physicians' limited capabilities is formed by a variety of "clinical decision support systems."[1]

As firm steps in this digital transformation of healthcare, increasing attention has been focused on support systems whose operating mode is based on machine learning (ML) algorithms. Connected with increasingly powerful processor technologies, these ML-based systems are already enabling improvements in accuracy and efficiency across several medical disciplines (cf. Topol, 2019). To date, they have served experimentally to identify and evaluate differential diagnoses and determine "best" therapy options. However, to warrant a high quality of medicine and healthcare in the future, the expectation of technical progress in this field must undergo a serious examination of the moral challenges and their epistemic presuppositions. Consequently, this paper seeks the epistemic foundations of ML-generated medical information and its normative implications for the doctor-patient relationship.[2] On the one hand, some authors have emphasized that users—in this case, physicians—need to *understand* in a certain way *why* an algorithm delivers its outcome to maintain the patients' trust in its resulting decisions. Therefore, they demand that ML applications provide explicable or interpretable processes and/or outcomes (cf. Bjerring & Busch, 2021; Heinrichs & Eickhoff, 2020; Holzinger et al., 2020; Rudin & Radin, 2019; Tsamados et al., 2021). On the other hand, voices have been raised that consider such explicability for ML in healthcare to be an overvalued aim and therefore consider that merely the proof of a certain accuracy (cf. London, 2019) or reliability (cf. Durán & Jongsma, 2021) sufficiently justifies its use.

[1] "Clinical decision support systems" will be used for "software that [is] designed to be a direct aid to clinical decision-making, in which the characteristics of an individual patient are matched to a computerized clinical knowledge base and patient-specific assessments or recommendations are then presented to the clinician or the patient for a decision" (Sim et al., 2001).

[2] The term "doctor-patient-relationship" is used here as more common in the clinical jargon. It can also be applied for relationships between patients and other therapeutic and nursing agents in the healthcare field.

By examining the epistemic foundations of medical knowledge, this paper aims to explain why, even though accuracy is often prima facie sufficient in medical contexts, deeper transparency in ML-generated information is normatively necessary in some medical decisions, even if it may sometimes remain out of reach. This paper demonstrates that this is mainly due to the peculiarity of medical knowledge, which Solomon (2015) calls "untidy pluralism": medical knowledge is thus characterized by a variety of methods and heuristics (e.g., "consensus conferences," "evidence-based medicine," "translational medicine," and "narrative medicine"; cf. Ibid.). Due to its ability to provide multiple appropriate methods for solving the same problem, different results and thus incoherence may occur. Hence, attempting a moral assessment in such cases is difficult because these different methods may represent different goals. To avoid falling prey to one-sided reliabilistic forms of knowledge (to which agents may be tempted by using ML-generated information; cf. Bjerring & Busch, 2021), transparency and interpretability of the goals pursued as well as the information provided can better address the epistemic and normative requirements of deliberation in medical practice. Acknowledging the variety and, in some situations, equivalence of methods utilized for medical knowledge leads to the recognition that only a clearer view on the genesis of and reasons for a specific ML-generated outcome enables its necessary epistemic contextualization and normative evaluation.

This paper is structured as follows: First, a brief overview of the current state of ML in healthcare and its obstacles in interactions between physicians and patients is provided (Sect. 2). This presentation is a starting point for the debate about basic conditions of medical knowledge and understanding in general and resulting epistemic challenges in utilizing ML-generated information to support clinical decisions (Sect. 3). Disagreement exists, especially regarding the criteria of information that necessarily underlies responsible medical decisions. Under keywords such as *opacity*, *black box*, *reliability*, *accuracy*, *transparency*, *describability*, *explicability*, and *interpretability*, attempts have been made to problematize or resolve the insight into ML-generated information. However, the plausibility and normativity requirements applied to a medical decision—and thus to an ML-supported decision—are fundamentally determined, I argue, by the scope of the decision to be made. Medical knowledge and information generated by ML do not differ categorically, but the latter manages to represent only part of the former. This results in significant normative and communicative implications for the deliberative doctor-patient relationship (Sect. 4). A short conclusion summarizes the relevant findings and provides perspectives on subsequent questions (Sect. 5).

2 Machine Learning in Healthcare Contexts: a Short Overview

Clinical decision support, including technical forms, are not a novelty in healthcare (consider laboratory medicine, imaging techniques, and interprofessional consultations). Hence, their implementation in the diagnosis and treatment process does not imply a categorical change in physicians' tasks. However, the dynamically evolving field of ML-based systems promises quantitative and qualitative expansions in decision support. These extensions are enabled by advances in generating and processing information as well as breakthroughs in ML in which non-rule-based algorithms “learn,” that is, generate insights, by identifying specific patterns and regularities in a defined set of raw data (cf. Hinton, 2007; Schmidt-Erfurth et al., 2018).[3] The goal of ML is to “intelligently” link sets of data and in doing so generate information that allows users to identify and interpret correlations, draw conclusions, and make predictions. In contrast to traditional, rule-based algorithms, ML-based support systems do not “require specific instructions that detail every step the program must take” (Ahuja, 2019). Inspired by the human brain, research on ML continues to “imitate the neural structure of the nervous system” by creating artificial neural networks (Pearson, 2017, quoted by Ahuja, 2019; cf. Hinton, 2007). Different forms of such networks are being developed to suit different areas of life, but the performance of all of these networks improves with an increasing amount of data (Hinton, 2007).

The usage of ML-based systems pledges considerable opportunities to integrate the constantly growing medical knowledge and the significant amounts of data into medical practice to substantially improve the benefit to the patient. Recently, thanks to this potential, research on ML and its implementation in the clinical setting has proliferated in various fields of healthcare (cf. e.g., De Fauw et al., 2018; Esteva et al., 2019; Johnson et al., 2018; Krittanawong et al., 2017; Patel et al., 2021; Salto-Tellez et al., 2018).

The list of expectations and hopes for the healthcare potential of ML-based systems is long. Medical practice could become more accurate and individually customized with less harm and fewer side effects; additionally, it could be less costly in the long term by preempting diseases. In this respect, medicine could be more preventive. For example, “the research firm Frost & Sullivan estimates that AI has the potential to improve patient outcomes by 30% to 40% while reducing treatment costs by up to 50%” (quoted by Ahuja, 2019). The rapid technological progress has led to largely substitute the question of whether such systems could partially or completely replace human practitioners in certain tasks (e.g., Ahuja, 2019; Pearson, 2017), only by the question of when this could be possible. Debates about such speculative “replacements” of certain professional groups hardly seem fruitful[4]; instead, it seems more realistic that the usage of ML-based systems—as with previous technological developments—will lead to sometimes more and sometimes less pronounced shifts in the tasks of human medical agents.

[3] According to the European High-Level Expert Group on AI (2019), the most widely used approaches are, first, “*supervised learning*,” in which results are learned from input examples; second, “*unsupervised learning*,” in which patterns are identified by the machine itself from raw data; and third, “*reinforcement learning*,” in which the system itself develops recommendations and these are subsequently evaluated by third parties through positive or negative reward signals, with the goal of maximizing positive rewards. Without going into more detail about the variety of systems and applications at this point, this overview already illustrates the range of systems under discussion.

[4] Despite all evidence for equivalent or even better performance of ML systems compared to clinical experts, the lack of validity of such comparisons between physicians and machines has already been pointed out. For example, according to some meta-analyses, many studies do not sufficiently consider external validation, are only retrospective, and were conducted outside the clinical setting (Liu et al., 2019; Nagendran et al., 2020).

However, these goals are obtainable only if physicians' knowledge and capabilities and the outcomes provided by ML complement each other in the optimal manner. This belief is based on the known limitations of ML: for example, because of its deficient contextual knowledge, ML-based systems may make errors and propose faulty recommendations (Cabitza et al., 2017; Orwat, 2019), which would be clear to human decision makers and thus would likely be bypassed. There is also a major threat in validating the ML algorithms from raw data, which may lead to unfair treatment of certain groups of patients due to different biases.[5] Consequently, Grote and Berens (2020) state, "[w]hile machine learning algorithms will not cure any disease by themselves any time soon, there is clear potential to improve diagnostic decision-making based on the progress we are seeing today." Even in cases in which ML seem to outperform the physician, an accompanying evaluation and review by a physician is not superfluous or substitutable per se. However, identifying the ideal manner of cooperation and implementation in specific clinical contexts are associated with epistemic, normative, and communicative challenges concerning, for example, the transparency of data generation, the insight into aspects considered in decision-making, and the allocation of responsibilities. This field of central challenges is analyzed and evaluated in the following sections by approaching how ML-generated medical information could shape our knowledge and understanding in healthcare.

3 Resolving the Epistemic Gap: Why All the Effort of Transparency?

As mentioned, some of the main barriers to the usage of non-rule-based algorithms in clinical practice are their lack of transparency and level of traceability of given outcomes, which is considered insufficient. The complex architecture of ML-based algorithms—especially those utilizing deep neural networks—makes it difficult or even impossible to understand *how* variables are combined to generate outcomes such as predictions or recommendations (cf. Rudin & Radin, 2019; Zednik, 2021). Basically, "deep neural networks consist of layers of nodes that each use simple mathematical operations to perform a specific operation on the activation of the layer before, leading to the emergence of increasingly abstract representations" of the input data (Grote & Berens, 2020). In these multi-layered networks, the larger the underlying dataset, the better the outcome produced by the algorithm. With the growing accumulation of input instances, the relative weights of the various nodes

[5] Some of the numerous potential biases have been adequately pointed out elsewhere (London, 2019; Hutson, 2021; Dalton-Brown, 2020). Developing fair ML systems is difficult because "there is no value-neutral way to select the training data, the objective function, the model, the benchmark task, the appropriate notion of fairness, and so on" (Genin & Grote, 2021). Nevertheless, such biases are not a novelty in medical knowledge generation. Consider, for example, the long-known but nonetheless—even today—inadequately considered appeals that sex-based physiological differences should lead to different therapies (Baggio et al., 2013). Clearly, our medical knowledge is also biased toward other factors, such as ethnicity and the like. But at least such undesired biases *can be detected* and subsequently taken into account when interpreting the data; however, if biases remain undetected, they cannot be considered at all when interpreting and applying them to individual cases.

in the neural network adjust themselves to produce the most accurate mathematical representation possible of the input information. Classifications tested on extensive data enable the system to produce highly accurate statements about probabilities for the presence of a certain finding (e.g., diagnostic image analysis) or the occurrence of a certain event (e.g., prognostic chances of therapeutic success or failure). Since the algorithm is fed by collections of overwhelmingly large amounts of data, the coming into existence of its probability statements is often "neither foreseeable nor transparent to the programmer" (Heinrichs & Eickhoff, 2020). London (2019) illustrates this phenomenon of "black boxes" as follows:

> Even when techniques are used to identify features or a set of features to which a model gives significant weight in evaluating a particular case, the relationships between those features and the output classification can be both indirect and fragile. A small permutation in a seemingly unrelated aspect of the data can result in a significantly different weighting of features. Moreover, different initial settings can result in the construction of different models.

Such systems are therefore characterized by a certain degree of epistemic opacity, which means that the complex and multi-dimensional mathematical processing performance of the algorithm is not comprehensible or is comprehensible only to a limited extent via the language of human agents. Of course, some systems may be more accessible to IT experts than, for example, to a doctor or a patient, which is why we can also speak of a "relative concept" (Smith, 2021). However, as long as a system is epistemically opaque, its outcomes elude sufficient interpretation; they remain inaccessible to some degree to everyone.

3.1 The Urgent Call for Explicable or Interpretable Algorithms

To optimally counter the opacity of the algorithms, researchers are attempting to trace the types of patterns, the identified statistical correlations, and the pathways taken in the process. However, this research, forming the field of so-called explicable AI (xAI) or interpretable ML (iML) (Hutson, 2021; Tsamados et al., 2021; Holzinger et al., 2020), is still in its infancy. Nonetheless, the white papers and recommendations of numerous large companies, including Microsoft and Google, and policy guidance institutions, including the World Economic Forum and the EU Commission, contain a principle called *explainability* or *explicability*. Especially in morally sensitive contexts, they say, it is important for decision-making to provide explanations for how certain recommendations came about. Only if, when applying an ML-based recommendation, we can rule out that this recommendation is not based on inappropriate considerations and biases, does such a recommendation seem to be acceptable. "If the algorithm is not explicable," Robbins (2019) states, such inappropriate considerations "may be used without our knowledge."

Consistently, a wide range of methods have been developed to increase the transparency and traceability of opaque algorithms and to turn some black boxes into gray ones. To date, two possible approaches have been helpful (Heinrichs & Eickhoff, 2020; Hutson, 2021; Molnar, 2021)[6]:

On the one hand, the goal of the *explicability of the operating mode of an entire algorithm* ("global" or "model explicability") is to provide the interacting user (presumably, due to the IT knowledge required, this refers especially to ML developers) with the best possible insight into how the algorithm works, typically by enabling the user to trace the foundations on which a model develops recommendations. For this, the user must know which data basis was employed, which aspects were considered, and how these aspects were weighted and balanced to question the resulting recommendations concerning their plausibility. Such global explicability is mostly achieved by utilizing iML systems to approximate the predictions of the black box ML. Then, by interpreting the iML, we can draw conclusions about the black box model itself (cf. Molnar, 2021).

On the other hand, the *explicability of certain individual results* ("local" or "result explicability") focuses on the features selected and weighted in the specific case, which may be of particular interest to the physician and patient. Currently, to convert singular opaque outcomes into interpretable ones, local surrogate models, such as local interpretable model-agnostic explanations (LIME; cf. Molnar, 2021; Visani et al., 2020), are utilized. With numerous tests performed on the opaque system, LIME approximates what happens with the individual outcome when the underlying dataset fed into the black box is altered several times. On this basis, LIME generates "a new dataset consisting of perturbed samples and the corresponding predictions of the black box model" (Molnar, 2021).

Nonetheless, the efforts of both kinds of explicability demonstrate that the more complex the algorithm, the more difficult the explicability is to be realized without inappropriate oversimplification. Consequently, a wide range of differently complex and explicable ML systems can be imagined, "with a general trade-off between performance and interpretability reflecting model complexity" (Heinrichs & Eickhoff, 2020).

Despite the intuitively plausible wish for explicability, at the epistemic center is the question of what makes an ML-based system or an individual decision explicable at all. According to Floridi et al. (2018), systems should be called *explicable* only when they grant "a factual, direct, and clear explanation of the decision-making process, especially in the event of unwanted consequences." When we interrogate a decision-making process regarding its explicability, we seek for its reasons. In the majority of cases, it is insufficient for us to acquire a descriptive explanation of *what* happens and *how* it occurs. Instead, we ask for the causally and intentionally relevant and effective reasons of a decision, that is, *why* a specific decision was made (Robbins, 2019). The answer regarding the explanations of the results, which can extend to all ML-based algorithms that the algorithm has "learned" it from its training data or instances, does not satisfy our need for answers regarding why the decision was made. What we really need to evaluate ML-generated information is "the considerations that contributed to the result in question" (Ibid.). Consequently, we

[6] For a more detailed overview of the explanatory methods and tools currently available, see Molnar, 2021.

see that the explicability of an algorithm is not an end in itself but rather a helpful and possibly necessary feature to satisfy the desire of the interacting human for the fullest possible *understanding*. To bring this concern for a deeper insight that goes beyond simple description to the fore, some authors employ instead of explicability the term *interpretability* (Heinrichs & Eickhoff, 2020; Rudin & Radin, 2019). Leaving aside the ambiguity of the term, it can be said that iML systems should provide detailed and preferably exhaustive information about their functioning and their process of generating individual outcomes. Therefore, the physician is able not only to trace how the information of the ML-based system came about but also to present it to the patient and thus employ the information in shared decision-making. If necessary, both the physician and the patient could then offer feedback about incorrect information or weightings that are evaluated as inappropriate in the specific situation. To achieve this, Rudin and Radin (2019) argue that we should not depend on algorithms that are analyzed ex post for being explicable; rather, they advocate for systems that are already constructed to be interpretable.

Yet, while for some processes forms of explicable or interpretable ML-based systems can be developed, other complex procedures that utilize immense amounts of data may remain not explicable or interpretable to human actors due to their limited processing capacity. Consequently, should the desire for a perhaps unreachable explicability impede the usage of such algorithms?

3.2 Why All the Fuss? Accuracy Has Often Been Enough Thus Far…

In contrast to commentators who emphasize the importance of interpretable ML-generated outcomes, some criticize the importance of explicability in its sweeping nature (Durán & Jongsma, 2021) or question its significance entirely in the medical field. A representative of the latter position is Alex J. London, who unfolds his questioning of the "explanatory power" of medical knowledge as follows (London, 2019): in other fields, a comparatively high completeness of causally significant interrelations has been or can be achieved; however, the knowledge about underlying causal interrelations in medicine is merely "in its infancy" (Ibid.). Pathomechanisms and therapeutic modes of functioning are often unknown or poorly understood; consequently, "decisions that are atheoretic, associationist, and opaque are commonplace in medicine" (Ibid.). Considerable parts of empirical medical knowledge have been applied for many years in spite of the lack of causal insight into their mechanism of action, as in the case of aspirin or lithium (Ibid.). Other therapies that were based on causal hypotheses—that is, theoretical attempts of explanation—were later revealed to be incorrect. Rigorously gathered empirical findings are therefore said to be *more reliable* and to *reflect causal interrelations better* than "theoretical claims that purport to ground and explain them" (Ibid.). Our medical knowledge and practice, London concludes, would be primarily "a mixture of empirical findings and inherited clinical culture," causing recommendations based on them to "reflect [the] experience of benefit without enough knowledge of the underlying causal system to explain how the benefits are brought about" (Ibid.).

Without presenting London's full illustration of this point, I summarize his argument roughly as follows: Uncertainty, especially involving causal interrelations, is the rule rather than the exception in medical practice. Clinical decisions often originate in the physician's apparently opaque and often inaccessible neural network. Therefore, equally opaque information or recommendations from artificial neural networks are "not radically different" (Ibid.) from the physician's recommendations. Consequently, because our ability to consider numerous features remains fragmentary and thus limited, non-rule-based algorithms are comparatively superior to us in terms of their accuracy and are therefore sometimes preferable. Hence, the focus of our attention should be less on efforts toward explicability or interpretability of ML-generated medical outcomes and more on the empirical validation of their accuracy or reliability (Ibid.).

These fundamentally different positions about the necessity of insight into ML-based outcomes,[7] both of which indicate an intuitive plausibility, require that we examine the epistemic presuppositions and their normative relevance within the communication process between physicians and patients. Although the current debate about ML-generated information is recent regarding its application, the argument can be traced to the foundations of medical knowledge and philosophy of science.

3.3 Efforts to Make Medical Information Accessible: the Normative Relevance of Understanding

According to Clemens Sedmak (2003: 10 f.; author's translation), working epistemically includes "all the efforts we have to make finding our way in the world." It is "work on orientation by introducing differentiations" and thus generating an "order" in which we can meaningfully classify the phenomena we encounter every day (Ibid.). Sedmak emphasizes that the grasp of reality, which functions as a prerequisite for successful action, may not be completely available to us but requires effort on our part. Gaining knowledge in medicine through effort is certainly nothing new. Sedmak, however, does not mean the efforts that relate to the daily production of vast amounts of information but its correct *understanding*. According to Sedmak, understanding is the *embedding* of new information *in the order* of our already existing individual knowledge.

Considerable parts of modern medical knowledge rely first on empirically collected data to evaluate the effectiveness of clinical interventions, that is, they are evidence based. With various methods—preferably randomized controlled trials (RCTs)—reliable statements should be made about probabilities for certain predictors, certain progressions, and certain endpoints. In this view, the entirety of our medical information leads us increasingly closer to a coherent system of statements

[7] Ultimately, the divergent positions represent a fundamental dispute about the epistemic nature of justification in their application: Whereas proponents of high-accurate but opaque systems for obtaining true beliefs seem to require exclusively the reliability of evidence, proponents of interpretable systems tend to emphasize the justification of propositions among themselves.

about reality (epistemic achievements), with the aim of describing it as ideally as possible.[8] Many of our best scientific theories to date are likely, strictly speaking, incorrect; thus, we are only close to the truth, that is, close to reality as it really is, or approximately true (cf. for approximate truth, e.g., Hardin & Rosenberg, 1982; Putnam, 1982; Smith, 1998). Nevertheless, there is inevitably what I call an *epistemic gap* between this information, which is based, for example, on generalizations, categorizations, and cancellations of so-called statistical outliers and the concrete case at hand with its unique circumstances.

The individual patient embodies reality as it really is; he is not a statistical representation of reality. The patient therefore eludes statistical simplifications, cannot fully be captured in categories, or may just *be* a statistical outlier. Due to their inherently simplifying design of collection, statistics, even if considered all together, categorically cannot exhaustively describe reality as it really is. This criticism is not a novelty but is well-known in the epistemology of evidence-based medical information as the problem of external validity (cf. Solomon, 2015: 141 ff.; Worrall, 2007).[9] To address this problem sufficiently, applying trial results to clinical decision-making requires "a good deal of background knowledge" (Ibid.; cf. Cartwright, 2007a/2007b) to produce an overall judgment that fits the situation. Consequently, medical knowledge does not comprise solely probabilistic results, but its usability needs further causal reasoning based on multiple methods with different kinds of evidence, justifications, and heuristics (cf. Solomon, 2015; Worrall, 2007).

Accordingly, even an ML system with access to many statistics that "take all this evidence into consideration – a feat that might not be possible for individual practitioners –" (Bjerring & Busch, 2021) does not achieve a sufficiently complete picture of the patient due to the epistemic gap of the data on which it is based.[10] The treatment of an individual patient therefore requires an interpretative capacity that examines, evaluates, and selects all the evidence regarding its relevance for the individual case.

What do I mean when I speak of the relevance of *understanding* in clinical decision-making? I mean the attempt of grasping the relationship between different pieces of information, especially the relations between causes and effects or "dependency relations" (Grimm, 2005, 2011; cf. Zagzebski, 2009: 142 ff.). Only by "embedding" evidence into one's network of other evidence can one achieve

[8] My argument here follows scientific realism, according to which "our best scientific theories give true or approximately true descriptions of observable and unobservable aspects of a mind-independent world" (Chakravartty, 2017).

[9] Reasons for the lack of external validity of medical studies include variability in patients (age, sex, severity of disease, risk factors, comorbidities, ethnicity, socioeconomic status), treatments (dose, timing of administration, duration of therapy, other medications), and setting (quality of care) (cf. Rawlins, 2008; quoted by Solomon, 2015: 143).

[10] Thus, the problem here lies primarily with the information underlying the ML system. An opaque ML system runs the risk of further obscuring the limitations of external validity (but also internal validity) through its mechanism of linking such data together. The supposed progress of making limitations of research as transparent as possible to interpretation would thus potentially be abandoned.

understanding (=individual knowledge).[11] Surely it is possible that we also integrate at least some false beliefs about propositions into our network, that is, we develop a "false understanding."[12] Nevertheless, such understanding as an interpretative capacity has a greater practical value than mere true beliefs because of its stability for following individual life goals (cf. for the value of understanding, see Zagzebski, 2009; Pritchard, 2009; de Regt et al., 2009; Grimm et al., 2017). Understanding enables us to integrate new information not only into our background knowledge about reality as it really is but also—based on this—into our life-guiding network of moral beliefs. As Zagzebski posits, "[I]f we adopt a moral belief on testimony [e.g., from an opaque recommendation] without understanding the broader moral reasons behind it, we will be unable to generalize our judgment to make relevantly similar judgments" (Zagzebski, 2009: 148). To make reasonable sense about anything in life, it is necessary to understand it and integrate it in the "order" of the manifold dimensions of one's personality, such as biographical, situational, social, and moral aspects.

These ever-present normative implications cause challenges in employing epistemically opaque ML-based outcomes: addressing evidence in an individual case is a formula that could be solved with ML-based representations and processing steps if, and only if the defined, expected benefit, that is, the goal, can be assumed to be universally generalizable. Otherwise, there would be a risk of imposing a goal (e.g., the highest possible medical outcome) on the patient at hand, which could run counter to the patient's goals. Certainly an ML-based algorithm, due to its high-capacity performance, could better calculate what is "best" for an individual case from a one-sided, statistical, and medical perspective, such as increasing the patient's lifetime. However, the interests, values, and goals of the majority of individuals extend beyond this goal of lifetime. Unfortunately, the search for a more universal goal also offers little hope. The determination of a single universally valid and operationalizable ultimate goal, which ideally could be programmed as the goal perspective in an algorithm, has long been controversial in philosophy and ethics regarding its existence, and its verbalized concretion and application may be impossible. If one notes the plethora of existing and potential human goals, however, then the practical possibility of registering and considering them in an algorithm to a large extent—and thus not simply mono- or oligo-dimensionally—remains questionable.[13]

Decisions regarding patient medical treatment concern not only a truncated "statistically captured view" but also an integral picture of the whole patient;

[11] I explicitly do not mean here the abstract knowledge of the professional community or the like, which may expand diachronically and possibly also falsify the earlier "understanding.".

[12] There is disagreement whether a belief must merely be subjectively appropriate or whether it must also be objectively appropriate, that means, whether it must be a reliable matter of truth in order to be knowledge (cf. e. g. Grimm 2011, 90). I will not be able to conclude this discussion either. But even if the latter is true, this belief only acquires practical relevance for the individual person if she integrates it into her own "network" of beliefs.

[13] To be understood correctly: I do not want to deny the potential existence of extensive or perhaps even universal goals, but I doubt their machine-adequate formulation and complete operationalizability such that they could justify a universal application of these "programmed" goals to every individual for whose treatment a ML-based recommendation is made.

consequently, the other dimensions relevant to the patient must be included in medical decisions. The debate opening up here regarding possible definitions of quality of life is obvious. Even if we could collect and operationalize the regularities and categorizations of a large part of the dimensions relevant to humans[14] and if, thanks to large amounts of data, we were empirically closer to so-called personalized medicine, even then there would still be this epistemic gap between the statistical (in comparison to reality as it really is always reductionistic) data and the concrete individual person, who cannot be statistically predicted with high certainty regarding her personal circumstances. The long-held truism of medical practice—treating not a disease but a person *with* a disease—is more relevant for the implementation of ML systems in healthcare than has been seen so far.

3.4 How Much Insight Is Needed in ML-Generated Outcomes for What Kind of Medical Decisions? An Approximation

What does this seemingly lofty digression about people's ultimate goals and therefore medicine's goal perspectives mean for the necessity or non-necessity of insight into ML systems? The advantage of interpretable systems is that the agent utilizing them, that is, the treating physician, optimally obtains insight into the adequate or perhaps non-adequate factors considered, the fitting of the employed data clusters to the case at hand, the weightings made, and, if necessary, the dimensions of personal life, which, to date, are not considered by the algorithm.[15] Thus, it empowers the physician and the patient to either reject or undertake attempts to understand, that is, to integrate the ML-generated information or recommendation into their existing orders of knowledge and experience (or to reject it) and, one could say, to align it with the patient's circumstances, interests, values, and goal perspectives. Admittedly, the physician's interpretative approach to the individual case is prone to error and sometimes a highly demanding capacity, but it enables the physician to estimate the existing statistical probabilities and risks for the patient and to deliberate them with the patient for a moral assessment.

ML-generated information which cannot or can barely be integrated into this order due to its lack of transparency can therefore be problematic: if the coming to existence of an information is inaccessible, then it is difficult to form a conviction or a belief based on it, particularly one that enables us to take responsibility for the decisions and actions based on this information. This is especially problematic if the ML-generated information contradicts our individual knowledge and experience and therefore causes plausible integration of the opaque information to seem

[14] Of course, theoretically, scenarios are conceivable in which at least large parts of the numerous relevant factors could be recorded and operationalized by means of a catalog of questions presumably comprising several thousand operators, in order to subsequently feed them into the algorithm and thus determine one or more individual ultimate goal(s) for their best possible achievement. With regard to practical feasibility, however, numerous questions remain unanswered.

[15] For Example, Krishnan (2020) has pointed out that the term "interpretability" often masks only other ends pursued, such as justification or non-discrimination.

impossible. In medical practice, disagreements about diagnostic findings as well as outcome evaluations and predictions are related not only to interpretative deficiencies "but also to an intrinsic ambiguity in the observed phenomena" (Cabitza et al., 2017; cf. Spreckelsen & Spitzer, 2008: 153).[16] However, that to which one has no access cannot be interpreted or embedded into one's epistemic order and therefore cannot be epistemically and normatively evaluated for oneself, much less conveyed to another person for evaluation. Nevertheless, these interpretative and evaluative aspects represent central facets of the doctor-patient relationship and its division of responsibility.

Of course, healthcare decisions are markedly diverse. An extensive differentiation cannot and should not be made here. Many clinical decisions are rightly made without considering the many dimensions of the patient's personal life and the diversity of individual goals. Other decisions, however, cannot avoid such consideration. Abstractly, one could say that the greater the scope of a medical decision for a patient, the more normatively decisive the patient's insight into the factors relevant to this decision and their interpretation in the context of the patient's personal life. This raises questions like the following: For what kind of medical decisions is insight into ML-generated outcomes by the physician or the patient necessary for a sufficient assumption of responsibility? Are there exceptions to this?

As I mentioned previously, both positions regarding the accuracy and interpretability of ML-based systems seem intuitively plausible. The examples utilized on both sides illustrate that there are healthcare decisions that require in-depth insight into their generation as well as those that do not require any real insight. As London (2019) posits, we are content with the demonstrated accuracy of, say, preparations such as aspirin or lithium, even if we have at times lacked or still lack explanations or certainties about their functioning (London, 2019). However, the depth of intervention and the scope of the decision to employ a medical preparation (once) are, excluding the most important contraindications, exceedingly manageable and can actually be compensated to a large extent afterward. The situation is different with normatively weightier therapeutic decisions, such as admission and treatment with agents that severely restrict or potentially endanger one's way of life, especially when these may be necessary for extended periods. Other examples include decisions regarding whether a patient is capable of giving consent or decisions on the continuation of life-sustaining measures, as these encompass numerous dimensions beyond empirically ascertained findings and a patient's operationalizable characteristics.

[16] Grote and Berens (2020) refer to possible cases of "peer-disagreement," according to which equally competent clinicians can nevertheless come to different—but possibly equally plausible—conclusions in one and the same case. Transferring this phenomenon to a "peer-disagreement" between a physician and an ML-based system they illustrate: "After assessing the evidence, she concludes that the patient has disease x, where she has a confidence of 0.8 in her proposition. However, when the machine learning algorithm screens the evidence, it states that the patient has disease y, with a similar degree of confidence. Now, when trying to make a well-informed decision, how much weight should the clinician assign to the algorithm's diagnosis? […] There is very little that the clinician might do on epistemic grounds to resolve the disagreement in question.".

With this diversity of healthcare decisions, it seems appropriate to identify for possible non-interpretable ML-support those decisions that depend solely or largely on statistically based, normatively uncontroversial information (especially in diagnostics or low-risk and reversible treatment decisions). This is likely the case, for example, with image-based diagnostic procedures, such as the detection of skin cancer (e.g., Esteva et al., 2019) or the analysis of radiological findings (e.g., Houssami et al., 2017). Even in such diagnostic decisions, clues to explanations would be helpful for advances in medical practice and theory, but they would be of less normative significance. However, the more normative implications involved in a decision (for example, about its scope, the risks and opportunities contained, the potential alternatives, reversibility), the more likely its generation and thus its interpretability will gain importance.[17] Robbins (2019) summarizes the aim of explicability:

> A principle of explicability, then, is a *moral* principle that should help bring us closer to acceptable uses of algorithms. This provides a human with the information they need in order to exercise that control. Explicability, therefore, is an attempt to maintain meaningful human control over algorithms. Only human beings can be held morally accountable so it should be human beings that are in control over these decisions.

Medical practice regards the treatment of a patient's life based on objective and subjective values, quality of life, and sometimes diffuse and ambiguous clinical phenomena that one encounters. The accuracy or reliability of ML-generated information and recommendations may be sufficient in a bundling of the objective facts from a statistical perspective; however, it cannot provide sufficient justification for decisions from a moral perspective, which is necessary to "maintain meaningful human control" (Robbins, 2019) over the decision to be made. This justification requires explanations beyond the description of the ML-generated outcome.

4 Normative and Communicative Challenges for the ML-Supported Doctor-Patient Relationship

I have offered a point of reference with my proposed formal measure regarding the necessary insight into ML-generated information, according to which interpretability becomes the more significant the more normative implications are involved in the decision to be made; however, its concretion quickly indicates central difficulties because the extent and meaning of the normative implications of even the most

[17] Robbins (2019) puts it very strictly here: "Knowing that a specific decision requires an explanation […] gives us good reason not to use opaque AI (e. g. machine learning) for that decision. Any decision requiring an explanation should not be made by machine learning (ML) algorithms. Automation is still an option; however, this should be restricted to the old-fashioned kind of automation whereby the considerations are hard-coded into the algorithm." In my view, a graded conception of the normative necessity of explicability/interpretability seems more appropriate, since, for example, circumstances of an emergency or situations of no alternative may justify the use of opaque ML-generated recommendations. However, this does not contradict Robbin's position but rather renders it more precise.

basal concepts of medicine (e.g., health, disease, and quality of life) are highly contested (cf. e.g., Seidlein & Salloch, 2019).

Our discursive practice between physicians and patients allows us to obtain an impression of the meaning of terms, to determine normative implications at different levels of shared decision-making and thus, in an ideal–typical manner, to sharpen the physician's picture of the patient's interests, values, and goals.[18] This enriched picture subsequently contributes to the evaluation, hierarchization, and selection of possible evidence-based diagnostic and therapeutic alternatives. Consequently, decisions can be made that connect different kinds of medical information with evaluative judgments. Furthermore, in this interaction, the physician learns about the patient's epistemic requirements for decision-making. While one patient wants the completest understanding possible, another finds it sufficient to ensure that the physician represents an epistemically trustworthy authority who can potentially be questioned and who is accountable for recommendations and decisions.[19]

This complex interaction, in which each decision to be made presents different justification requirements, does not preclude ML-based recommendations, even opaque ones. The challenge of an ML-supported doctor-patient relationship now consists in the identification of precisely those decisions that largely do not require normative justification, such as identifying the most appropriate surgical access (cf. Zhou et al., 2020) after deciding to operate. Such decisions could be supported analogously to evidence-based guidelines—in compliance with defined duties of care and quality standards in addition to the exclusion of undesirable biases by the IT developers—which bundle the current empirical state of knowledge and presort it according to criteria suitable for the patient (age, sex, or preexisting conditions). The treating physician could then utilize these outcomes considering the other non-operationalizable factors. For those decisions that require a higher normative justification due to their scope and depth of intervention, forms of interpretability and insight into the genesis of the information or recommendation are important for implementing them in the discursive practice of physicians and patients.

Of particular epistemic and normative explosiveness could be those situations in which the result of an ML-based system contradicts those obtained conventionally—via established instruments and on the basis of existing medical knowledge—and (possibly due to the lack insight) cannot be validated. How likely such situations are remains questionable. Nevertheless, they represent a possibility to be considered, which presents physicians with a normative challenge: they may lack the tools or the medical knowledge to either verify or falsify the result produced by the ML system.

[18] In this sense, it is correct that statistics on disease progressions, average survival times, recoveries, and deaths cannot provide such an enriched picture of a patient. To what extent other aspects mentioned could be operationalized at all—and thus sufficiently taken into account in the future—I am not able to judge.

[19] In this respect, trustworthiness will (at least potentially) be demonstrated in the deliberative process with the person of trust. The epistemic expertise of the physician may be largely inaccessible to the layperson in terms of content, but it is at least communicatively interrogable in terms of its general criteria of rationality (evidence, adequacy, consistency, deliberation) and its normative implications. See, e.g., the discussion on epistemic authorities by Goldman, 2001 and Martini, 2020.

Such a conceivable constellation also offers the potential for treatment improvement, since the physician inevitably must reexamine and reevaluate the measure that has been favored to date to exclude possible errors (Grote & Berens, 2020). However, if the physician still concludes that the measure recommended continues to deviate from that one of the ML-based system, then this is some kind of a stalemate situation in which only epistemically different justification strategies can be invoked for resolving the divergent results. Here it cannot be determined whether and to what extent such situations should be circumvented to avoid overburdening the patient with decisions about the preferred "path of knowledge" or to prevent problems of taking responsibility for such an ML-generated outcome.[20] However, even for such a case, it becomes evident that the embedding of ML-based information into the deliberative and communicative decision-making process would clearly benefit from a maximum possible insight into the genesis and thus interpretability of this outcome. Therefore, once again, the more normatively far-reaching and important the clinical decision (and thus the higher the justification requirements), the more significant the traceable and interpretable processes of the ML-based generation of information and recommendations.

The requirements for justification of recommendations or decision-making, though perhaps only rudimentary or even sometimes retrospectively proven wrong, do not constitute just some kind of "opium for the conscience" but form the normative basis of interactions between people, especially in the necessarily trust-based doctor-patient relationship.

5 Conclusion

ML-based systems and human agents differ in their assets in clinical decision-making. The former, thanks to their processing capacities, are potentially capable of data-based synthesis performances that surpass a human many times over. The latter, in turn, are potentially capable of a discernment or integration performance by considering the practically relevant aspects beyond operationalizable data (social-relational, psychological, moral, and religious factors) and by perceiving and processing uncertainties and ambiguities that will remain inaccessible to ML-based systems for the foreseeable future. Both ML-based systems and human agents have different error-proneness and weaknesses that the other seems to be able to improve. The transparency and interpretability of ML-generated information can enable human agents—here, physicians and patients—to identify, circumvent, or at least adequately exploit some of the system's error-proneness and weaknesses in clinical decisions. However, unlike a highly accurate or reliable but opaque ML system, interpretability allows understanding of the inherent normative implications during the genesis of ML-based information and to accept, modify, or reject them.

[20] In any case, it seems that the physician cannot be held accountable here, since qua her own knowledge and experience, she comes to a conclusion and epistemic *conviction* that contradicts that one of the ML-based system.

The task for the future in the development and implementation of ML-based systems is therefore to identify that equilibrium in which the skills of physicians and ML-based systems optimally complement each other. To achieve this, ML developers and physicians must engage in close collaborations. On the one hand, regulatory quality standards and performance criteria to evaluate the achievement of medical benefits and compliance with other relevant aspects (privacy, liability, etc.) must be determined; on the other hand, it is necessary to seek and formulate precise implementation opportunities in clinical practice. Only by considering the concrete potential field of application, *those* aspects and goals of the clinical decision can be identified whose processing must, due to their normative implications, remain comprehensible, assessable, and communicable for the decision-makers. As much as the implementation of ML-based systems may facilitate some tasks of everyday medical practice, it makes the interaction between patients, physicians, and algorithms an even more challenging task. This requires physicians to heighten their sensitivity and increase their skills in addressing the social, communicative, and ethical aspects and issues of their medical practice. And to apply these skills in dealing with ML systems requires extensive training. In clinical decision-making, this will allow both physician and patient to assess, avoid, or consciously accept possible risks in a responsible manner.

Author Contribution Single author, all own work.

Funding Open Access funding enabled and organized by Projekt DEAL.

Availability of Data and Material Not applicable.

Code Availability Not applicable.

Declarations

Competing Interests The author declares no competing interests.

References

Ahuja, A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ, 7*, e7702. https://doi.org/10.7717/peerj.7702

Baggio, G., Corsini, A., Floreani, A., Giannini, S., & Zagonel, V. (2013). Gender medicine: A task for the third millennium. *Clinical Chemistry and Laboratory Medicine, 51*(4), 713–727. https://doi.org/10.1515/cclm-2012-0849

Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology, 34*, 349–371. https://doi.org/10.1007/s13347-019-00391-6

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA, 318*(6), 517–518. https://doi.org/10.1001/jama.2017.7797

Cartwright, N. (2007a). Are RCTs the gold standard? *BioSocieties, 2*(2), 11–20. https://doi.org/10.1017/S1745855207005029

Cartwright, N. (2007b). *Evidence-based policy: Where is our theory of evidence?* Center for Philosophy of Natural and Social Science, London School of Economics, Technical Report 07/07.

Chakravartty, A. (2017). Scientific Realism. The Stanford Encyclopedia of Philosophy (Summer 2017 Edition), Retrieved January 6, 2022, from https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine, 24*(9), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

de Regt, H. W., Leonelli, S., & Eigner, K. (Eds.). (2009). *Scientific Understanding: Philosophical Perspectives*. University of Pittsburgh Press.

Densen, P. (2011). Challenges and opportunities facing medical education. *Transactions of the American Clinical and Climatological Association, 122*, 48–58.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics, 47*, 329–335. https://doi.org/10.1136/medethics-2020-106820

Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*, 24–29. https://doi.org/10.1038/s41591-018-0316-z

Floridi, L., Cowls, J., Beltrametti, M., Chatile, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds & Machines, 28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Genin, K., Grote, T. (2021). Randomized controlled trials in medical AI. A methodological critique. *Philosophiy of Medicine* 2, 1–15. https://doi.org/10.5195/POM.2021.27.

Goldman, A. I. (2001). Experts: Which ones should you trust? *Philosophy and Phenomenological Research, 63*, 85–110.

Grimm, Stephen R. (2005). Understanding as an epistemic goal, Dissertation (University of Notre Dame).

Grimm, S. (2011). "Understanding". In *The Routledge Companion to Epistemology*. Edited by S. Berneker D. Pritchard, 84–94. New York: Routledge, 2011.

Grimm, S. (Ed.). (2017). *Making Sense of the World*. Oxford University Press.

Grimm, S., Baumberger, C., & Ammon, S. (Eds.). (2017). *Explaining understanding: New perspectives from epistemology and philosophy of science*. Routledge.

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

Hardin, C. L., & Rosenberg, A. (1982). In Defence of Convergent Realism. *Philosophy of Science, 49*(4), 604–615. https://doi.org/10.1086/289080

Heinrichs, B., & Eickhoff, S. B. (2020). Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping, 41*, 1435–1444. https://doi.org/10.1002/hbm.24886

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11*, 428–434. https://doi.org/10.1016/j.tics.2007.09.004

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: The system causability score (SCS). *KI – Künstliche Intelligenz,* 34, 193–198. https://doi.org/10.1007/s13218-020-00636-z.

Houssami, N., Lee, C. I., Buist, D. S. M., & Tao, D. (2017). Artificial intelligence for breast cancer screening: Opportunity or hype? The Breast, 36, 31–33.https://doi.org/10.1016/j.breast.2017.09.003.

Hutson, M. (2021). Lyin' AIs: The opacity of artificial intelligence makes it hard to tell when decision-making is biased. IEEE Spectrum, 58(2), 40–45. https://doi.org/10.1109/MSPEC.2021.9340114

Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E., & Dudley, J. T. (2018). Artificial intelligence in cardiology. *Journal of the American College of Cardiology, 71*(23), 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521

Krishnan, M. (2020). Against interpretability: A Critical examination of the interpretability problem in machine learning. *Philosophy & Technology, 33*, 487–502. https://doi.org/10.1007/s13347-019-00372-9

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology, 69*(21), 2657–2664. https://doi.org/10.1016/j.jacc.2017.03.571

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet – Digital Health*, 1(6), E271–E297. https://doi.org/10.1016/S2589-7500(19)30123-2.

London, A. J. (2019). Artificial intelligence and black-box. Medical decisions: Accuracy versus explainability. *Hastings Center Report*, 49(1), 15–21. https://doi.org/10.1002/hast.973.

Martini, C. (2020). The Epistemology of Expertise. In M. Fricker, P. J. Graham, D. Henderson, & N. J. L. L. Pedersen (Eds.), *The Routledge Handbook of Social Epistemology* (pp. 115–122). Routledge.

Molnar, C. (2021). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Retrieved August 20, 2021, from https://christophm.github.io/interpretable-ml-book/.

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., & Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ, 368*, m689. https://doi.org/10.1136/bmj.m689

Orwat, C. (2019). *Studie Diskriminierungsrisiken durch Verwendung von Algorithmen*. Retrieved April 11, 2021, from: https://www.antidiskriminierungsstelle.de/SharedDocs/Downloads/DE/publikationen/Expertisen/Studie_Diskriminierungsrisiken_durch_Verwendung_von_Algorithmen.html.

Patel, S., Wang, J. V., Motaparthi, K., & Lee, J. B. (2021). Artificial intelligence in dermatology for the clinician. *Clinics in Dermatology*. In Press. https://doi.org/10.1016/j.clindermatol.2021.03.012.

Pearson, D. (2017). *Artificial intelligence in radiology: the game-changer on everyone's mind. Radiology business*. Retrieved April 11, 2021, from: https://www.radiologybusiness.com/topics/technology-management/artificial-intelligence-radiology-game-changer-everyones-mind

Pritchard, D. (2009). *Knowledge*. Palgrave Macmillan.

Putnam, H. (1982). Three Kinds of Scientific Realism. *Philosophical Quarterly, 32*(128), 195–200. https://doi.org/10.2307/2219323

Rawlins, M. (2008). De testimonio: On the evidence for decisions about the use of therapeutic interventions. *Lancet, 372*(9656), 2152–2161. https://doi.org/10.1016/S0140-6736(08)61930-3

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines, 29*, 495–514. https://doi.org/10.1007/s11023-019-09509-3

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2). https://doi.org/10.1162/99608f92.5a8a3a3d.

Salto-Tellez, M., Maxwell, P., & Hamilton, P. W. (2018). Artificial intelligence – The third revolution in pathology. *Histopathology*. https://doi.org/10.1111/his.13760

Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B. S., Waldstein, S. M., & Bogunović, H. (2018). Artificial intelligence in retina. *Progress in Retinal and Eye Research, 67*, 1–29. https://doi.org/10.1016/j.preteyeres.2018.07.004

Sedmak, C. (2003). *Erkennen und Verstehen. Grundkurs Erkenntnistheorie und Hermeneutik*. Tyrolia Innsbruck.

Seidlein, A. H., & Salloch, S. (2019). Illness and disease: An empirical-ethical viewpoint. *BMC Medical Ethics, 20*(1), 5. https://doi.org/10.1186/s12910-018-0341-y

Sim, I., Gorman, P., Greenes, R. A., Haynes, R. B., Kaplan, B., Lehmann, H., & Tang, P. C. (2001). Clinical Decision Support Systems for the Practice of Evidence-based Medicine. Journal of the American Medical Informatics Association, 8, 527–534. https://doi.org/ 10.1136/jamia.2001.0080527

Smith, P. (1998). Approximate truth and dynamical theories. *British Journal for the Philosophy of Science, 49*(2), 253–277. https://doi.org/10.1093/bjps/49.2.253

Smith, H. (2021). Clinical AI: Opacity, accountability, responsibility and liability. *AI & Society*. https://doi.org/10.1007/s00146-020-01019-6

Solomon, M. (2015). *Making Medical Knowledge*. Oxford University Press.

Spreckelsen, C., & Spitzer, K. (2008). *Wissensbasen und Expertensysteme in der Medizin. KI-Ansätze zwischen klinischer Entscheidungsunterstützung und medizinischem Wissensmanagement*. Medizinische Informatik. Vieweg + Teubner.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The ethics of algorithms: Key problems and solutions. *AI & Society*. https://doi.org/10.1007/s00146-021-01154-8

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7

Visani, G., Bagli, E., & Chesani, F. (2020). OptiLIME: Optimized LIME explanations for diagnostic computer algorithms. *Proceedings of ACM Conference '17*. ACM New York.

Worrall, J. (2007). Evidence in medicine and evidence-based medicine. *Philosophy Compass, 2*(6), 981–1022. https://doi.org/10.1111/j.1747-9991.2007.00106.x

Zagzebski, L. (2009). *On Epistemology*. Wadsworth.

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology, 34,* 265–288. https://doi.org/10.1007/s13347-019-00382-7

Zhou, X.-Y., Guo, Y., Shen, M., & Yang, G.-Z. (2020). Application of artificial intelligence in surgery. Frontiers in Medicine, 14, 417–430. https://doi.org/10.1007/s11684-020-0770-0.