Philosophy & Technology (2022) 35: 50
https://doi.org/10.1007/s13347-022-00546-y

RESEARCH ARTICLE

On the Ethical and Epistemological Utility of Explicable AI in Medicine

Christian Herzog[1]

Received: 21 December 2021 / Accepted: 18 May 2022 / Published online: 30 May 2022

Abstract
In this article, I will argue in favor of both the ethical and epistemological utility of explanations in artificial intelligence (AI)-based medical technology. I will build on the notion of "explicability" due to Floridi, which considers both the intelligibility and accountability of AI systems to be important for truly delivering AI-powered services that strengthen autonomy, beneficence, and fairness. I maintain that explicable algorithms do, in fact, strengthen these ethical principles in medicine, e.g., in terms of direct patient–physician contact, as well as on a longer-term epistemological level by facilitating scientific progress that is informed through practice. With this article, I will therefore attempt to counter arguments against demands for explicable AI in medicine that are based on a notion of "whatever heals is right." I will elucidate my elaboration on the positive aspects of explicable AI in medicine as well as by pointing out risks of non-explicable AI.

Keywords Artificial intelligence · Explicable AI · Explainable AI · Evidence-based medicine · Patient compliance · Good health

1 Introduction

The vision of computer-aided decision-making in medicine has a tradition of at least 60 years (Shortliffe & Sepúlveda 2018). Increasingly capable algorithms, computational resources, and increases in available data have let the vision resurface. Artificial intelligence (AI) and deep learning algorithms in particular have shown potential to surpass human-level classification accuracy (Topol 2019), but widespread clinical adoption is still hampered by a number of reasons, one of which being that black-box algorithms are widely deemed unacceptable, e.g., (Shortliffe & Sepúlveda 2018), since they are believed to be incapable of providing information to the user that supports accountable decision-making. However, considering the promise

✉ Christian Herzog
christian.herzog@uni-luebeck.de

[1] Ethical Innovation Hub, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany

that black-box algorithms may perform even better, the demand for making them explainable has been challenged, e.g., (Durán & Jongsma 2021; London 2019).

In this article, I will focus on the rather rich notion of explicability due to (Floridi et al. 2018) as a means to disentangle notions of interpretability, transparency, intelligibility or, simply, explainability. Even though the status of "explicability" as an ethical principle within the ethics guidelines of the EU's High-Level Expert Group on Artificial Intelligence (Independent High-Level Expert Group on Artificial Intelligence Set Up By the European Commission (2019)) may be debatable (Robbins 2019), the notion brings to the fore an important consideration for the combination of both intelligibility and accountability. In these terms, the notion of "explicability" can be understood to demand explanations that can be understood and utilized in practice and, hence, support responsible use. I take it that such demands should not only account for various stakeholders, but also for various time frames. As such, there might be immediate benefits from allowing the use of non-explicable AI algorithms in the present, which, however, may diminish long-term benefits such as epistemological gain and, ultimately, improved health care in the future. Not unlike arguments put forward in favor of strong measures for mitigating climate change justified by the responsibility for future generations, see, e.g., (Desjardins 2006), it is my desire to advocate for a wide-arching consideration of short- to long-term effects within the debate about whether or not AI should be explicable in the medical context. To this end, the present paper can be regarded as contributing also to the growing literature dedicated to the risk and potential of AI to be opposed or contribute to a positive and sustainable development, see, e.g., (Holzinger et al. 2021; Sætra 2021). As upcoming AI regulation in the EU is presumably based on risk stratifications (Floridi 2021), a balanced and all-embracing consideration of risks is in order.

However, discussions surrounding the ethical implications of so-called black-box AI algorithms have revolved mostly around arguments that reside in the present situation or near-term future. Be it in terms of the way medical practice unfolds in direct physician to patient and physician to physician encounters, forensics and accountability issues or in terms of a general notion of responsible medical decision-making, there seems to be a lack of consideration for long-term ethical and epistemological risks. London (2019), e.g., vehemently argued that insisting on medical AI being explainable would risk applying poorer treatments. According to London, a preference for simpler, interpretable models without assurances of additional benefits to the patient would even constitute a "lethal prejudice." Admittedly, such a line of thinking is hard to argue against. Similarly, Bjerring and Busch (2021, p. 351) postulate that reports on the success of (predominantly deep-learning-based) AI in medicine[1] suggest that practitioners will very likely have an "*epistemic obligation* to rely on these systems".

Certainly, I am not proposing to deliberately withhold available and effective treatments from patients in need, just because they stem from non-explicable AI systems. Rather, my agenda is to argue that just because non-explicable AI can help or

[1] See, e.g., (Liu et al. 2019; Loh 2018; Topol 2019) for reviews on advances of AI in medicine and how their performances compares to professional healthcare practitioners.

even heal now, we should not lose sight of what may support progress in the future as well as what will constitute sustainable efforts towards maintaining health also in terms of the entire breadth of individual conceptions of it.[2] As will become clear in the sequel, for this purpose, I am rejecting a reductionist notion of what constitutes progress in medicine, e.g., one that is simply based on indicators such as longevity. Instead, if AI is to be truly a tool for enhancing human autonomy, it should be malleable enough to account for different notions of beneficence, and hence, more individual accounts of good health. To make this more concrete, I will argue in favor of AI systems, e.g., in primary care, or—more generally—narrative-based medicine, that can take into account individualistic notions of good health or life (even with diseases), supporting shared decision-making between physician, patient and AI. To some extent, this issue has already been raised by others. Bjerring and Busch (2021), e.g., argue that black-box AI in medicine is incompatible with a patient-centered approach, for which they cite Epstein et al. who delineate patient-centric medicine as aiming at establishing "a state of shared information, shared deliberation, and shared mind" (Epstein et al. 2010, p. 1491). McDougall (2019) discusses both opportunities as well as risks for shared decision-making in the face of medical AI-based recommendation systems. In light of this, I would like to expand on this argument by outlining the ethical and epistemological utility of either augmenting black-box AI in medicine by means of explanatory interfaces, or by relying on inherently interpretable AI in the first place. With respect to the latter, however, I have argued elsewhere (Herzog 2021) that using inherently interpretable AI does not justify to forgo the devision of explanatory interfaces for meeting the requirements of responsible and accountable use in practice.

By more clearly articulating the utility of explicable AI in medicine,[3] I hope to contribute to and enable further discussion about black-box AI in medicine in a differentiated manner: As the practical medical utility of AI—and black-box approaches in particular—seems to progress at an increasing pace, strong arguments are needed to determine when it is justifiable to proceed with the deployment and when there are valid grounds for pause. For instance, Di Nucci (2019) highlights the general potential of machine learning for personalizing treatment options, which should not be underestimated. However, the way corresponding AI systems are designed can influence potential biases skewing recommendations towards specific treatments, or even result in a systematic exclusion of alternatives, which may be highly dependent on the medical or care issue at hand and the specific situation of the patient itself. Hence, the assessment of what kinds of explanation are required and whether they can even be dispensed with is likely to vary between the fields of application and subdisciplines in medicine and care. This is why a comprehensive discussion is beyond the scope of this paper. Rather, I can only hope to contribute to the discussion by highlighting the particular epistemological and ethical utility that I focus on in the sequel.

[2] For a connection between the United Nation's Sustainable Development Goal associated with "Good Health and Well-Being" please see, e.g., (Muller et al. 2021).

[3] I will try to more thoroughly motivate the choice of term "explicable" in Sect. 3.

Hence, I will try to provide a nuanced perspective on the utility of explicable AI in medicine. For this purpose, I will establish epistemological—and ultimately, ethical—arguments in favor of AI whose outputs and models can be explained in some sense in Sect. 5. I will argue that by relinquishing a preference for explainable AI in medical practice, we may be sabotaging important cross-fertilization between theory and empiricism informed by practitioners and patients. This appears plausible in light of the observations that have spawned initiatives for demanding a "learning health care system" (Olsen et al. 2007) that integrates practice and research. Secondly, in terms of being supportive of "good health",[4] both mechanistic or correlative explanations may constitute an essential tool in improving patient compliance or increasing patient autonomy by allowing more individual decision-making and lifestyle-compatible interventions. Before presenting an ethical and epistemological viewpoint in favor of explicable AI in medicine, I will first delineate the basic terminology of "explainable," "interpretable," "intelligible," "transparent," and—ultimately—"explicable" AI in Sect. 3. I will then summarize the case against a preference for explicable AI in medicine and present counterarguments in Sect. 4.

2 Explainable, Interpretable, Intelligible, Transparent and Explicable AI

Different notions of rendering opaque AI approaches understandable in some sense go by the terms "explainable," "interpretable," "intelligible," and "transparent," to name the dominant ones. Meanings often differ slightly, are sometimes even used interchangeably or have been adopted habitually by referring to a common sense of language. However, for the purposes of this paper, it is necessary to first establish a clear distinction of terminology. A broad distinction can be made regarding the notion of "transparency" and the notions of "interpretability," "explainability," and "intelligibility."

"Transparency" typically refers to laying open certain aspects (ideally all) of an algorithm or piece of software to a specific audience. This may incorporate training routines, data sets, neural network designs, or even less technical aspects such as the composition of the development team, stakeholders, funding, etc. Clearly, typical addressees of "transparency" measures are neither those that use, or are affected by, the algorithms in question. Rather "transparency" typically addresses groups such as auditors, forensics, or other entities that should provide oversight and safeguard contestability. Even though "transparency" has been sometimes advocated as a means to enhance accountability (Weller 2019), concerns have been raised that it is neither desirable for privacy, security, and economic reasons as well as that it does not contribute much in terms of accountability as complex algorithms are often inherently opaque (de Laat 2018).

[4] I choose this term to specifically highlight the possibility to render healthcare practice amenable to a patient's individual conception of both good health and good healthcare.

It is often stated that it is this opacity that should be tackled by some means of endowing algorithms with "explainability," cf. (Robbins 2019). Even though different authors use different definitions and even often use "explainability" and "interpretability" exchangably, for the purposes of this paper let "explainability" refer to means to "summarize the reasons […] for behavior […] or produce insights about the causes of […] decisions," whereas "interpretability" refers to descriptions of "internals of a system in a way which is understandable to humans" (Gilpin et al. 2019). Hence, while interpretable AI models may utilize reasonably understandable mathematical models as their basis, explainable AI models often rather offer post hoc visualizations that aim to shed light on the behavior of models built from a vast number of variables. In the sense of these meanings, authors such as Rudin (2019) ardently advocate the use of interpretable models in high stakes domains, such as medicine, since interpretable models encode a priori knowledge about the causal or correlative connections that can be verified and validated before being put to use, citing—among others—drawbacks of post hoc means of explanation. Others, such as Krishnan (2020), claim that knowledge about the "inner workings" of an algorithm has limited power to contribute to issues such as fostering trust, avoiding biases, or prevent other kinds of system failure. I concur, and it will become clear in the sequel that, what may be important is to devise context-specific means of allowing for relaying information to human agents for enhancing responsible use.

In this vein, the notion "intelligibility" is often regarded as a generic term, encompassing the possibility of referring to "interpretability" or "explainability" (or both), cf. (Marcinkevičs and Vogt 2020; Weld and Bansal 2018), Weld and Bansal (2018) acknowledge a necessity of interdisciplinary cooperation when trying to achieve "intelligibility" by stating that the "key challenge for designing intelligible AI is communicating a complex computational process to a human".

As stated above, the notion of "explicability" combines the desiderata to effectively communicate information to human agents and to do so in a manner that allows accountable use. In acting as a term that implies both the epistemological demand for intelligibility as well as the ethical demand for accountability (Floridi et al. 2018), the notion of "explicability" transcends specifics of a technical realization while emphasizing responsible use in its broadest sense. The notion of "explicability" hence remains malleable with respect to the stringency of the means to alleviate an algorithm's opacity, while staying firmly committed to demanding accountable human agency.

Elsewhere, I have defended the status of "explicability" as an ethical principle against objections form Robbins (2019). For the purposes of this paper, I will only briefly recapitulate the main points. In fact, Floridi et al. stop short of making precise demands regarding the extent of explanations of the inner workings of an AI algorithm that would constitute intelligible and accountable systems. As a normative claim that is independent of the application domain, this perhaps seems wise. However, for the medical context that I am concerned with, it is instructive to briefly discuss what amounts to explanations that could be denoted intelligible and that support accountability. As I will discuss in the sequel, medical evidence is often not supported by mechanistic explanations, but may also be based on correlative data (London 2019). However, even if that is the case, there may nevertheless be "some

useful information that a practitioner can pass on to patients to encourage rational and informed deliberation" (Bjerring and Busch 2021, p. 364). Simply speaking, the least that an explicable medical AI system should do (if nothing else is available) is to provide information about the statistical correlations that are being invoked to support a certain recommendation. This would give both the medical professional and the patient some grounds to arrive at an informed decision about whether to concur with or reject the AI-based recommendation. If situated in a shared decision-making setting, such kinds of information would also provide grounds for exchange and developing a common understanding, cf. Section 4.2. Of course, it is not my intention to argue for the least amount of information any explanation must convey. Rather with this illustration, I would like to stress that the normative epistemological and ethical demand for intelligibility and accountability translate to a human-centric approach, in that concrete demands in intelligibility and accountability must be assessed based on the explanandum, recipient, and overall context. Consequently, in proposing the normative principle of explicability, Floridi et al. are stopping short of demanding that an AI algorithm may be explainable in every single detail.

While in this paper, I will mostly consider the accountability of the medical practitioner and—in a shared decision-making setting—of the corresponding patient. However, the principle of explicability invites to broaden the scope towards the entire socio-technical system involved in conceiving, designing, developing, marketing, deploying, maintaining, certifying, utilizing, and scrutinizing a specific AI system. In this sense, Floridi et al.'s principle of explicability is different from demanding that AI systems be fully explainable and, hence, non-opaque, but rather requires its accountable use, where all relevant actors are both sufficiently able and comfortable to acknowledge and take responsibility (Kiener 2021). In the medical domain, this would mean that even non-black box systems (e.g., expert systems), which are interpretable in principle, would require explanatory interfaces due to the stakes and demands on privacy involved in concrete medical decision-making. Patients must mostly be able and required to trust and confide in the treating medical team only. Any referral to outside bodies is ethically questionable, if a patient finds her or himself in a vulnerable position with the necessity to reach immediate and highly personal decisions for which he or she should share accountability only with the treating medical practitioners.

The characterization of explicable AI as solutions that actually yield intelligible explanatory interfaces that enable accountable use may not be precise. However, as stated above, such epistemological and ethical demands need to be malleable to the context at hand. This is similar to demanding that AI solutions should respect human autonomy. Consequently, the principle of explicability needs to be understood on a different normative level than explainable, interpretable, or transparent AI. These latter three are mostly referring to providing of some kind of explanation (usually referring to an algorithm's mechanics rather than to those of the inference domain), resorting to simpler pre-hypothesized mathematical models, or giving insight into the AI development process and result. Compared to that explicability is considerably less concerned with the technical realization, but with the epistemological and ethical impact.

However, explicability's status as an epistemological or ethical principle is disputed. For instance, Robbins (2019) argues against letting "explicability" constitute an ethical principle, even though he does not dismiss explicability as useless. Furthermore, Robbins claims that the "epistemic value of explicability is not under dispute" (Robbins 2019, p. 501). We will later see that there is dispute about the epistemic value—at least when in trade-off against other beneficial values—but I will remain on its ethical value here. In order to dispute the ethical value of explicable AI, Robbins argues that (i) explicability should attach to an decision rather than the corresponding entity making it, (ii) the need for explicability is context-dependent, and (iii) explicability should not restrict AI to decision support systems in which possible explanations are deemed acceptable a priori. I maintain that commenting on these arguments is insightful to how explicability should be meaningfully conceived. Hence, with respect to (i) I do not believe that there is a fundamental difference in demanding single morally relevant AI-made decisions to be explained or demanding that an AI should be generally capable of providing explanations at least for the morally relevant ones. In fact, I believe that this distinction between high-risk decisions that require explanations and low-risk decisions that do not require one is implicit in the framing of explicability as intelligibility that supports accountability. A high degree of accountability is only necessary for the morally relevant decisions. I maintain that demanding AI to be explicable is tacitly sensitive to a potential risk stratification, e.g., similar to the one currently proposed in the European draft AI regulation (Floridi 2021). Hence, the principle of explicability may readily be attached to AI in general. The same argument holds with respect to refuting (or rather acknowledging) (ii). Low-risk purposes do not require a strong support of who is accountable. However, in high-risk settings, such as medicine, it should be in either the physicians' or the medical device providers' own interest to be able to take responsibility and also know when and why one's own responsibility needs to be invoked. Low-risk settings—as the designation readily foretells—still incur risks, even if they are minor ones. A simple solution to a potential responsibility gap, cf. (Matthias 2004), could exist in some human actor willingly accepting accountability for the effects of an AI whose decisions or recommendations are not provided alongside intelligible explanations, cf. (Kiener 2021). I maintain that this could be acceptable in low-risk settings, but would currently generate issues in high-risks settings—very likely also legal ones. With respect to (iii), I believe that again explicability's characterization as demanding intelligibility to the point that it supports accountability holds the key. Ethically speaking, it can be argued that we should refrain from using AI if it does not provide us with sufficient information to take responsibility. Of course, we may debate whether we would be willing to relinquish responsibility to some artificial entity. However, we might as well debate whether we should refrain from insisting on respecting human autonomy, which is—of course—absurd. The point is that ethical principles need not necessarily be universally agreed upon. This does not make them any less an ethical principle for those that choose to adopt them—for good reasons, I might add.

In the sense of the discussion above, I therefore propose to conceive of demands for explicable AI as implying a sensible built-in stratification of the intensity of the demand for intelligibility and accountability that is context-dependent and balanced

with regard to other ethical demands. As such, one may argue that explicability as a principle becomes rather vague. It is my take on this that we are challenged to engage in discourse about what is at stake both ethical and in epistemological terms with the purpose to provide clarity to the utility of explicability. Even though legally a five-step risk stratification may be meaningful, both philosophically and scientifically we should strive to differentiate this utility further. This is exactly what I set out to do with this article.

Hence, for the purposes of this paper, it is sufficient to consider the difference between "explainability," "interpretability," and "explicability" as described above. Krishnan (2020) has acknowledged the problem that the notions of "explainability" and "interpretability" (which she uses interchangeably) do not in themselves solve the ethical challenges presented by opaque AI. Unfortunately, Krishnan has also not distinguished terminology precisely enough to reveal the epistemological and ethical demands inherent to the notion of "explicability." Accordingly, the present paper is concerned with the instrumental value of intelligible AI algorithms in medicine to achieve accountability in a broad sense. In order to do so, objections against the instrumental relevance of explanations or interpretative approaches in AI-supported medicine first need to be refuted. This is the objective of the following section.

3 Refuting the Case Against a Preference for Explicable AI in Medicine

One of the most all-encompassing suite of arguments against a preference for explicable AI in medicine is due to (London 2019). London's core argument rests on the idea that calls for explicable AI in medicine are based on a misconception of medicine as a productive science (*techne*) that primarily rests on identifying and invoking causal relationships and principles to bring about an effect. Under the provision that underlying theories can be broken down and explained to non-experts, London concedes that "explanations […] help to foster social trust […], accountability [… and] autonomy" (London 2019, p. 16). However, London argues that knowledge about causal relations in medicine is severely limited and that "decisions that are atheoretic, associationist, and opaque are commonplace" (London 2019, p. 17). Drawing on several examples both from antiquity and the present, London argues that medicine is not a productive science in the sense of a *techne*, as it routinely operates efficaciously despite high (or even complete) uncertainty about the underlying mechanisms. London goes even further by maintaining that an undue emphasis on the demand for explanations leads to situations, where "patients suffer, resources are wasted, and progress is delayed" (London 2019, p. 18).

In addition, Durán and Jongsma (2021) suggest an epistemic definition of opaque algorithms in terms of the impossibility of an algorithm to be "surveyed by humans." This notion of technological opacity is related to one of Burrell's (2016) concepts of opacity, which results from a mismatch between the mathematical procedures of operation characteristics of machine learning systems (i.e., the sheer scale of different parameters required) and human styles of semantic interpretation. Burrell also mentions opacity from intentional secrecy (Pasquale 2015) and technical illiteracy.

Considering even further notions of opacity, such as one that results when a process too complex to be scrutinized from a particular point of view (Herzog 2019) may also be enlightening, especially considering asymmetric levels of knowledge, capabilities and agency as may be the case in medicine. Hence, it may be important to notice that, from the perspective of particular stakeholders, an algorithm may be opaque even though, theoretically, there are ways for experts to scrutinize it. Still—and because transparency is not a viable solution—Durán and Jongsma (2021) are justified in narrowing down their discussion on black-box algorithms that are simply impossible to survey, when considering the question whether black-box algorithms should be allowed in healthcare and medicine or not. In essence, Durán and Jongsma (2021) then argue for a reliability criterion sufficient for accrediting algorithms with trustworthiness, even when the algorithm's inner workings cannot be scrutinized.

In the following, I will introduce two new aspects that strongly suggest that explanatory AI in medicine should not be dismissed: (i) An epistemological view that explicable AI supports critical appraisal in practice and hence maintains an effective feedback loop between medical practice and research and (ii) an ethical view that explicable AI in medicine supports patient compliance and autonomy by facilitating individual notions of good health care. First, however, I will attempt to refute in more detail some of the arguments brought forth by London, Duran, and Kongsma.

3.1 Explainability Risks Deteriorating Accuracy

There is a wide-spread conception that in the development of AI systems, interpretability is often only achieved at the expense of accuracy, see, e.g., (DARPA 2016). As derived from this trade-off, the decision to either go with systems that promote interpretability and, hence, accountability in medical decision-making or choosing a black-box system with the best available "performance" is not an easy one. In his article, London (2019) puts the burden of proof on critics of non-explicable AI. He demands that the risk of deteriorating AI accuracy—and presumably deteriorating efficacy of AI health systems—due to the demand for interpretability should be offset by benefits to patients. These benefits should first be substantiated, before they would warrant a consideration of extra efforts and potentially worse performance due to means for rendering medical AI interpretable or explainable.

Irrespective of the fact that there exist said benefits (albeit less easily quantifiable ones than classification accuracy metrics), which I will elaborate on in the sequel, recent research has called into question the existence of the interpretability-accuracy trade-off itself. (Rudin 2019; Rudin & Radin 2019), for instance, go as far as calling the trade-off "a myth," albeit conceding that "[i]nterpretable models can entail significant effort to construct in terms of both computation and domain expertise" (Rudin 2019, p. 210). Other examples exist, see, e.g., (Caruana et al. 2015), that not only call into question the general existence of the trade-off, but also elucidate how interpretable models can help find flaws in the AI system design that pose potentially dangerous risks to patients. This seems to not only rather make the trade-off merely one between interpretability and the speed in the development of innovations, but

there also seem to be additional risks incurred due to a lack of interpretability, especially in high-risk domains such as health care.

One such risk is the perpetuation of biases, i.e., discriminatory tendencies in the outputs of medical AI, that could be more easily addressed by means of explainability interfaces or inherently interpretable approaches, both of which are allowing for easier debugging and finetuning of an algorithm's fairness (Yoon et al. 2021). Worries are that common accuracy measures cannot account for the minimization of such biases and neither can accuracy necessarily guarantee scientific reproducibility and generizability (Topol 2019).

Accordingly, we may as well assume that designing for interpretability pays off doubly. However, critics may argue that interpretability is an elusive concept, especially in medicine, where mechanistic explanations may not even be available. But most commonly, explainable AI methods are not aspiring to provide insight into the physical (or, in this case, medical) phenomenon, but rather into the algorithm and its way of mapping input data to actionable outputs.

As a brief note, one may also counter that efforts in devising explainable AI will ultimately render the issue of a trade-off between explainability and accuracy obsolete. The day this happens would definitely mark an achievement. However, with regard to the considerable economic benefits of forgoing either the time-consuming efforts to base AI on inherently interpretable models, or devising appropriate explanatory interfaces (and being potentially legally permitted or condoned to do so), make the issue of this paper one that is urgent now. Furthermore—and as will become clear in the sequel—it is mandatory to consider the quality and capability of explanatory interfaces to accommodate practitioners in combining the technical explanation with relevant non-technical factors that should influence decision-making in medicine, such as a patient's preferences, life situation, and compliance supporting/diminishing circumstances, cf. Section 4.2.1, and potentially even a sense of a patient's conception of good health, cf. Section 4.2.2

3.2 A Lack of Clarity in the Goals of Interpretability

(London 2019) asserts that calls for interpretability often suffer from a lack of precise goals it should cater to. In terms of post hoc rationalizations of decisions taken, London, following Lipton (2018), claims that machines and humans are very much interpretable, even though these ex post reflections on reasons are not identical to the reasons responsible for and during the decision. Yet other concepts of interpretability that demand the possibility of step-by-step analyses may make deep learning as well as other sufficiently complex software systems look non-interpretable as well.

I concede that uncritical demands for interpretability may miss the point of what is ethically and legally relevant. While a post hoc step-by-step tracing of the inner mechanics of a decision algorithm is useful for forensic purposes when time is not of the essence, in time-critical situations, AI systems should deliver explanations tailored to the addressee and depend on the context. Consequently, there is a need to not conflate the meanings of interpretability and explicability. Especially in

high-risk situations, however, some form of interpretability will definitely be needed to hold AI system designers accountable. However, this does not necessarily imply that AI algorithms must always rely on models that are interpretable in terms of domain-specific and strictly causal relations.

3.3 The Danger of Causal Interpretations

In relation to the previous section, (London 2019) cautions that in domains that lack causal knowledge, the demand for interpretability is all too easily contributing to misconceiving correlations for causations. Consequently, London questions the value of interpretability, because causal inferences are not to be expected from associationist approaches to AI systems in health care.

In fact, London portraits medical decision-making as often relying "on an associationist model encoded in the neural network in the clinician's head that is opaque and often inaccessible to others", (London 2019, p. 18). He further notes that "[l]arge parts of medical practice frequently reflect a mixture of empirical findings and inherited clinical culture. In these cases, even efficacious recommendations of experts can be atheoretic in this sense: they reflect experience of benefit without enough knowledge of the underlying causal system to explain how the benefits are brought about.", (London 2019, p. 17). Taking these remarks in the context of AI systems development implies a particular and very direct use-case for AI systems in medicine: London seems to envision AI systems directly mapping data to actionable medical decisions on a broad range of specializations and contexts. While at the same time demanding specifically delimited goals and contexts for being able to maintain appropriate metrics that can ensure reliability, in likening the sometimes atheoretic nature of medical experience London implicitly proposes AI applications that output actual decisions, rather than merely further data points human physicians would base decisions on, This does not seem to reflect the current practice of successful medical AI, which is much more focused on specific and rather limited problem definitions that amount to relatively small parts within the decision chain, for which a good deal of heuristics, empirical evidence, and even causal relationships are known and available, cf. (Topol 2019). Surely, few would dare to put AI into use in a completely atheoretic and associationist inference setting, e.g., training only based on anamnestic and corresponding decision data without taking into account existing guidelines, actual evidence from medical research, etc. The reference to "inherited clinical culture," however, rightly hints at the dangers attempting to transfer an AI trained on one particular "clinical culture" on to another, cf. (Smith & Funk, 2021). That, however, does not mean that such an approach could not be utilized effectively in research to find out about new correlations and possibly give rise to the forming of new theories—all with proper caution and scientific rigor. Any of those studies, however, would obviously need substantive evidence before being put into medical practice.

Such use of AI in medical research aside, there seems to be a preconception towards interpretability necessarily relying on causal relations in the (medical) domain. Surely, AI can be trained purely based on empirical evidence, in which case

interpretability would amount to answering question about which empirical evidence weighed in to lead to a particular balancing decision. For instance, AI-based rapid meta-analyses constitute a field, which may completely rely on assessing and evaluating all (or large extents) of medical evidence and guidelines and may thus legitimately constitute completely non-causal inference, see, e.g. (Michelson et al. 2020). Nonetheless, interpretability would require an AI system to transparently inform, among others, about the scope of the meta-analysis, the most highly valued evidence, levels of uncertainty, and how the data is generally assessed in principle.

Thus, and contrary to what is often implied, interpretable medical AI may mean transparent models that reflect causal relationships in terms of mechanistic evidence and/or correlational evidence. Thus, interpretability rather demands to lay open the causal relationships with regard to how the algorithmic decision was reached rather than necessarily the causal relationships with respect to the domain the AI system is being applied to.

3.4 AI is Often Theory-Agnostic

There is a prevailing claim that black-box approaches in AI—especially deep learning systems—are theory-agnostic. Admittedly, claims about AI systems' theory agnosticity are usually qualified to mean that the underlying models would not reflect the causal structure of a problem (London 2019, p. 16). However, even this is only partly true, since at the very least, the models reflect a simple input–output structure and data selection rationales particular to the problem.

Furthermore, there is now a host of methods to incorporate prior knowledge in both post hoc interpretation, see, e.g., (Montavon et al. 2018), as well as in model selection and training, e.g., (Diligenti et al. 2017). Incorporating prior knowledge has even been shown to improve performance (Diligenti et al. 2017). Using iterative AI development processes that make use of evaluation cycles incorporating domain knowledge is further also likely to boost performance or eradicate issues—which may, in fact, be already general responsible AI development practice. A simple positive example is given by Caruana et al. (2015), who elaborate on comparing both a black-box and rule-based AI system for predicting the mortality risk of pneumonia patients delivered to the hospital. Their comparison led to a contextual analysis revealing that internal workflows unaccounted for in the data resulted in erroneous risk assessments. More specifically, pneumonia patients with asthma had always been transferred directly to the intensive care unit, which improved their probability of survival. This informal procedural rule was not reflected in the data acquisition, pre-processing, and interpretation such that both AI systems associated lower mortality risks with an asthma comorbidity.

This shows that AI systems design involves interpretative processes, most of which will be either implicitly or explicitly informed by theory. Mittelstadt and Floridi (2015) write that "[a]t each step the data undergoes a transformation by passing through an interpretive framework, yet custodians act as though it remains an objective analogue of reality," stressing that all-too-often traces of theory are only implicitly acknowledged to the detriment of an AI system's contribution to

responsible and effective practice. Thinking that certain AI approaches are fully theory-agnostic may therefore be denying that both explicit and implicit (tacit) theory and potentially even its forming are at play.

3.5 Trust is Only Grounded in Reliable and Justifiable Results

London questions the potential benefits for patients arising from a demand for explicable AI in medicine. He portraits this as a question of trust. London grounds reasons to trust in an expert's or a system's ability to "produce certain results and justify their actions."

For London, justification is interpreted as an explanation based on knowledge in terms of a "domain model" and "domain principles." The domain model captures causal relationships, whereas the domain principles consider the dynamics of the domain in question, i.e., how to acquire knowledge. London dismisses the need for justification based on the lack of (known) causal relationships in medicine. I have already commented on domain knowledge in medicine in the prequel by contending that it need not be purely causal. Hence, I would like to turn to maintaining that we need to invoke a richer notion of both trust and explanations (and therefore justification), to be able to perceive more far-reaching benefits from explicable AI in medicine. In relation to this, Yoon et al. (2021) emphasize both professional and public acceptance as the potentially "most difficult challenge [machine learning] will face in a high-stakes environment like healthcare." For this purpose, it is instructive to analyze the moral dimension of trust that transcends reliance.

In discussing the European Union's Ethics Guidelines for Trustworthy AI, Rieder et al. (2020) attribute an important moral dimension to trustworthiness that distinguishes it from the notion of reliability. They define "reliance […] in terms of the rational expectation of a dependent person about the person (or entity) being depended upon," which is also found in the rational-choice account of trust (Nickel et al. 2010, pp. 431–433) that is based upon weighing costs and benefits when relying on someone or some artifact to produce results. (Nickel et al. 2010) also add a "motivation-attributing" account to trust. This introduces a moral dimension, which may not be attributed to technological artifacts. However, it can be meaningfully attributed to socio-technical systems, if it is interpreted indirectly and in terms of trust in technology derived from trust in the human agents responsible for the development, deployment and use of the technology under consideration (Rieder et al. 2020). This notion of trust, then, clearly goes beyond a grounding in the mere reliable functioning of a system intended to deliver results. Thus, in evaluating the function of explanations in AI on a limited rational-choice account of trust, the importance of additional functions of explanations remains hidden.

Instead, and even in the face of technological opacity from complexity, it is conceivable that there exist justifications of a quality that does not rely on full technical transparency either on the algorithm's inner workings or on (potentially even unknown) medical mechanisms. Examples could be found in invoking domain models that capture correlative relationships rather than causal ones and domain principles that constitute experimental procedures with uncertain outcomes. During

the use of medical AI there very likely is non-negligible utility in communicating these to the user and patient. Even though explanatory interfaces for black-box algorithms may not be sufficient to alleviate opacity from complexity, they could still act as ways to transport both what is known as well as what is not known as a means to justify a certain decision. This way, explanations serve a different purpose than establishing transparency, perhaps in terms of a regulatory context. Rather, it provides useful domain context during a medical AI system's use in practice that might not be complete, but sufficient to allow for a kind of shared responsibility between medical practitioner, medical AI provider, and patient.

Again, this is in alignment with (Rieder et al. 2020) view on both the epistemic and moral component of trustworthy socio-technical systems. In brief, this consists in being aware of the capacity and limitations as well as openly communicating them in addition to the goals being pursued. From this point of view, I propose to pursue a more differentiated stance on black-box algorithms in medicine. This stance would also not ban non-explicable algorithms altogether. However, rather than purely favoring diagnostic and predictive accuracy in algorithms like (London 2019), this stance would favor explicable algorithms and require to pursue their development whenever non-explicable algorithms are the only ones available for a certain application. This stance is grounded in emphasizing the need to provide explanations from an account of trustworthiness beyond mere reliability and for sustaining intelligibility and accountability. The next sections will exemplify what this may mean in more concrete terms and further give reasons for why explicability in medical AI is beneficial.

4 Ethical and Epistemological Reasons for Favoring Explicable AI

Having attempted to refute the main arguments against a requirement for explanations in medical AI from the literature, I will now turn to novel arguments highlighting the utility of explicable AI in the context of long-term progress in both medical research and practice. The first argument hinges on a viewpoint on the relation between medical research and practice that emphasizes the need for research to be informed, scrutinized, and inspired by practice. Medical AI that provides explanations aids in maintaining feedback between practice and research. The purport of the second argument consists in the view that explanations also improve patient compliance and allow that a chosen therapy can reflect a patient's individual conception of good health. Hence, explanations yield significant contributions to the effectiveness of medical intervention and prevention.

4.1 Establishing Feedback Between Medical Practice and Research

It appears that current discussions on medical AI for decision support are mostly framed in terms of allowing to reap the benefits of evidence-based medicine more efficiently and with less diagnostic and therapeutic errors, see, e.g., (Dias & Torkamani 2019; Gómez-González et al. 2020; Topol 2019). All the while, evidence-based

medicine is mostly portrayed being based on *correlative* evidence for which AI is the perfect vehicle. Not only does this neglect that AI—and even the subfield of machine learning—consists of more than a collection of methods capable of being trained to recognize correlations and patterns. It further misinterprets evidence-based medicine as relying purely on correlative empirical evidence (Maclure 1998). Instead, evidence-based medicine is an approach that requires evidence about actual (clinical) benefits as a replacement for a mere reliance on experience, deductive reasoning from mechanistic theories or even tradition (Webb 2018). Hence, evidence-based medicine does not preclude a rationalist approach to theory forming. Rather, it only demands empirical proof that theories can be rendered into effective treatment options.

However, it would be misleading to deduce from evidence-based medicine's demand for empirical proof that theory forming is entirely unnecessary. If this were the case, all that was left for empiricism would be trial and error. Still, as historical treatises indicate, evidence-based medicine seems to have fallen prey to an over-reliance on empirical findings (Greenhalgh et al. 2014). The original intention behind evidence-based medicine, however, relied on the notion of critical appraisal and integrating evidence into practice rather than blindly following it. Sackett et al. (1996) write:

> Evidence based medicine is not "cookbook" medicine. Because it requires a bottom up approach that integrates the best external evidence with individual clinical expertise and patients' choice, it cannot result in slavish, cookbook approaches to individual patient care. External clinical evidence can inform, but can never replace, individual clinical expertise, and it is this expertise that decides whether the external evidence applies to the individual patient at all and, if so, how it should be integrated into a clinical decision.

Given practical difficulties, time pressures, and the vast array of medical literature, it is understandable that the critical appraisal of medical evidence for each and every patient is not viable and rather necessitates consolidation of some sort, e.g., in terms of theory (Webb 2018) or guidelines. Given the vast amounts of literature and case studies, this may even be an area, where AI may be of assistance, either in the form of inferring viable theories, providing meta-analyses, e.g., (Michelson et al. 2020), or in terms of collecting evidence relevant to the specifics of a particular patient.

Hence clearly, instead of being an approach focused on merely population-based correlative evidence, evidence-based medicine is widely recognized as supporting individual decision-making in which the challenge remains to diverge from the algorithmic guidelines when appropriate, e.g., especially in the face of multimorbidity (Greenhalgh et al. 2014). Critical thinking and reflection are required for its application, but it can undoubtedly inform good and consistent medical practice. In line with this, the notion of integrative medicine calls for a combination of both theory and evidence to cater to the needs of an individual therapy that considers individual factors as well as population-based evidence and in which the therapeutic relationship between physician and patient is not neglected. However, advancing integrative medicine requires the acting

therapeutician to take part in the act of synthesizing evidence for a specific patient to integrating this into individualized treatment options. Such an integration is hardly possible with medical AI systems that remain non-verbose about specific details of their decision support rationale, invoked evidence, and potential mechanistic theories taken into account.

However, there is an open-endedness associated with this approach that, in turn, requires medical practice to also inform the derivation and revision of guidelines and act as an initiator to evidence-based medical research. In reference to Guyatt et al. (1992), Webb (2018) writes:

> […] overemphasis on empirical findings on the part of [evidence-based medicine] would derogate research in the basic and mechanistic sciences in favor of large clinical studies, a move that could forestall progress in biomedical knowledge.

The concept of a research-practice feedback in medicine is also present in what is termed a "learning health care system" (Olsen et al. 2007). In the Institute of Medicine's workshop summary on the Round-table on Evidence-Based Medicine, a proclaimed goal consists in the quickening of "efforts to position evidence development and application as natural outgrowths of clinical care—to foster health care that learns" (Olsen et al. 2007, p. x). The authors further write that "[c]apturing and utilizing data generated in the course of care offers the opportunity to bring research and practice into closer alignment and propagate a cycle of learning that can enhance both the rigor and the relevance of evidence" (Olsen et al. 2007, p. 151).

The learning health care system as envisioned by (Olsen et al. 2007, p. 151) may not explicitly entail that practitioners are necessarily provided with intelligible explanations of evidence-based medical guidelines or decision support systems fed by the latest evidence. However, I will still maintain it is not sufficient to let physicians act as mere executive agents, carrying out the suggestions of opaque decision support systems and collecting data for their improvement. I build this argument on the view that evidence-based medicine requires theory to suggest novel and meaningful empirical investigations, while, in turn, empirical results can inform new and question existing theory. Running randomized clinical trials on account of next to no prior mechanistic assumptions and theories has proliferated, leading authors to call for "science-based medicine rather than evidence-based medicine" (Gorski & Novella 2014). In light of these considerations, it is the creative and, at times, unstructured work of theory forming that is supported by keeping AI-supported medical personnel informed via explicable systems. For this approach to hold value, it would not even take explanations to necessarily be either complete, based on mechanistic evidence or entirely certain, even though this may be desirable. Instead, for serving as a link between medical research and practice, explicable AI could provide even incomplete contrastive, dialogic or, perhaps, even quite selective accounts of explanations, cf. (Miller 2019), in which case, however, rigorous validation of the AI system's reliability under defined nominal circumstances would be a prerequisite. In a sense, this would reconcile both Durán and Jongsma (2021) and London's (2019) perspective with the one I have proposed in this article: Of course, we want

medical AI to be reliable, but there are very important reasons for wanting it to be explicable, as well.

But what are the implications for explicable AI systems? How could explicable AI really contribute to enhancing feedback from practice to research and is there truly a necessity for it? Let us consider the case of clinical guidelines more closely. Clinical guidelines have been shown to improve medical practice (Grimshaw & Russell 1993). However, there is also evidence that compliance with clinical practice guidelines is lacking (Barth et al. 2016). Challenges contributing to this state exist in a potential multitude of guidelines of different quality (Ariel Franco et al. 2020), awareness, and lack of resources, such as time, cf. (Barth et al. 2016), among others. In its epistemological sense, explicable AI could help because giving intelligible explanations generally does not aim at making practitioners reliant on the system but could also contribute to their education. Allowing physicians to memorize adequate interventions and the reasons behind them would allow them to be quicker than forcing them to always use an AI tool. Hence, it is intuitive that explicable AI can help bring research to fruition in practice. But what about the reverse?

In clinical and primary care practice, reality is often not entirely concordant with study or guideline parameters. For instance, multimorbidity can render clinical practice guidelines to be no longer applicable (Ariel Franco et al. 2020). Simply following guidelines in parallel could be harmful, as the simultaneous application of independent clinical practice guidelines has been even shown to potentially yield adverse effects, see, e.g., (Dumbreck et al. 2015). Explicable AI can indicate the grounds on which it recommends certain medical interventions, allowing physicians to disagree and go for alternative treatment options that attempt to take comorbidity into account. In order to enable practitioners to use AI accountably, an AI system must make itself intelligible with the purpose to make an informed decision about when to divert from a standard guideline. In that regard, explicable AI would contribute to the avoidance of overly strict adherence to clinical guidelines. This scenario implies that AI systems are mainly conceived as effective implementations of clinical practice guidelines to speed up medical decision-making, achieve increased compliance with high medical standards and, hence, systematize rigorous and consistent treatment according to the best available evidence, see, e.g., (Terenziani et al. 2003). If such AI systems are clearly delimited in terms of specific goals and contexts, in which they should be applied—as proposed by London (2019) with the purpose to clearly assess reliability and positive outcomes—their application in practical scenarios, involving a multitude of factors that possibly violate these delimiting specifications, would require human intervention.

This is the domain, in which (general) practitioners have to excel: The meaningful, often experience-based/supported and sensible combination of independent indicators or recommendations in cases of unforeseen and badly documented multimorbidity is a highly relevant part of daily medical practice. However, evidence on successfully treating patients with multimorbidity is limited and there is a need for research—and perhaps for practical guidelines—to be better able to cope with these situations (S. M. Smith et al. 2012). Hence, input from practice on how to deal with these complex medical situations is necessary. Inexplicable AI that deprives practitioners of the possibility of an informed rejection of AI-based recommendations

could thus not only incur patient harm, but clearly hampers the forming of experience and individual theories that could end up being the spark that drives systematic research. In turn, explicable AI, conceived not only as a way to better inform practitioners of the reasons behind medical recommendations, but also in terms of allowing probing inputs, could even take an active part in facilitating the refinement or creation of guidelines for specific cases of multimorbidity. Consider the following actual developments and systems as examples that present promising directions in that vein.

The so-called The DECIDE-AI Steering Group (2021) is making a case for small-scale clinical trials of AI-based systems in terms of a rigorously governed process that ensures a positive impact on the health outcome. The group also mentions that "users might wish for an additional key variable to make sense of the algorithm's recommendations, which in turn would require developers to access a totally different section of the electronic patient record" (The DECIDE-AI Steering Group 2021, p. 186). This indicates the necessity to intimately connect the practitioners' perspectives and research on medical AI systems. This may only constitute direct feedback from medical practice to medical AI research. However, medical AI research is already influencing medical research *and* medical practice (Meskó & Görög 2020).

One example of a system that is a first attempt at realizing a meaningful, almost dialogic, medical recommendation system is AsthmaCritic (Kuilboer 2002). The system is designed to analyze data recorded by physicians in primary care, adding critiquing comments to the patient record with respect to the treatment of the comorbidity of asthma and chronic obstructive pulmonary disease. Not only does the very idea of the system rest on analyzing patient records documented during medical practice. The system has also been shown to positively influence physician's data recording and treatment behavior (Kuilboer et al. 2006).

More generally, approaches that involve the mining of electronic health records, cf. (Hernandez Medrano et al. 2018; Lauritsen et al. 2020; Ramakrishnan et al. 2010), present a form of direct feedback from medical practice onto health system development and medical research. Because "it cannot be assumed that users' decisions will mirror the algorithm's recommendations" (The DECIDE-AI Steering Group 2021, p. 186), using the data may present a comparatively rigorous way of closing the loop that allows to keep track of the effectiveness of deviations from current guidelines. But why is explicability even necessary, when AI may—in theory—continuously learn from health records and related health outcomes and one may proceed to update medical guidelines accordingly? Why is there a need to make AI-based recommendations intelligible? As elaborated on above, human intervention can be thought of as injecting theories into such a kind of AI-augmented learning health system. Apart from requiring human professionals as the main bearers of accountability, without the provision of substantive human intervention in unclear cases, the advancement of medical research may be at risk. However, in order to allow these interventions to be made in a responsible way, professionals require at least basic forms of explanations in order to challenge AI-based recommendations and even be able to reject them as not applicable. For instance, Lauritsen et al. (2020) have devised an AI-based early warning score to predict acute critical illness that points out on which electronic health record data the warning is grounded.

All of the above suggests that non-explicable AI in medicine may well aid in the application of medical treatments for some undisputed means, but remains, epistemologically speaking, a dead end, because it cannot make substantial contributions in the theory-empiricism feedback cycle. A seemingly completely different area of advancing medical research and embedding this into clinical guidelines further stresses the need for explanatory interfaces. Involving patients and the public is seen as essential by some to guideline development (Armstrong et al. 2018). Hence, explanations for both medical phenomena and how treatment options can be inferred are very likely to be necessary to meaningfully include lay persons into the process of guideline development. Hence, even if I am mistaken and physicians can advance medical research by both merely carrying out AI-supported medical decisions and collecting data on outcomes without being provided explanations, it can still be argued that the absence of explanations severely limits advances in health on an individual level: Inexplicable AI deprives physicians of the option to meaningfully engage in shared decision-making with their patients, cf. (Bjerring & Busch 2021). As such, inexplicable AI is also preventing feedback from practice in the domain of medical research focused on patient-centered medicine. How can we analyze and further develop responsible patient-centered medical practice, when the medical AI tools are inhibiting shared decision-making in the first place? This leads to the second general thesis that being able to respect patient autonomy provides an ethical argument in favor of explainable AI in medicine. Explicable AI combines an epistemological with an ethical perspective via demanding both intelligibility and accountability. Taking accountability as the acknowledgement of responsibility for action, shared decision-making requires this responsibility to be shared among both physician and patient, thus, respecting patient autonomy. I am going to elaborate on this argument next.

4.2 The Importance of Explanations for Good Health

In the following, I will present arguments elucidating the benefits of explicable AI in medicine that goes beyond a narrow—and perhaps measurable—conception of improving health outcomes. The bottom line is that an explanatory interface between the AI system and physician—or even between AI and patient, directly—contributes to at least two interrelated and highly relevant aspects: (i) to improve patient compliance and (ii) to make space for adapting standardized treatments to individual needs and individual conceptions of good health as well as open possibilities for a wider and more systematic consideration of individual conceptions of health as implicit metrics and assessments of successful health care. In turn, if I am correct, demanding that AI decision support in medicine can help achieve these aspects requires explicability as a necessary principle for responsible AI-facilitated health care. Intelligible explanations form the prerequisite to engage in meaningful shared decision-making with patients, hence, respecting their autonomy. However, intelligible explanations allow both patients and physicians to acknowledge responsibility, hence, allowing a shared way to engage in mutual and accountable decision-making.

4.2.1 Improving Patient Compliance

Explanations in medicine can take many forms beyond being purely based on mechanistic evidence. Given the sometimes selective nature of everyday explanations (Miller 2019), it should be clear that explanations in medical AI can meet a wide range of different qualitative criteria. In turn, it may be right to not put too much emphasis on scientifically correct explanations, but rather consider the patients' and physicians' needs. Accordingly, empirical research on what can be considered useful, e.g., (Berry et al. 1995; Darlington 2011), have shown that patients appreciate explanations generally and can judge usefulness based on content.

Thus, achieving meaningful explanations in medical AI may not necessarily imply simpler and inherently interpretable models, if we step beyond regulatory demands of transparency. Obviously, though, there needs to be some way to verify that both an AI's decision support as well as the explanations offered are beneficial and are not inducing harm, even if verifying correctness may not be possible. Different quantitative and qualitative assessments of beneficence and non-maleficence are required that cover the full spectrum of what medical practice entails. However, discourse on medical AI and a need for (or the dispensability of) explanations seems rather limited on the technological artefact itself as a focus on accuracy and reliability is called for only in terms of the AI system's output. However, proper "reliability" (or rather, beneficence) assessments should also incorporate qualitative information on the patient's uptake, compliance with and successful completion of treatment plans as well as other quality of life factors that involve, e.g., relatives. While (London 2019) argues against an "overreliance on plausible theoretical explanations," since they have led "to treatment practices that harmed patients and consumed scarce resources precisely because key causal claims in those theories were false," it is similarly likely that a lack of plausible explanations can lead to a waste of resources, because patients may not comply with proposed therapy plans.

Studies have long revealed that a physician's concern as well as efforts towards providing explanations contribute to patient compliance (Falvo et al. 1980) and satisfaction (Tarn et al. 2013). This holds true for both non-standard therapy, e.g., such as hirudotherapy (Kim et al. 2017), as well as popular diseases such as diabetes (Koenigsberg & Corliss 2017). The World Health Organization issued a report on patient adherence, outlining the complexity and socio-economic scale of the challenge. Specifically, the report states that (Sabaté 2003, p. 156)

> [s]tudies show that people who receive explanations from a concerned doctor are more satisfied with the help they receive and like the doctor more; the more they like the doctor, the better they follow a treatment plan.

By citing (Haynes et al. 2002), the report also points out that increasing patient adherence to therapies may more thoroughly improve population health than other advances in specific medical treatments. Relating this to medical AI is straightforward: (Triberti et al. 2020) argue that a lack of understanding of AI recommendations and the inability to explain them to patients can "delay or paralyze" clinical decisions. Thus, if an AI system cannot offer or support any explanation for a proposed treatment plan, physicians are either left with their own knowledge or will

have to resort to a stark paternalistic stance by simply asking patients to comply without reasons. The latter appears fundamentally anachronistic given the paradigm of a learning health care system as well as the active part patients' are supposed to play in contributing to detailing ethical requirements associated with the framework, questions of informed consent, and shared decision-making (Faden et al. 2013).

Further, if, in turn, we assume that explanations *are* demanded and *will be* provided by physicians, the danger of plausible, but false explanations outlined by London is even increased if medical AI remains mute on explanations. When physicians find themselves in a position, where explanations are demanded by patients, but the AI system does offer little in the way of supportive mechanistic or correlative evidence, the alignment between the actual proposed treatment plan and any explanation given to the patient despite this lack of support may be severely off. Otherwise, if there is little to no mismatch, one might argue that the AI system may not be of much use anyway. This calls for AI decision tools in medicine to actively contribute to keeping the physicians' way of offering plausible accounts in alignment with the correlative or mechanistic ways medical AI utilizes to offer suggestions.

In addition, McKinley and Middleton (1999) have shown that patients consult physicians (in their case, general practitioners) with their own agenda, i.e., own ideas of which disease they might be suffering from as well as initial ideas on potential explanations. The consultations should take this into account to be effective, given that patients may need to realign their conception of what medical issue is at stake, in order to remain convinced about the treatment plan.

While it is often stated that AI systems can help increase the available time for physicians to interact with patients on a social and empathetic level, see, e.g., (Fogel & Kvedar 2018), all-too-often it is not made clear, how this is going to be achieved. Patient compliance and information on what the individual health issue might be about may just be part of the qualitative content of the interaction between patient and physician that is highly useful in creating a shared understanding and mutual agreement with respect to a treatment plan. In fact, it has been shown that a relationship between a patient and a health care practitioner needs to be based on "mutual trust, respect, and commitment" (Nagy & Sisk 2020, p. 395) as well as that it needs to be nurtured to remain so, in order to improve the health outcome, see (Nagy & Sisk 2020) and references therein. Clearly, non-explicable AI cannot possibly support this important aspect of trying to guarantee therapeutic success. At best, it would not interfere, either. However, this is rather unlikely as there would not be much use for an AI-based recommender system, if the physician was knowledgeable enough to provide every bit of explanatory background information to the patient needed to establish and maintain a trusting relationship.

While it is beyond the scope of this paper to fully explore what it takes to make a medical AI system guaranteedly explicable in the sense that it both provides intelligible outputs as well as strongly promotes its fully accountable use, I will turn to one particular aspect that appears to be essential. It may be an obvious essential ingredient to provide information about the AI's past performance, capabilities and limitations, and design objectives. It may further be useful to provide an interpretive reconstruction of how the AI has produced its output. However, in the sequel, I will argue that giving both physician and patient means to adapt a recommendation

system to an individual conception of good health—one that could have been developed as an implicit result of the shared deliberation and decision-making process—is key to fully realize the utility of explicable AI in medicine.[5]

4.2.2 Adapting to Individual Conceptions of Good Health

So far, I have argued that the provision of mechanistic or correlative explanations may well constitute an essential tool in improving patient compliance and health care outcomes. Beyond that, explanations may also facilitate patient autonomy by allowing more individual decision-making and, hence, lifestyle-compatible interventions. This takes into account that, e.g., in primary care, providing patients with suggestions about adaptations to their lifestyle must involve sensible explanations about why changes are (or have become) advisable and to which extent they can provide a lasting increase in health and well-being. Only on this account can patients either reject medical suggestions or adapt them to their needs in an informed manner. This connection is also made in the above-mentioned report of the World Health Organization, which takes into account that patient compliance is intertwined with bringing suggested therapies in alignment with personal conceptions of health. For instance, the report states that "interventions that target adherence must be tailored to the particular illness-related demands experienced by the patient" (Sabaté 2003, p. XIV).

Hence, the purpose of explanations in medicine is not solely about improving patient compliance, but rather seeks to boost a self-determined approach to personal health, which, in turn, facilitates adherence and improves the chances of success. In this view, by contributing significantly to success in promoting health, explanations may even be regarded part of the treatment or intervention itself. Consequently, non-explicable medical AI would not be able to sustain and build upon this aspect of effectiveness.

A demand for explicability in medical AI then turns out to be an even more important mediator that allows to harness the benefits of AI and the interpersonal relation between patient and physician. At the very least, physicians will need correct and relevant input from an AI decision support system that can form the basis of explanations and dialog-supported interventions tailored to the patient. The power of AI would rest in being able to combine vast amounts of information, being up-to-date about the most promising treatment options and factoring in individual parameters. However, if we are to harness these benefits, we must realize that explanatory interfaces of a certain quality are required.

Explanatory interfaces may be dismissed too casually as counting as just another piece of evidence in the daily routine of physicians and care workers, and, hence, do

[5] Clearly, this level of cooperation between patient, physician, and AI-system that I have hinted at here, imagines a particular medical setting, in which decisions need to be taken that involve significant trade-offs, could determine the patient's future quality of life, or similar aspects. Less consequential use-cases, such as automated registration in radiological imaging, might be significantly less demanding on the capabilities of explanatory interfaces.

not hold particular epistemic power and authority over the user. Clearly, if it were that simple, future research in "explainable AI" would guaranteedly offer methods that allowed physicians to select from a vast source of evidence to convey to the patient as a means to convince of the appropriateness of a treatment. This would only refer to a kind of "technical explainability" that concerns the parts that have contributed to model accuracy, a particular prediction, or perhaps even provide an interpretive reconstruction of how the system may have produced its output, cf. (Anderson & Anderson 2019). While this seems anything but useless, it also occurs the danger of quickly leading to an automation bias, i.e., the tendency to overly trust technology-supported input when trying to reach a decision (Goddard et al. 2011). As illustrated by Bjerring and Busch (2021), reports are increasingly published about medical AI surpassing the power of even experienced clinicians, leading to the likely state that medical personnel does in fact have an *epistemic obligation* to follow AI-based advice, perhaps even irrespective of explanatory interfaces. Hence, both in terms of overtrust in the capabilities of AI technology in general as well as in terms of a particular AI system's past performance, scrutiny of medical personnel could vane. Hence, it might appear increasingly difficult to weigh medical-technical contributors to an AI-based decision support against other, human factors, such as individual conceptions of good health.

Furthermore, it appears unlikely that a medical AI system—however complex—will be able to factor in every kind of individual characteristic and circumstance to fully tailor treatment options to a particular patient (Clancey 1995). This excess information will have to be handled by human physicians, whose task consists also in synthesizing both AI-based decision support with additional patient-individual factors the AI could not consider. It follows that non-explicable AI severely hinders practitioners in adapting the proposed treatment based on patient-individual factors.

Attempts to overcome this issue, e.g., involve the concept of "causability" as the measurable extent to which an explanation supports causal understanding of the issue (Holzinger 2021). For instance, Holzinger and Muller (2021) propose to utilize facial expression and gaze analysis when experts are in interaction with explainable AI interfaces to determine what features of the explainability routines contribute to a positive mapping on an expert's existing mental model. Such approaches aim at making it easier to accommodate AI-based explanations with the human strength in conceptual thinking. Fusing both explanations for AI-based recommendations and a patient's idea of good health on a conceptual level could possibly promote shared decision-making and, hence, respecting a patient's autonomy by allowing both the physician and patient to acknowledge their shared responsibility.

Consequently, an explanatory interface of medical AI should not only feature a uni-directional output of possible causal or correlative evidence, information about the system's past performance, etc. Instead, the interface should allow for probing inputs, providing counterfactual explanations and the possibility to alter other relevant parameters that reflect both medically and ethically relevant weights influencing the final recommendation. While this may sound like high and rather abstract demands on explicable AI in medicine, certain developments seem to already point into this direction. For instance, Madumal et al. (2018) are working towards a dialog model for explainable AI that takes into account both the cognitive as well

as the social process involved in successful explanations. Wachter et al. (2017) have elaborated not only on the utility of counterfactual explanations, but also on how to technically generate them. This would provide a straightforward way to probe even opaque decision support systems and let both physician and patient theorize about sensible alternatives. However, to the best of the author's knowledge, systems that can directly process patient preferences with respect to, e.g., the desired goal of therapy or what constitutes a good quality of life have not been realized, though conceptual descriptions exist, e.g., in radiation oncology (Lambin et al. 2017). However, systems that more generally aspire to provide patient-centric and individualized decision support exist, see, e.g., (Peleg et al. 2017). It is conceivable that similar designs could incorporate interfaces to input preferences for trade-offs, e.g., occurring from the administration of multiple drugs.

Another very important factor involves well-being. Since 1946, the constitution of the World Health Organization states that "[h]ealth is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity." Considering this comprehensive definition, Asadi-Lari et al. (2004) argue that, in order to assess health care intervention properly, including the patient perspective is necessary. This perspective can focus on health needs and (perceived) quality of life for which there are systematic assessment tools (Leplège 1997). However, rather than being something that can be firmly resolved with quantitative approaches (Asadi-Lari et al. 2004), culturally diverse conceptions of health need also be accounted for explicitly by health care professionals (Levesque & Li 2014).

This adaptation certainly requires a mix of both tacit knowledge, experience and generally difficult to assess (and potentially even quantify) parameters that will also not be covered by elaborate individualized or precision medicine-based approaches for some time. For practitioners to make a responsible and well-informed decision on when to deviate from standardized treatments, it thus seems that explanations facilitate both truly individualized care and legal security for physicians by allowing for shared decision-making.

Only if AI-based decision support can allow, or even facilitate to incorporate individual conceptions of good health and well-being into therapeutic decisions, it will be possible to maintain and potentially advance the trend to incorporate individual accounts of health also into research. As indicated above, the provision of explanatory interfaces between patients, physicians, and AI systems plays a key role in making this connection.

5 Outlook—Explicable AI for Good Health

As I have tried to argue, explanations in medical AI are highly relevant for achieving good health outcomes: First, explanatory interfaces can keep physicians within the feedback loop between practice and research, which facilitates epistemological and scientific progress. Second, explanations also support patient-physician communication, e.g., by promoting patient understanding and compliance as well as by aiding in providing the informational basis for integrating the patients' individual conceptions of good health with more standardized treatment plans.

This view also highlights how—given potent explanatory interfaces—AI can, in fact, not only maintain, but improve on how patient-physician and patient-AI interaction can support better health outcomes. For instance, explicable AI, whose outputs can be tailored to specific addressees can cater to the patients' demands to get involved (Deber 1994; Strull 1984). If part of the more trivial diagnoses and interventions can be carried out by patients themselves, this would hold the potential to allow for more interpretive, empathetic, and detailed patient-physician interaction. However, instead of focusing solely on explainable AI in the sense of the view that it should be based on interpretable models, explicable AI is required here that provides intelligible information and (self-)accountable action.

Furthermore, making explicable medical AI systems directly available to the patients can also empower them in maintaining or reflecting upon their own health ideal. Obviously, self-diagnosing and self-therapy come with their own list of ethically relevant implications, on which I would like not to dwell. Rather AI-based independent intelligible explanations can support patient-physician interaction, in which there can be some sense of precaution that the information conveyed is not expressed in a manipulative way (Say & Thomson 2003). The idea of using AI as a second opinion may provide yet further challenges, see, e.g., (Cabitza 2019). However, explanations are certainly needed to address them.

Regarding medical decision-making, it has further been stated that uncertainty is often not assessed or communicated, but maybe highly relevant for weighing different treatment options (such as physician made or AI-based ones) against each other (Kompa et al. 2021). Where possible and not in danger of deluding about non-quantifiable aspects, explicable AI can improve medical decision-making by facilitating acceptance through an appropriate explanatory interface.

Finally, and especially when considering individual conceptions of health, explicable AI can help address the changing support needs of patients with chronic diseases (Thompson et al. 2001). This may happen implicitly, if AI-based medical decision support is increasingly designed with the importance of providing means to align standard therapy with individual conceptions of health in mind, or explicitly, when AI-based self-diagnoses and self-therapy tools can potentially actually factor in a patient's individual characteristics. While the latter may be still farther off, explicable AI in typical physician–patient interaction may already now facilitate shared decision-making by providing a basis on which a decision can be made against or in terms of variations of standardized treatments.

6 Conclusions

This article has scrutinized arguments that suggest that explanations in medical AI should not be regarded as overly important, but that instead accuracy and reliability should rather be the ultimate criteria based on which medical AI can be considered useful. In these discussions the notions of explainable and interpretable AI are often conflated, meaning that explanations are only considered in terms of invoking scientific evidence rather than in terms of the full breadth of explanations that can be encountered and considered useful in medical practice and research. Consequently,

the discussions are often confined to considering an approach to accuracy and reliability that rests on overly simple and technical notions of seemingly rigorous inference problems. In this article, I have elaborated on the significance of explanations of various kinds as a means to widen the discussion about what it actually means to design a reliable medical AI system. The notion of explicability as encompassing both a demand for intelligibility and accountability then leads to the demand that medical AI should support epistemological progress as well as good health understood as being a synthesis of both a scientific and an individual conception of health. To reliably support good health thus means to keep medical practice in feedback with medical research and to account for a patient-centered shared decision-making that allows individual adaptations of standardized treatments and facilitates improved patient compliance. Explicable medical AI is definitely needed to achieve this.

Author Contribution Not applicable.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Anderson, M., & Anderson, S. L. (2019). How should AI be developed, validated and implemented in patient care? *AMA Journal of Ethics, 21*(2), 125–130. https://doi.org/10.1001/amajethics.2019.125

Ariel Franco, J. V., Arancibia, M., Meza, N., Madrid, E., & Kopitowski, K. (2020). Clinical practice guidelines: Concepts, limitations and challenges. *Medwave*, e7887–e7887. https://doi.org/10.5867/medwave.2020.03.7887

Armstrong, M. J., Mullins, C. D., Gronseth, G. S., & Gagliardi, A. R. (2018). Impact of patient involvement on clinical practice guideline development: A parallel group study. *Implementation Science, 13*(1), 55. https://doi.org/10.1186/s13012-018-0745-6

Asadi-Lari, M., Tamburini, M., & Gray, D. (2004). Patients' needs, satisfaction, and health related quality of life: Towards a comprehensive model. *Health and Quality of Life Outcomes, 2*, 1–15. https://doi.org/10.1186/1477-7525-2-32

Barth, J. H., Misra, S., Aakre, K. M., Langlois, M. R., Watine, J., Twomey, P. J., & Oosterhuis, W. P. (2016). Why are clinical practice guidelines not followed? *Clinical Chemistry and Laboratory Medicine (CCLM)*, *54*(7). https://doi.org/10.1515/cclm-2015-0871

Berry, D. C., Gillie, T., & Banbury, S. (1995). What do patients want to know: An empirical approach to explanation generation and validation. *Expert Systems with Applications, 8*(4), 419–428. https://doi.org/10.1016/0957-4174(94)E0033-Q

Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy and Technology, 34*(2), 349–371. https://doi.org/10.1007/s13347-019-00391-6

Burrell, J. (2016). How the machine "thinks:" Understanding opacity in machine learning algorithms. *Big Data and Society, 3*(1), 1–12. https://doi.org/10.2139/ssrn.2660674

Cabitza, F. (2019). *Biases affecting human decision making in ai-supported second opinion settings* (pp. 283–294). https://doi.org/10.1007/978-3-030-26773-5_25

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 1721–1730. https://doi.org/10.1145/2783258.2788613

Clancey, W. J. (1995). The learning process in the epistemology of medical information. *Methods of Information in Medicine, 34*(1–2), 122–130.

Darlington, K. W. (2011). Designing for explanation in health care applications of expert systems. *SAGE Open, 1*(1), 1–9. https://doi.org/10.1177/2158244011408618

DARPA. (2016). *Broad Agency Announcement Explainable Artificial Intelligence (XAI) DARPA-BAA-16–53*. 1–52.

de Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy and Technology, 31*(4), 525–541. https://doi.org/10.1007/s13347-017-0293-z

Deber, R. B. (1994). Physicians in health care management: 8. The patient-physician partnership: Decision making, problem solving and the desire to participate. *Cmaj, 151*(4), 423–427.

Desjardins, J. R. (2006). Responsibilities to future generations: Sustainable development. In *Environmental Ethics: An Introduction to Environmental Philosophy* (4th ed., pp. 70–93). Thomson/Wadsworth.

Di Nucci, E. (2019). Should we be afraid of medical AI? *Journal of Medical Ethics, 45*(8), 556–558. https://doi.org/10.1136/medethics-2018-105281

Dias, R., & Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine, 11*(1), 1–12. https://doi.org/10.1186/s13073-019-0689-8

Diligenti, M., Roychowdhury, S., & Gori, M. (2017). Integrating prior knowledge into deep learning. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 920–923. https://doi.org/10.1109/ICMLA.2017.00-37

Dumbreck, S., Flynn, A., Nairn, M., Wilson, M., Treweek, S., Mercer, S. W., Alderson, P., Thompson, A., Payne, K., & Guthrie, B. (2015). Drug-disease and drug-drug interactions Systematic examination of recommendations in 12 UK national clinical guidelines. *BMJ, 350*(mar11 2), h949–h949. https://doi.org/10.1136/bmj.h949

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*, 329–335. https://doi.org/10.1136/medethics-2020-106820

Epstein, R. M., Fiscella, K., Lesser, C. S., & Stange, K. C. (2010). Why the nation needs a policy push on patient-centered health care. *Health Affairs, 29*(8), 1489–1495. https://doi.org/10.1377/hlthaff.2009.0888

Faden, R. R., Kass, N. E., Goodman, S. N., Pronovost, P., Tunis, S., & Beauchamp, T. L. (2013). An ethics framework for a learning health care system: A departure from traditional research ethics and clinical ethics. *Hastings Center Report*, *43*(SUPPL. 1). https://doi.org/10.1002/hast.134

Falvo, D., Woehlke, P., & Deichmann, J. (1980). Relationship of physician behavior to patient compliance. *Patient Counselling and Health Education, 2*(4), 185–188. https://doi.org/10.1016/S0738-3991(80)80101-7

Floridi, L. (2021). The European legislation on AI : A brief analysis of its philosophical approach. *Philosophy & Technology*, *0123456789*. https://doi.org/10.1007/s13347-021-00460-9

Floridi, L., Cowls, J., Beltrametti, M., Chatila, 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*. *28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fogel, A. L., & Kvedar, J. C. (2018). Artificial intelligence powers digital medicine. *Npj Digital Medicine, 1*(1), 5. https://doi.org/10.1038/s41746-017-0012-2

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning.ArXiv:1806.00069 [Cs, Stat] http://arxiv.org/abs/1806.00069

Goddard, K., Roudsari, A., & Wyatt, J. C. (2011). Automation bias – A hidden issue for clinical decision support system use. International Perspectives in Health Informatics. *Studies in Health Technology and Informatics, 164*, 17–22.

Gómez-González, E., Gomez, E., Márquez-Rivas, J., Guerrero-Claro, M., Fernández-Lizaranzu, I., Relimpio-López, M. I., Dorado, M. E., Mayorga-Buiza, M. J., Izquierdo-Ayuso, G., & Capitán-Morales, L. (2020). Artificial intelligence in medicine and healthcare: A review and classification of current and near-future applications and their ethical and social Impact. *ArXiv*.

Gorski, D. H., & Novella, S. P. (2014). Clinical trials of integrative medicine: Testing whether magic works? *Trends in Molecular Medicine, 20*(9), 473–476. https://doi.org/10.1016/j.molmed.2014.06.007

Greenhalgh, T., Howick, J., Maskrey, N., Brassey, J., Burch, D., Burton, M., Chang, H., Glasziou, P., Heath, I., Heneghan, C., Kelly, M. P., Lehman, R., Llewelyn, H., McCartney, M., Milne, R., & Spence, D. (2014). Evidence based medicine: A movement in crisis? *BMJ (online), 348*(June), 1–7. https://doi.org/10.1136/bmj.g3725

Grimshaw, J. M., & Russell, I. T. (1993). Effect of clinical guidelines on medical practice: A systematic review of rigorous evaluations. *The Lancet, 342*(8883), 1317–1322. https://doi.org/10.1016/0140-6736(93)92244-N

Guyatt, G., Cairns, J., Churchill, D., Cook, D., Haynes, B., Hirsh, J., Irvine, J., Levine, M., Levine, M., & Nishikawa, J. (1992). Evidence-based medicine a new approach to teaching the practice of medicine. *JAMA - Journal of the American Medical Association, 268*(17), 2420–2425.

Haynes, R., McDonald, H., Garg, A., & Montague, P. (2002). Interventions for helping patients to follow prescriptions for medications. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.cd000011

Hernandez Medrano, I., Tello Guijarro, J., Belda, C., Urena, A., Salcedo, I., Espinosa-Anke, L., & Saggion, H. (2018). Savana: Re-using electronic health records with artificial intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence, 4*(7), 8. https://doi.org/10.9781/ijimai.2017.03.001

Herzog, C. (2019). Technological opacity of machine learning in healthcare. *2nd Weizenbaum Conference: Challenges of Digital Inequality - Digital Education, Digital Work, Digital Life*. https://doi.org/10.34669/wi.cp/2.7

Herzog, C. (2021). On the risk of confusing interpretability with explicability. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00121-9

Holzinger, A. (2021). Explainable AI and multi-modal causability in medicine. *I-Com, 19*(3), 171–179. https://doi.org/10.1515/icom-2020-0024

Holzinger, A., & Muller, H. (2021). Toward human–AI interfaces to support explainability and causability in medical AI. *Computer, 54*(10), 78–86. https://doi.org/10.1109/MC.2021.3092610

Holzinger, A., Weippl, E., Tjoa, A. M., & Kieseberg, P. (2021). Digital transformation for sustainable development goals (SDGs)—A security, safety and privacy perspective on AI. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (Vol. 12844, pp. 1–20). Springer International Publishing. https://doi.org/10.1007/978-3-030-84060-0_1

Independent High-Level Expert Group on Artificial Intelligence Set Up By the European Commission. (2019). *Ethics Guidelines for Trustworthy AI*.

Kiener, M. (2021, July 8). Can "taking responsibility" as a normative power close AI's responsibility gap? *CEPE/IACAP Joint Conference 2021: The Philosophy and Ethics of Artificial Intelligence*.

Kim, K. S., Sim, H. S., Shin, J. H., Hwang, J. H., & Lee, S. Y. (2017). The relationship between explanation and patient compliance in hirudotherapy. *Archives of Craniofacial Surgery, 18*(3), 179–185. https://doi.org/10.7181/acfs.2017.18.3.179

Koenigsberg, M. R., & Corliss, J. (2017). Diabetes self-management: facilitating lifestyle change. *American Family Physician*, *96*(6).

Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: Communicating uncertainty in medical machine learning. *Npj Digital Medicine*, *4*(1). https://doi.org/10.1038/s41746-020-00367-3

Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy and Technology, 33*(3), 487–502. https://doi.org/10.1007/s13347-019-00372-9

Kuilboer, M. M. (2002). Feasibility of AsthmaCritic, a decision-support system for asthma and COPD which generates patient-specific feedback on routinely recorded data in general practice. *Family Practice, 19*(5), 442–447. https://doi.org/10.1093/fampra/19.5.442

Kuilboer, M. M., van Wijk, M. A. M., Mosseveld, M., van der Does, E., de Jongste, J. C., Overbeek, S. E., Ponsioen, B., & van der Lei, J. (2006). Computed critiquing integrated into daily clinical practice affects physicians' behavior: A randomized clinical trial with AsthmaCritic. *Methods of Information in Medicine, 45*(04), 447–454. https://doi.org/10.1055/s-0038-1634103

Lambin, P., Zindler, J., Vanneste, B. G. L., De Voorde, L. V., Eekers, D., Compter, I., Panth, K. M., Peerlings, J., Larue, R. T. H. M., Deist, T. M., Jochems, A., Lustberg, T., van Soest, J., de Jong, E. E. C., Even, A. J. G., Reymen, B., Rekers, N., van Gisbergen, M., Roelofs, E., & Walsh, S. (2017). Decision support systems for personalized and participative radiation oncology. *Advanced Drug Delivery Reviews, 109*, 131–153. https://doi.org/10.1016/j.addr.2016.01.006

Lauritsen, S. M., Kristensen, M., Olsen, M. V., Larsen, M. S., Lauritsen, K. M., Jørgensen, M. J., Lange, J., & Thiesson, B. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications, 11*(1), 3852. https://doi.org/10.1038/s41467-020-17431-x

Leplège, A. (1997). The problem of quality of life in medicine. *JAMA: The Journal of the American Medical Association, 278*(1), 47. https://doi.org/10.1001/jama.1997.03550010061041

Levesque, A., & Li, H. Z. (2014). The relationship between culture, health conceptions, and health practices: A qualitative-quantitative approach. *Journal of Cross-Cultural Psychology, 45*(4), 628–645. https://doi.org/10.1177/0022022113519855

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36–43. https://doi.org/10.1145/3233231

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

Loh, E. (2018). Medicine and the rise of the robots: A qualitative review of recent advances of artificial intelligence in health. *BMJ Leader, 2*(2), 59–63. https://doi.org/10.1136/leader-2018-000071

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

Maclure, M. (1998). Mechanistic versus empirical explanations and evidence-based medicine. *Acta Oncologica (stockholm, Sweden), 37*(1), 11–12. https://doi.org/10.1080/028418698423113

Madumal, P., Miller, T., Vetere, F., & Sonenberg, L. (2018). Towards a grounded dialog model for explainable artificial intelligence. ArXiv:1806.08055 [Cs] http://arxiv.org/abs/1806.08055

Marcinkevičs, R., & Vogt, J. E. (2020). Interpretability and explainability: A machine learning zoo minitour. ArXiv:2012.01805[Cs]http://arxiv.org/abs/2012.01805

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

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics, 45*(3), 156–160. https://doi.org/10.1136/medethics-2018-105118

McKinley, R. K., & Middleton, J. F. (1999). What do patients want from doctors? Content analysis of written patient agendas for the consultation. *British Journal of General Practice, 49*(447), 796–800.

Meskó, B., & Görög, M. (2020). A short guide for medical professionals in the era of artificial intelligence. *Npj Digital Medicine, 3*(1), 126. https://doi.org/10.1038/s41746-020-00333-z

Michelson, M., Chow, T., Martin, N. A., Ross, M., Tee Qiao Ying, A., & Minton, S. (2020). Artificial intelligence for rapid meta-analysis: Case study on ocular toxicity of hydroxychloroquine. *Journal of Medical Internet Research, 22*(8), e20007. https://doi.org/10.2196/20007

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt, B. D., & Floridi, L. (2015). The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics, May*. https://doi.org/10.1007/s11948-015-9652-2

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing A Review Journal, 73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Muller, H., Mayrhofer, M. T., Van Veen, E.-B., & Holzinger, A. (2021). The ten commandments of ethical medical AI. *Computer, 54*(7), 119–123. https://doi.org/10.1109/MC.2021.3074263

Nagy, M., & Sisk, B. (2020). How will artificial intelligence affect patient-clinician relationships? *AMA Journal of Ethics, 22*(5), E395-400. https://doi.org/10.1001/amajethics.2020.395

Nickel, P. J., Franssen, M., & Kroes, P. (2010). Can we make sense of the notion of trustworthy technology? *Knowledge Technology and Policy, 23*(3–4), 429–444. https://doi.org/10.1007/s12130-010-9124-6

Olsen, L., Aisner, D., & McGinnis, M. J. (2007). The learning healthcare system. In *IOM Roundtable on Evidence-Based Medicine – Workshop Summary*. National Academies Press. https://doi.org/10.17226/11903

Pasquale, F. (2015). *The Black Box Society—The secret algorithms that control money and information*. Harvard University Press.

Peleg, M., Shahar, Y., Quaglini, S., Fux, A., García-Sáez, G., Goldstein, A., Hernando, M. E., Klimov, D., Martínez-Sarriegui, I., Napolitano, C., Parimbelli, E., Rigla, M., Sacchi, L., Shalom, E., & Soffer, P. (2017). MobiGuide: A personalized and patient-centric decision-support system and its evaluation in the atrial fibrillation and gestational diabetes domains. *User Modeling and User-Adapted Interaction, 27*(2), 159–213. https://doi.org/10.1007/s11257-017-9190-5

Ramakrishnan, N., Hanauer, D., & Keller, B. (2010). Mining Electronic Health Records. *Computer, 43*(10), 77–81. https://doi.org/10.1109/MC.2010.292

Rieder, G., Simon, J., & Wong, P.-H. (2020). Mapping the stony road toward trustworthy AI: Expectations, problems, conundrums. *SSRN Electronic Journal*, 1–14. https://doi.org/10.2139/ssrn.3717451

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines, 29*(4), 495–514. https://doi.org/10.1007/s11023-019-09509-3

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

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

Sabaté, E. (Ed.). (2003). *Adherence to long-term therapies: Evidence for action* (Issue February 2003). World Health Organization.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ, 312*(7023), 71–72. https://doi.org/10.1136/bmj.312.7023.71

Sætra, H. S. (2021). AI in context and the sustainable development goals: Factoring in the unsustainability of the sociotechnical system. *Sustainability, 13*(4), 1738. https://doi.org/10.3390/su13041738

Say, R. E., & Thomson, R. (2003). Clinical review decisions—Challenges for doctors. *British Medical Journal, 327*(September), 542–545.

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA Journal of the American Medical Association, 320*(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163

Smith, S. M., Soubhi, H., Fortin, M., Hudon, C., & O'Dowd, T. (2012). Managing patients with multimorbidity: Systematic review of interventions in primary care and community settings. *BMJ, 345*(sep03 1), e5205–e5205. https://doi.org/10.1136/bmj.e5205

Smith, G., & Funk, J. (2021). AI has a long way to go before doctors can trust it with your life. *Quartz*. https://qz-com.cdn.ampproject.org/c/s/qz.com/2016153/ai-promised-to-revolutionize-radiology-but-sofar-its-failing/amp/. Accessed 17 Mar 2022

Strull, W. M. (1984). Do patients want to participate in medical decision making? *JAMA The Journal of the American Medical Association, 252*(21), 2990. https://doi.org/10.1001/jama.1984.03350210038026

Tarn, D. M., Paterniti, D. A., Orosz, D. K., Tseng, C. H., & Wenger, N. S. (2013). Intervention to enhance communication about newly prescribed medications. *Annals of Family Medicine, 11*(1), 28–36. https://doi.org/10.1370/afm.1417

Terenziani, P., Montani, S., Bottrighi, A., Torchio, M., Molino, G., Anselma, L., & Correndo, G. (2003). Applying artificial intelligence to clinical guidelines: The GLARE approach. In A. Cappelli & F. Turini (Eds.), *AI*IA 2003: Advances in Artificial Intelligence* (Vol. 2829, pp. 536–547). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-39853-0_44

The DECIDE-AI Steering Group. (2021). DECIDE-AI: New reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nature Medicine, 27*(2), 186–187. https://doi.org/10.1038/s41591-021-01229-5

Thompson, M., Gee, S., Larson, P., Kotz, K., & Northrop, L. (2001). Health care professional support for self-care management in chronic illness: Insights from diabetes research. *Patient Education and Counseling, 42*(1), 81–90. https://doi.org/10.1016/S0738-3991(00)00095-1

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Triberti, S., Durosini, I., & Pravettoni, G. (2020). A "third wheel" effect in health decision making involving artificial entities: A psychological perspective. *Frontiers in Public Health, 8*(April), 1–9. https://doi.org/10.3389/fpubh.2020.00117

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3063289

Webb, W. (2018). Rationalism, empiricism, and evidence-based medicine: A call for a new Galenic synthesis. *Medicines, 5*(2), 40. https://doi.org/10.3390/medicines5020040

Weld, D. S., & Bansal, G. (2018). The challenge of crafting intelligible intelligence. ArXiv:1803.04263 [Cs] http://arxiv.org/abs/1803.04263

Weller, A. (2019). Transparency: Motivations and challenges.ArXiv:1708.01870 [Cs] http://arxiv.org/abs/1708.01870

Yoon, C. H., Torrance, R., & Scheinerman, N. (2021). Machine learning in medicine: Should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics*, medethics-20