Ethics and Information Technology (2022) 24:20
https://doi.org/10.1007/s10676-022-09646-x

ORIGINAL PAPER

Relative explainability and double standards in medical decision-making

Should medical AI be subjected to higher standards in medical decision-making than doctors?

Hendrik Kempt[1] · Jan-Christoph Heilinger[1] · Saskia K. Nagel[1]

Accepted: 10 March 2022 / Published online: 5 April 2022

Abstract

The increased presence of medical AI in clinical use raises the ethical question which standard of explainability is required for an acceptable and responsible implementation of AI-based applications in medical contexts. In this paper, we elaborate on the emerging debate surrounding the standards of explainability for medical AI. For this, we first distinguish several goods explainability is usually considered to contribute to the use of AI in general, and medical AI in specific. Second, we propose to understand the value of explainability relative to other available norms of explainable decision-making. Third, in pointing out that we usually accept heuristics and uses of bounded rationality for medical decision-making by physicians, we argue that the explainability of medical decisions should not be measured against an idealized diagnostic process, but according to practical considerations. We conclude, fourth, to resolve the issue of explainability-standards by relocating the issue to the AI's certifiability and interpretability.

Keywords Explainability · Heuristics · Double standards · Certifiability · Interpretability · Responsibility · Diagnostics · Medical decision-making

Introduction

With the increased presence of medical AI in clinical use, a variety of ethical questions has arisen regarding the acceptable and responsible implementation of said technologies. One main concern lies with the presumed lack of sufficient or satisfactory explainability of the AI's decision-making processes. This lack of explainability is usually thought to be a *theoretical* challenge in itself (Durán, 2021), as we ought to be able to explain the tools we are working with; but it is also gives rise to several *practical* challenges, as the lack of explainability can be a source of uncheckable bias (Panch et al., 2019), of soft paternalism by recommending certain treatment-plans without accounting for individual preferences (McDougall 2019; Grote & Berens, 2020), and of a presumed responsibility gap by physicians being unable to fully understand the machine whose suggestion they may have to rely on (including potential mistakes) (e.g., Habli et al., 2020).

In this paper, we discuss these concerns with the intention (a) to compare and contrast AI-based medical practice with the established practices of human physicians, (b) to test whether a lack of explainability poses genuinely new problems, and (c) to ask if the generally discussed standards for medical AI constitute higher requirements, and thereby a "double standard" (Zerilli et al., 2019) and, if so, if such double standard may be permissible.

To address these questions, we proceed as follows. *First*, we discuss the issue of explainability from the four different types of goods it supposedly provides: preconditional, instrumental, conditional, and democratic. Such distinction appears to be both necessary and useful: necessary, because many arguments within the XAI debate appear to consist of a mix those goods, and useful because it allows us to structure the debate in the field of medical ethics, as issues of consent and responsibility are especially pertinent to explainability-questions.

✉ Hendrik Kempt
Hendrik.kempt@humtec.rwth-aachen.de

[1] Applied Ethics Group, RWTH Aachen University, Theaterplatz 14, 52062 Aachen, Germany

With a growing number of authors arguing for a decreased role of explainability (London (2019), Zerilli et al (2019), Durán (2021), Sand et al. (2021), Ferreira (2021), and others for an increased role (McDougall 2019; Grote & Berens, 2020; Bjerring & Busch, 2020), the relevance is given to analyze the value of explainability closer. We introduce the distinction of "*absolute explainability*" and "*relative explainability*" to characterize the debate further: "absolute explainability" refers to the value of explainability on its own, while "relative explainability" of AI holds that the ethical significance of explainability can only be inferred from comparing it to our other methods of explaining decisions.

Second, we discuss current norms and realities of medical explanations in the clinic. We find that many of the points made against AI are present in a physician's decision-making. Not the communicative limits of individual physicians are of interest for these considerations, but the epistemologically justified heuristics and uses of bounded rationality. These illustrate, along with certain accepted unintentional biases or random mental events, that the optimization process of diagnostics is an idealized, impractical demand rarely, if ever, fulfilled. From this, we infer that the requirements for physicians are justified through reliabilist assumptions rather than the deductive-diagnostic process. Coming from the distinction of relative explainability, then, we ought to compare the explainability-requirements against these practical compromises rather than the idealized process.

Third, if we understand using AI as comparable to using heuristics, we can ask whether the fact that physicians choose to use heuristics is relevant from the perspective of responsibility. Physicians can only take responsibility for a diagnosis they can theoretically at least explain, and it appears that for this fact, AI ought to be more advanced in its explainability than a physician's self-chosen heuristic can be. Especially in the context of the ability to take responsibility for mistakes, these clarifications are necessary. Thereby, a simple transfer of the practical levels of explainability to the use of AI is unsuccessful. We may hold physicians responsible for their lack of explainability and potential mistakes, but we cannot do the same with AI. Thereby, a higher standard of AI may be required, while not constituting a double standard.

Fourth, we propose to locate the goods provided by explainability of AI machines in the certification process and the interpretative work done by physicians, instead of considering explainability as a core normative requirement in the use of medical AI. If we can find acceptable levels of precision, accuracy, and reliability, we may speak of an AI being certifiable. The ability to explain as to why it is so accurate may not be required after all, as we are used to certifying pharmaceuticals as well based on their efficacy and side-effects. On the other hand, the interpretative work of what evidence a machine produces ought to be meaningful for the clinical use and oriented along the needs of physicians and patients. Thus, it needs to be interpretable. Having some medical AI certified according to the acceptable levels of accuracy dissolves the need for it to be explainable in detail to patients to confirm the conditional good of explainability.

Relative explainability

The problem of explainability of tools

Not every technological invention yields genuinely new ethical and epistemological challenges. However, philosophers agree that explainability (or, as Floridi et al (2018) put it, "explicability"), and the lack thereof constitute such a challenge (for an overview on the different dimensions of this debate, see Mittelstadt et al., 2019). AI, especially deep learning, has led to largely autonomous decision-making processes in algorithms that can rarely be fully explained in mechanistic-causal terms. This warrants the questions to what degree we should be able to explain these tools and under which conditions we ought to use them, if at all.

Different positions have been put forward in response to this question, with most granting that explainability of AI poses serious ethical and epistemological challenges that ought to be addressed *before* implementing them. Especially in discussions surrounding medical decision-making (in diagnostics, risk-assessments for operations and other medical interventions, as well as prescribing certain treatments like medical drugs), the use of AI poses a high-stakes opportunity to save lives. However, without properly being able to explain the tools used to make these decisions, AI appears to simultaneously endanger long-held norms of responsibility attribution. In the following, we reconstruct the goods increased explainability is supposed to provide to medical decision-making.

Four goods of explainability

The assumption that we ought to be able to explain how a tool works before we use it and while we use it has immediate intuitive appeal. It seems prima facie irresponsible to even consider using a machine the workings of which we do not fully understand, let alone using it in high-stakes contexts like medical diagnostics. From dangers of misuse to misunderstanding its results, there are many reasons to find concern in the lack of such ability, and for which we assign value when present.

This intuitive appeal to insist on high levels of explainability for the AI in use, however, can lead to insufficient differentiation between the different types of good provided by explanations. Additionally, the question of what actually triggers the strong expectation about the ability of experts to explain the tools they are using may require some closer examination than what intuitions provide at first glance.

In opposite to approach this debate from different kinds of explanations like Mittelstadt et al. (2019), we propose to disentangle four different goods that increased explainability of medical AI can provide: preconditional, consequentialist-instrumental, deontological-conditional, and democratic goods. This distinction allows to analyze exactly what kind of good is provided by increased explainability of an AI, and how these different goods are usually included in arguments for or against certain methods of AI (i.e., deep learning), the underlying training data, or the tools (i.e., diagnostics algorithms), especially in respect to medical uses. From these goods, we usually come to value-judgments about the inclusion of explainability-norms in medical decision-making. We do not claim these lists of good to be the only four, or that they cannot be subsumed under more inclusive terms. Rather, we chose these four goods as especially relevant to explainability.

The distinction between kinds of explanations, such as causality, simulatability, decompositionality or post-hoc rationalizability (Lipton, 2016) is very useful to the issue of explainability: by providing different methods to increase our understanding of how specific machines work, we can increase the desired explainability for differently programed, machine-learned algorithms or our use-specific knowledge requirements in using them. Different ways to program algorithms, their uses, and our knowledge-requirements can necessitate different kinds of explanations.

However, we stress that approaching the issue of explainability from a value-theory perspective can clarify why explainability as such, i.e., indifferent to its different forms, is considered a morally desirable endeavor and can introduce arguments that uncover tacit presumptions about our standards of medical explanations. Thus, the debate around the kinds of explanations and the value of explanations is not mutually exclusive but pursue different, equally important purposes: distinguishing kinds of explanations can help explain more algorithms in different forms, distinguishing goods of explainability can help explain why we should explain in the first place.[1]

Preconditional good

As a general epistemic rule, experts ought to understand the tools they are using. Without looking at the consequences of using a machine, it appears a priori correct that it is better to know more about the machine's functions than less when operating it. This preconditional understanding or knowledge of tools in use can be understood as a conceptual connection between "expertise" and tool-knowledge as an epistemic virtue. This condition has probably the biggest intuitive appeal ("an expert should know what they are doing!"). We can question the strength of this preconditional knowledge, however, as experts in using the machine are not the same as experts in constructing and hence explaining the workings of a machine. In medical contexts this shared labor of medical and technical experts becomes more apparent, as the operator of an MRI machine may know how the machine works, but does not know how to diagnose a patient based on the machine's results, while a radiologist may be able to explain how to diagnose the patient, while not being able to explain the machine that provides the patient's data.

An expert that not only knows what a machine does, but also how the machine does it, is usually considered better equipped to deal with the machine than someone who can "merely operate " it. Both are considered valuable: the former usually is considered part of the explainability requirement, while the latter is less about the machine and more about the discipline, in our case medical diagnostics. And though this may constitute a supererogatory duty of experts, in high stakes contexts like medical diagnostics, these epistemic duties of physicians to possess the know-what and the know-how on the machines they are using seem to have a certain intuitive appeal.

The issue with medical AI currently lies with the lack of such assistants or technical staff that can be called to explain the workings of a machine. This is why some have called for specifically educated medical-technical staff that work on medical AI alone (Jha & Topol, 2016, e.g., call for the conceptualization of radiologists and other diagnostic professionals as "information specialists").

Another example demonstrating the value of preconditional knowledge emerges in the case of the "never failing oracle": If we imagine an oracle whose diagnoses are always correct without fail, the practical need to explain its workings are diminished—we simple would ask it to diagnose patient after patient and work from there. However, even then philosophical curiosity will insist on answers that explain how it works and how the oracle can be correct, as it must have access to knowledge that we would like to share. Theoretically speaking, then, if a machine would never fail a given task, there would be *no practical need* to explain its workings. And yet, the cognitive purpose of humanity's natural inquisitiveness to explain it would persist nevertheless.

Consequentialist-instrumental good

Increased explainability comes, so goes the argument, with a better understanding of the potential errors of the machine, its biases, malfunctions (Lombrozo, 2011). Therefore, in

[1] We would like an anonymous reviewer for pointing towards this distinction.

order to further improve the performance of the diagnostic process at large, and to avoid introducing new issues previously not present, we ought to achieve a certain level of explainability of the AI used in clinical contexts. This is specifically independent of current methods—the goal of constantly improving diagnostic methods will potentially require to improve an AI's explainability.

Further, the higher the level of explainability the better it can be adjusted for specific purposes, increasing its usefulness even more. Deep-learned AI is usually fairly inflexible, having led to misleading promises about its performance in non-ideal (i.e., non-laboratory) contexts. With increased knowledge about the pathways and patterns of decision-making processes within the AI we may be able to adjust the decision-making according to real-life conditions.

Many of the technical research in explainability of performance-improvement is primarily consequentialist, as even non-ethical considerations of performance are affected by explainability: for example, the ability to sell algorithms for a wider range of applications or an increased robustness widens markets, which both requires an extended understanding of how algorithms work.

Medical AI can also profit from increased explainability of this kind, as the parameters of medical care vary heavily. Even small issues such as indoor-light can influence the performance of an AI at this stage as the attempt of using laboratory-tested eye-scanning algorithms have somewhat unfortunately impressively shown (Heaven, 2020). These adjustments are best implemented with an increased understanding of how the AI works, and what it gets wrong and why.

Deontological-conditional good

If the decision-making process of an AI is directly or indirectly affecting a patient, it appears reasonable to assume that explainability of the decision-making process is a condition for informed consent. While many patients probably do not care about the technological details of the diagnosing machine, as they already do not care about the technological details of current technology (but, if at all, about the diagnostic system of human physicians and their use of such technology (Montague et al., 2010)), the point of the conditional good of explainability is its potentiality. It is at least theoretically possible to explain the process of current technology used in diagnostic processes, while this explainability is both lost in current methods of training AI, as well as ever more needed as the AI is not only providing evidence on which a physician can base their diagnosis on, but may provide a (preliminary) diagnosis on its own.

The ability of a machine to not only take over more complex cognitive pre-instructed tasks, but to perform most parts of the decision-making process, appears to increase the conditional requirements for explainability from the patient's perspective, which is usually connected to making informed decision about the risks and benefits of certain treatments, to which an explainable AI can contribute (Bjerring & Busch, 2020).

However, the question of responsibility attribution may arise here as well. Explainability seems to be normatively connected to the level of explainable decisions (Grote & Di Nucci, 2020). In the case of medical AI, the physician ought to be able to explain their decision, and if medical AI was part of the decision-making process, physicians ought to be able to explain how the AI's process features in their own decisions. And while this may be less relevant in a context where role-responsibilities demand of physicians to take responsibility even for processes in part of our their control, the requirements for lowering this burden by making medical AI "inspectable" (Zerilli et al (2019) use this term) to the degree that physicians can take responsibility seems a justified demand.

Democratic good

Next to the specific goods that can emerge from interactions between technology and its users (physicians and their patients), the context in which these machines are created can be attributed as democratic good (e.g., Robinson, 2020). The higher the explainability of a machine, the more it can be subject to democratic discourse about its constitution.

In AI, these democratic goods of explainability expand on one side to its transparency, i.e., which data was used, how it was sourced, and how the algorithm was trained, and on the other it allows for increased bias detection. The High Level Expert Group for AI (HLEGAI) accordingly includes explainability and transparency as part of their "trustworthy AI " approach (HLEGAI, 2019).

Especially given the racially and gender-biased history of medicine, we should expect traces of these biases to be present in the medical training data of these algorithms. Only an open, democratically controlled and supervised process of training methods can ensure a decrease in biased diagnoses, which we consider distinct from the consequential and deontological good. Explainability can thus be a source of improving the social conditions of human–machine interactions, as increased explainability requires transparency of potential biases.[2]

[2] We owe the consideration of the democratic good as a good in its own right to the suggestion of Thomas Ploug.

Relative versus absolute explainability

These types of good are usually considered to constitute explainability as necessary for the ethically justified use of AI in medical practice and elsewhere. However, this necessity presupposes that the explainability of AI is a goal to be achieved on its own, as we ought to explain how the machine makes decisions to achieve the aforementioned goods. Following the assumption about this context-free relevance of AI-explainability, several authors conclude certain ethical (im-) permissibilities or formulate ethical concern about the use of AI that is insufficiently explainable. E.g., McDougall's claim that an unexplainable AI may reach levels of decision-making autonomy that can re-introduce an unjustified soft paternalism (McDougall 2019; Grote and Berens 2020, concurring), while Bjerring and Busch (2020) are equally concerned for a patient's ability to consent. We call the view that AI needs to be explainable "no matter what" the "*absolute explainability view.*"

However, the ethical contribution of explainability of AI can only be properly understood in comparison with other methods of decision-making and their explainability without having to deny the goods it can provide. Thereby, the value of explainability is best thought of to be relative, in which different decision-making processes and their levels of explainability ought to be compared (the "*relative explainability view*"). In the context of medical diagnostics, the explainability of AI ought to be assessed in the light of the explainability of physicians's decision-making processes and their norms.

This is not to say that the different types of good of explainability are relative; in fact, increased explainability appears to be a purpose striving for independent of other practical considerations. However, in the assessment of just how much we have to be able to explain the decision-making processes of AI, it is possible to presuppose an undue double-standard on AI when compared to established standards of medical explanations. While higher standards for AI (in contrast to physicians) might be justifiable, double-standards are to be avoided. The harm of double-standards, i.e., undue high standards for explainable AI, becomes apparent when they are used to discredit the potential of a promising technology because it is not achieving these undue high standards, or to increase the regulatory process to implement them in clinical contexts.

While normative standards should not be lowered to merely dream up a more technologically advanced, automated but ultimately unethical clinic, the chances of improving medical care both for the caring and the cared for ought to be taken into consideration when determining explainability standards.

Medical explanations in practice: heuristics and bounded rationality

Having established the idea that explainability of AI ought to be considered relative to the explainability of human reasoning, we now need to explore what the norms of explainability consist of. However, we also have to ask whether these accepted practical norms of decision-making and explainability are transferable to decision support systems, or whether considerations of backwards-looking responsibility justify different standards.

For this, we first need to distinguish between the *communicative* and the *epistemic* quality of human explainability. Many studies concentrate on the communicative practices of physicians to offer explanations to their patients on the level a patient can understand them (e.g., Elmore et al., 2010; McLafferty et al., 2006). This poses a vital issue as the ability to make informed decisions rests on a patient's understanding of the given medical situation and decision-making options. Increased explainability of this kind, then, means to be able to make others understand. This debate is vast and has many further differentiations we cannot consider here (in particular regarding how accurate the patient's understanding of the medical information actually is), that are mainly relevant for issues surrounding patients' ability to provide consent. And while the communicative skills of human physicians may be one feature to compare to an AI's ability to communicate its behavior, the subjective, idiosyncratic variance, including non-expertise related linguistic skills such as emotional intelligence may not yield a very useful measure to infer norms for AI from. Further, the communicative task of the decision-making process can be delegated to non-expert staff, i.e., nurses, while the epistemic quality of a decision, i.e. the reasons why the decision was made in a certain way, cannot be performed by non-experts.

Thus, the epistemic dimension of explainability of diagnosis affects the very ability of human doctors to explain their reasoning and decision-making processes. Thereby, clarifying the epistemological challenges for physicians is the key for setting justifiable norms for AI explanations.

The ideal medical explanation is deductive based on established rules of causal relations between certain symptoms and causes (Maung, 2017). The range of accepted methods of medical explanations for "optimal explanations" is vast and cannot be fully covered here. It is more instructive to analyze the limits of those explanations, to discern the rule of minimum-explanations that fulfill the ethical requirements of clinical processes in practice. And while it is certainly the case that many physicians surpass these minimum requirements, they certainly also fail to meet these standards, at least occasionally. This negative-approach to explanation incorporates the idea previously elaborated upon that the acceptability of explainability of AI ought *not* to be measured against an ideal world, or against absolutes, but in relation to the actually available explainability of similar decision-making processes that underlies the explanations offered by medical personnel.

Knowing the limits of when an explanation should be considered sufficient, then, seem to be a key ingredient of setting the minimal requirements for AI explanations. As many decision-making methods cannot be optimized due to practical constraints in the process, several authors have worked towards reconstructing acceptability-criteria, such as "bounded rationality" (Simon, 1957), or narration-approaches (Hunter, 1996). In the following, we discuss heuristical measures that help physicians make decisions when optimization of the accuracy diagnosis is not a viable option.

Heuristics as standards

We find limits to the optimization of reasoning-processes, as some of the insights produced may be implicit, unaware, spontaneous, or temporary (see also Ferreira (2021) for parallelization of human and AI "inscrutable processes" and a justification for both). Anyone who ever had a good idea will know the feeling that this idea may have "sprung out of nowhere", also researched in the philosophy of mind as "eureka moments". And if those reasonings are insufficiently supported by evidence, we ask experts to make educated guesses. An educated guess, however, is still based on some inferential background credence of the guesser (Horowitz 2019), consisting in a demonstrable history of being a good guesser, i.e. experience.

However, in not giving too much ground to heuristical explanations of decision-making, we usually require doctors to provide intelligible explanations of their reasoning, i.e. we expect the physician to be aware of and, upon request, to be able to explain the heuristics they deploy. These heuristics are then considered justified tools for decision-making, even if they are based on a reduced amount of information considered for the decisions in question (Marewski and Gigerenzer 2012). Heuristics for medical decision-making, and credence-supported educated guesses, thus indicate a solution to a necessary trade-off between explainability and other demands of clinical decision-making, such as urgency and accuracy.

As Marewski & Gigerenzer as well as Tverksy and Kahnemann (1974) point out, we usually are permitting these heuristics as (a) the acknowledgement that these are the only practicable methods of clinical decision-making under conditions of urgency, and (b) that they still yield reliable results in terms of accuracy. What Gigerenzer calls "fast-and-frugal-heuristics" are one way of justifying epistemic shortcuts to increase the number of decisions made while admitting a small increased inaccuracy. Of course, the application of those heuristics to decide certain cases is explainable as well. However, this moves the explanandum from the epistemically optimal diagnosis to the procedural method (the heuristic) that optimizes diagnostic decisions according to limits of clinical rationality while retaining good amounts of success (in accuracy and resource-saving). If a physician is explaining a diagnosis after using heuristics, they do not explain their decisions based on evidence, but they explain why they looked for certain evidence—they explain the process of decision-making, not the decision.

Obviously, this trade-off has limits, as we would not want to trade ever faster heuristics for an increasing lack of explainable reasoning behind them. Additionally, in cases of lower decisional urgency, optimization may be a reasonable way to go after all. The very debate surrounding the rational application of heuristics as a concession to the practical limitations of an ideal theory of medical decision-making, shows the relativity of explainability in its normative role for medical decisions.

Lastly, the entire debate surrounding biases and other cognitive hurdles and limits (Saposnik et al., 2016) can be viewed through the lens of relative explainability: differences in medical education and specialization, levels of experience, general distribution of biases both in medical education as well as in personal attitudes, and random contextual features like attention span, association chains, and priming may all feature into a diagnostic judgment, affecting its reasoned explainability. Yet, all these features are acknowledged and accepted in our norms of medical decision-making. And while these limiting features are not normatively justified, it goes to show that they are accounted for in a relative explainability: some physicians are better in providing explanations, and the minimum required explanation is a compromise of practical requirements.

AI as heuristic

Having established that the acceptable baseline of explainability lies considerably below the optimization process of causal and deductive reasoning, we conclude that standards of explainable AI must not be compared to the idealized standards of medical diagnostic explanations: the optimization process of diagnostics is never fully achieved, and cannot reasonably be expected to be achieved, either. The expectations of physicians lie in making decisions that balance explainable reasoning, procedural speed, reasonable levels of accuracy, and potentially a minimization of bias. Consequently, any expectation of explainability made towards an AI-based application requires justification, in particular if they diverge from the standard expectation that is made towards medical personnel.

While Gigerenzer argues against using an optimized diagnosis as the normative goalpost, in the medical context such optimized diagnosis may serve well as an ideal from which discounts for AI decision-making processes can be negotiated. That is why we ought to discuss whether AI should face similar expectations, or if the standards for explainability of AI diagnosis should be higher than relying on heuristics.

One of the initial conditional goods of explainability introduced above consisted in the ability to identify or take responsibility for mistakes when being able to explain the complex decision-making process. The ability to take responsibility, and to accept legal liability, is the basis for the processes of "error-management", in which reviews of decision-making processes weigh the decisions of doctors and their explanations for such decisions. This allows for clear distributions of responsibility of physicians—they may consciously decide to forgo the optimization process and rely on a heuristic.

The challenge of responsibility

Understanding medical diagnostic AI as a heuristic in itself, however, precludes physicians from taking responsibility, as physicians cannot choose to forgo the optimization process; instead, using AI-diagnostic tools is akin to using heuristics, just without the possibility to opt out: while a physician can still decide that the applied heuristic method did not yield a reliable result, and therefore different tests or diagnostics strategies must be taken, the use of an AI usually produces the data on which physicians may base their diagnosis in the first place.

While heuristics are sufficiently justified as a decision-making method, the ability to choose them as a way to move forward is the mechanism allowing physicians to take responsibility for their outcomes.

In having physicians rely on AI-as-heuristic, without plausibly being able to opt-out, we can reasonably demand a different, that is higher standard for the explainability of those machines. In order for physicians to reasonably take responsibility for mistakes a machine made, they ought to be able to understand how the machine operates, and how errors occur when they occur. Both the preconditional and the instrumental goods of explainability come into play in this manner: only by properly understanding the machine and its workings can physicians detect potential malfunctions, and avoid having to take responsibility for misdiagnoses they could not have prevented from happening, e.g. the propensity of an AI for diagnostic bias, if the AI is assigning too much weight to a person's gender based on the training data.

Lately, Sand et al. (2021) have put forward a list of requirements for physicians to take forward-looking responsibility for using medical AI. In their analysis, they state the need to update the "entrustable professional activities" (EPAs) of physicians to include the ability, among others, to connect image quality and likelihood of accuracy of output. Their list is a helpful way to ensure proper training of future medical professionals. However, while this list provides a useful update for AI-literacy requirements of medical professionals, it remains agnostic about the issue of explainability as a prerequisite for backward-looking responsibility.

However, if AI is to become an integral part of a physician's process to diagnose and treat a patient in a trade-off with an optimized process, they must be able to rely on the use of these devices. We argue, thus, that the requirements of the challenge of responsibility cannot be fulfilled by the physicians themselves, but ought to be addressed in the initial introduction of AI to clinical use (comparable to the introduction of new blood tests or other diagnostic tools).

Relocating explainability: certifiability and interpretability

The main task of this paper is to determine what levels of explainability medical AI has to achieve for it to be responsibly implemented in clinical use. So far, we have shown that the goods of explainable AI are best understood relative to the norms of actual explainability of physicians' heuristics, which suggests that the current, often implicit explainability requirements of AI should be lowered. However, considering the use of AI *as* heuristic, the ability of physicians requires their ability to rely on the theoretical explainability of AI, while also being able to convey the machine's results in an intelligible manner to their patients. For this, we propose a distinction that reflects the practical need of explainable AI in medical use: certifiability and interpretability. Taken these two together, both can fulfill the four goods explainability provides, while avoiding the creation of double-standards for machines.

Certifiability and risk assessment

It is unlikely, and—considering the previously discussed sharing of labor in many tech-rich contexts—almost impossible, that physicians will be able to fully explain how the machinery works that they operate. From this, we can assume that the factual explainability of a given diagnostic AI will not play a relevant role in clinical uses, as it already does not for any other technology. The question, then, is more where such explainability ought to be located. Sand et al. (2021) also point out that "entrustable professional activities" of physicians currently do not require physicians to be able to explain MRI and other technologies.

In analysing the acceptability of degrees of explainability of other tools, we again can use the concept of *relative explainability* to help find the norms for AI to be *certifiable*.

One such object of comparison may be pharmaceuticals and their certification process. While we acknowledge that the comparison of therapeutics and diagnostics may hold relevant differences in some areas of ethical considerations, the requirement for their explainability in order to be justifiably used appears to be similar enough to take insights from the one for the other.

For pharmaceuticals, for example, we usually favor a certification process that determines their therapeutic efficacy through their observable effects on humans. In fact, while the causal, chemical explanation of how such drugs work may be a helpful preconditional or instrumental good, the more important feature for patients and physicians alike when prescribing it is what the effects and side-effects could be.

These, however, are not based on the explanation of its chemical components, but through the previously performed testing and validation process. In reference to the epistemological theory of reliabilism, in which a conclusion is not only acceptable by its reasoned justification, but also if it is a result of a previously justified reasoning process, this holds for pharmaceuticals: when certifying them, we do not primarily justify the causal chemistry behind them, but the actual effects those pharmaceuticals had in extended testing processes.

In maintaining that relative explainability ought not only be considered in comparison to other methods of explaining the same decision, as we discussed above regarding with physicians, but also as a demand for coherence, we can turn towards potential certification-processes in medical AI. An approach with a similar goal has been put forward by Tutt (2017), who proposes a regulatory body to organize and guide the certification process. This kind of „FDA for algorithms " should, according to Tutt, possess the powers to classify, certify, and control the introduction of algorithms used for medical diagnostics. In fact, the regulatory side of introducing AI to medical services, known also under the concept of „Software as a medical device " (SaMD), connects quite well to the philosophical requirements introduced here.

In order to avoid double-standards here, we may propose that reliabilism is also applied in the introduction of medical AI (see Smart et al (2020) for a general introduction to this kind of epistemic and moral justification in AI, and de Fine Licht and de Fine Licht (2020) for a similar approach to explainability as source for public legitimacy of algorithms). Several authors have moved towards criticizing such differences as double-standard in reference to our intentional stances towards machines (Zerilli et al., 2019), while others justify them in reference to design-requirements, i.e. what we called absolute explainability (e.g., Günther & Kasirzadeh, 2021). By pointing towards reasonable standards of accuracy, precision, and reliability in their processes, we claim that AI ought not to face higher standards than the certification-process of using and producing pharmaceuticals faces. These standards are indeed high, but also allow physicians to morally justified "outsource" questions of explainability to the certification process. This positions us between those two camps: some higher standards are justified, but not to be understood as a double-standard.

Such a process may require an algorithm to provide an increased level of "answer justification" (see e.g., Sharp et al (2017)) and other advanced ways of explaining more precise decision-making pathways (like heatmaps or inspectability), as well as certain levels of transparency of data allocation and training (Felzmann et al., 2019, 2020). This addresses the democratic good, as a public or scientific discourse about the acceptable standards of imprecision can be negotiated. Yet, it is to be expected that these still do not have to reach the "optimization goal" of a fully explainable diagnostic process. As we have seen that even in the most ideal circumstances we would not expect physicians to be able to provide those—and yet, we praise, recommend, and trust physicians not merely on their ability to provide explanations. Thus, we may certify AI that is reliably precise, and that we can explain sufficiently for physicians to rely on it.

Similar to the certification of medical pharmaceuticals, the effects of using AI—their reliable results—are usually the measure for acceptability of risks, rather than the complete causal explanation for those results in every instance. And similar to the package insert included with pharmaceuticals, the conditional good of explainability can be reproduced through clear communication of the tested and validated precision and reliability of such AI. This standard is considerably higher than the norm for diagnostic explanations of physicians, but it does not seem unjustified: we certainly do not want to certify an AI that cannot claim superiority over human performances.

Interpretability

While locating the requirement of explainability in the certification process rather than clinical uses, the conditional good of physicians and patients interacting with machines still needs to be addressed. Reasonably, physicians always require some information from the machine and its decision-making process, with the same applying to the requirements for informed consent of patients. The fact that a machine is just *more accurate* than a human is usually not enough. While London (2019) argues against explainability in favor of precision, we have seen that there may be pre-conditional or conditional goods that suggest that, if the opportunity arises, an increase of explainability ought to be welcome and worked towards. These (pre-) conditional goods of explainability, however, are also achieved through a machine providing information that can lead to meaningful decision-making.

As long as both patients and physicians are satisfied with the information given (i.e., they are empowered to successfully access risks and benefits in the light of a certified AI), they should feel confident in coming to a decision (of course, making informed medical decisions will remain some of life's toughest decisions to make in the first place). However, the requirement for providing information for such decision-making is best understood as interpretability rather than explainability. Similar positions on the relevance of interpretability over explainability have been put forward before in other contexts (Rudin, 2019) and for medical imaging in particular (Reyes et al., 2020). In combination with the required certifiability, however, interpretability will cover the goods that explainability has previously been thought to provide.

The package insert indicating potential side effects is an illustrative example of the information required for an acceptable interpretability of patients in medical decision-making, and under consideration relative explainability, arguments of why AI should provide much more than presume a double-standard.

Interpretability on its own appears problematic due to its lack of connection to the actual workings of the interpreted machine-behavior (we may read anything into a machine's behavior if we do not know how it works). Certifiability on its own appears problematic because it is insufficient for actual clinical use, as the machine may only produce results with very little regard for the physician and patient's needs. Together, however, those offer an adequate answer for the concerns posed by unexplainability.

Conclusion

Understanding explainability not as contributing an absolute value to the ethical decision-making in clinical contexts but as one that is relative to several different factors allows for an analysis of what goods explainability actually provides. As we concluded, measuring the explainability-demands for AI against the idealized explainability of the diagnostics optimization-process of physicians is not only misleading but would most certainly constitute a double-standard. Higher standards, other than double-standards, however, can be justified in regards to considerations of responsibility for mistakes. To accommodate those, we proposed to locate the explainability not in physician–patient interactions, but in the certification-process in which the AI is being tested and explained to a reasonable degree. For clinical uses, with a certified accuracy and reliability in the background, interpretability becomes a much more important factor. With certifiability and interpretability in place, all the goods explainability provides are still present, but are more operationable, relative to our current customs.

Author contributions All authors contributed extensively to the paper. HK wrote the main parts of the manuscript, JCH and SKN commented substantially on all parts.

Funding This research has been developed and funded by the project ELSA-AID (grant number 01GP1910A) of the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung (BMBF)), as well as the project KIPeriOP (grant number 2520DAT10G) by the German Federal Ministry of Health (Bundesgesundheitsministerium (BMG)).

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest We have no conflicting or competing interests to declare.

References

Bjerring, J. C., & Busch, J. (2020). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology*. https://doi.org/10.1007/s13347-019-00391-6

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & Society*. https://doi.org/10.1007/s00146-020-00960-w

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science, 76*(4), 444–463.

Durán, J. (2021). Dissecting scientific explanation in AI (sXAI): a case for medicine and healthcare. *Artificial Intelligence*. https://www.sciencedirect.com/science/article/abs/pii/S0004370221000497v?via%3Dihub

Elmore, J. G., Ganschow, P. S., & Geller, B. M. (2010). Communication between patients and providers and informed decision making. *Journal of the National Cancer Institute. Monographs, 41*, 204–209. https://doi.org/10.1093/jncimonographs/lgq038

Felzmann, H., Fosch Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society, 6*(1), 1–14. https://doi.org/10.1177/2053951719860542

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics, 26*, 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Ferreira, M. (2021). Inscrutable processes: Algorithms, agency, and divisions of deliberative labour. *Journal of Applied Philosophy*. https://doi.org/10.1111/japp.12496

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*(3), 205–211. https://doi.org/10.1136/medethics-2019-105586

Grote T. & Di Nucci E. (2020). Algorithmic Decision-Making and the Problem of Control. In: Beck B., Kühler M. (eds) Technology, Anthropology, and Dimensions of Responsibility. Techno:Phil—Aktuelle Herausforderungen der Technikphilosophie, vol 1. J.B. Metzler, Stuttgart. https://doi.org/10.1007/978-3-476-04896-7_8

Günther, M., & Kasirzadeh, A. (2021). Algorithmic and human decision making: For a double standard of transparency. *AI & Society*. https://doi.org/10.1007/s00146-021-01200-5

Habli, I., Lawton, T., & Porter, Z. (2020). Artificial intelligence in health care: Accountability and safety. *Bulletin of the World Health Organization, 98*(4), 251–256. https://doi.org/10.2471/BLT.19.237487

Heaven, W. D. (2020). Google's medical AI was super accurate in a lab. Real life was a different story. *MIT Technology Review*. Retrieved 2021 from https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/

High Level Expert Group on AI of the European Union. (2019). Ethics guidelines for trustworthy AI. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

Hunter, K. M. (1996). Narrative, literature, and the clinical exercise of practical reason. *The Journal of Medicine and Philosophy, 21*(3), 303–320. https://doi.org/10.1093/jmp/21.3.303

Jha, S., & Topol, E. J. (2016). Adapting to artificial intelligence: Radiologists and pathologists as information specialists. *JAMA, 316*(22), 2353–2354. https://doi.org/10.1001/jama.2016.17438

Lipton, Z. C. (2016). The Mythos of Model Interpretability. *Communications of the ACM 61(10).*https://doi.org/10.1145/3233231

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass, 6*(8), 539–551. https://doi.org/10.1111/j.1747-9991.2011.00413.x

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

Maung, H. H. (2017). The causal explanatory functions of medical diagnoses. *Theoretical Medicine and Bioethics, 38*(1), 41–59. https://doi.org/10.1007/s11017-016-9377-5

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. J Med Ethics *45*, 156–160.

McLafferty, R., Williams, R. G., Lambert, A. D., & Dunnington, G. L. (2006). Surgeon communication behaviors that lead patients to not recommend the surgeon to family members or friends: Analysis and impact. *Surgery*. https://doi.org/10.1016/j.surg.2006.06.021

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279–288). https://doi.org/10.1145/3287560.3287574

Montague, E. N. H., Winchester, W. W., & Klein, B. M. (2010). Trust in medical technology by patients and health care providers in obstetric work systems. *Behaviour & Information Technology, 29*(5): 541–554. https://doi.org/10.1080/01449291003752914

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: Implications for health systems. *Journal of Global Health, 9*(2), 010318. https://doi.org/10.7189/jogh.09.020318

Reyes M., Meier R., Pereira S., et al. (2020). On the interpretability of artificial intelligence in radiology: Challenges and opportunities. Radiology: Artificial Intelligence. 2020;2(3):e190043. https://doi.org/10.1148/ryai.2020190043

Robinson, S. C. (2020). Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). *Technology in Society*. https://doi.org/10.1016/j.techsoc.2020.101421

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

Sand, M., Durán, J. M., & Jongsma, K. R. (2021). Responsibility beyond design: Physicians' requirements for ethical medical AI. *Bioethics*. https://doi.org/10.1111/bioe.12887

Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: A systematic review. *BMC Medical Informatics and Decision Making, 16*(1), 1–14. https://doi.org/10.1186/s12911-016-0377-1

Sharp, R., Surdeanu, M., Jansen, P., Valenzuela-Escárcega, M. A., Clark, P., & Hammond, M. (2017, August). Tell me why: Using question answering as distant supervision for answer justification. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 69–79).

Simon, H. A. (1957). *Models of man; social and rational*. Wiley.

Smart, A., James, L., Hutchinson, B., Wu, S., & Vallor, S. (2020). Why reliabilism is not enough: Epistemic and moral justification in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 372–377). https://doi.org/10.1145/3375627.3375866

Tutt, A. (2017). An FDA for algorithms. *69 Administrative Law Review*. https://doi.org/10.2139/ssrn.2747994

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32*(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6s

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.