Philosophy & Technology (2022) 35:100
https://doi.org/10.1007/s13347-022-00598-0

RESEARCH ARTICLE

Justice and the Normative Standards of Explainability in Healthcare

Hendrik Kempt[1] · **Nils Freyer**[1,2] · **Saskia K. Nagel**[1]

Received: 30 March 2022 / Accepted: 10 November 2022/ Published online: 23 November 2022

Abstract

Providing healthcare services frequently involves cognitively demanding tasks, including diagnoses and analyses as well as complex decisions about treatments and therapy. From a global perspective, ethically significant inequalities exist between regions where the expert knowledge required for these tasks is scarce or abundant. One possible strategy to diminish such inequalities and increase healthcare opportunities in expert-scarce settings is to provide healthcare solutions involving digital technologies that do not necessarily require the presence of a human expert, e.g., in the form of artificial intelligent decision-support systems (AI-DSS). Such algorithmic decision-making, however, is mostly developed in resource- and expert-abundant settings to support healthcare experts in their work. As a practical consequence, the normative standards and requirements for such algorithmic decision-making in healthcare require the technology to be at least as explainable as the decisions made by the experts themselves. The goal of providing healthcare in settings where resources and expertise are scarce might come with a normative pull to lower the normative standards of using digital technologies in order to provide at least some healthcare in the first place. We scrutinize this tendency to lower standards in particular settings from a normative perspective, distinguish between different types of absolute and relative, local and global standards of explainability, and conclude by defending an ambitious and practicable standard of local relative explainability.

Keywords Clinical decision support systems · Justice · Medical AI · Explainability · Normative standards · AI ethics

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

Nils Freyer
freyer@fh-aachen.de

Saskia K. Nagel
saskia.nagel@humtec.rwth-aachen.de

[1] RWTH Aachen University, Theaterplatz 14, 52062 Aachen, Germany

[2] FH Aachen, Eupener Straße 70, 52066 Aachen, Germany

1 Introduction: Explainability as a Normative Standard

Explainability has been identified as an ethical challenge *sui generis* to artificial intelligence (Floridi & Colws, 2019). The prevailing assumption, especially in cases of practical application with real-life stakes, is that there is an imperative to increase the explainability of artificial intelligent algorithms. The more explainable the decision-making processes of such algorithms are, the better for their justifiability. In the medical professions, the challenge to provide a sufficient level of explainability, or to reliably increase it, has been problematized as a challenge to the possibility of informed consent (Grote & Berens, 2020; McDougall, 2019) and has since then been contested on different grounds (Durán & Jongsma, 2021; Zerilli et al, 2019; Arbelaez Ossa et al, 2022). While Arbelaez Ossa et al. argue that acceptable standards of explainability are dependent on the individual clinical context, the general trend in these arguments seems to be moving from strict explainability standards towards alternatives (such as interpretability and its different forms, like contestability or rationalizability; cf. for an analysis Lipton, 2018). In these trends, the normative role of these alternatives remains the same: in proposing certain required features of a technology, standards are being established for their ethically justified use. Thus if a machine does not exhibit these features, there is a reason to criticize or even reject its use.

Standards of explainability arise in conjunction with other practical requirements of such technologies, such as their reliability and precision. Explainability comes as a gradient and not as a binary, and any standard of explainability is chosen on this gradient. It is these choices that require normative justification. For some uses we may be satisfied with the mere functional explainability that allows for simple corrections and bug fixes, since these allow for reliable, albeit not fully explainable machines. Some uses require higher levels of explainability to be ethically permissible: they can affect human autonomy or relationships by forcing patients to make decisions based on diagnoses that are insufficiently explainable, or to trust a physician's expert-opinion even though this opinion is based on evidence the physician cannot explain to the patient. In the field of medical AI, explainability and its role in evaluating the permissibility of deploying AI-based technologies have been controversially discussed (cf. Amann et al, 2022). Medical technologies call for particularly high standards, including those pertinent to explainability. The stakes involved in medical decision-making (which medical technology is often intended to aid) are much higher than for most other human activities: the right decision in medical contexts can decide life or death questions, can reduce or increase suffering manifold, can affect the autonomy and future prospects of individuals, and — in the case of epidemiology — can affect entire communities. Thus, not only is the medical decision-making process in itself under high normative justificatory pressure, but also any technology that will impact this process.

The rules developing from those efforts have been evolving around securing the health and well-being of those affected without putting too many restrictions on those attempting to help. The results, at least in the societies of the Global

North, are usually subsumed in the classical four principles of biomedical ethics suggested by Beauchamp and Childress (1989/2009): autonomy, beneficence, non-maleficence, and justice.

The addition of explainability as a genuinely new category to medical AI ethics can either be considered part of these four principles insofar as it affects a patient's ability to provide informed consent to diagnoses and treatments and thus their autonomy (McDougall, 2019; Grote & Berens, 2020; Andreotta et al., 2021), or stand as a medical ethics issue on its own (Arbalaez Ossa et al., 2022). Either way, the necessity to be able to explain the technology used for medical decision-making processes represents a normative standard unquestioned in its principle relevance.

In this paper, we will also consider normative standards of the explainability of medical decision-making processes from a global justice perspective, one insufficiently explored in the current ethical debate about medical AI.

2 Conceptual Clarifications: Explainability Far and Wide

An investigation into the connection between the issue of explainability as conceived in the discussion around AI ethics and questions of distributive and relational justice requires a clear analysis of the terminology at hand. While "explainability" is often distinguished alongside other standards for assessing explanations of machine-made decisions (i.e., interpretability, and contestability), there are useful internal differentiations that help to connect these previously separated concepts.

2.1 Explainability: Absolute or Relative Standards

While some authors argue that we should weigh the performance of a specific technology by the improvement it brings to the tasks it is used for (London, 2019; Rudin, 2019; Ghassemi et al., 2021), the general assumption is that explaining medical AI is necessary for its justifiable use (Bjerring & Busch, 2021; Grote & Berens 2020; Wadden, 2022). Considering that explainability is not a binary but a gradient, the question of just how much we should be able to explain these technologies emerges. As has been pointed out elsewhere (Arbalaez Ossa et al. 2022; Kempt et al., 2022), explainability is often only defined by the features of a technology itself. This conceptualization of "explainability" as a feature independent of explanatory standards for decisions comparable to human decisions can be called "absolute explainability" (Kempt et al., 2022). From this concept follows that there is a technology-specific standard of being transparent (in the sense of epistemic accessibility to understand its inner workings). This way, worries about the "black boxes" of AI ought to be reduced, as the increased explainability of machines can offer ways to clarify how the technology operates (Bjerring & Busch, 2021).

If we cannot explain the full workings of a technology and its decision-making process, and thus are not able to pre-determine the outcomes of the behavior of such a technology, so the argument goes (cf. Zednik, 2021), we encounter several ethical

concerns of responsibility, trust, and liability. Standards of absolute explainability require efforts to lighten up the black box to avoid these concerns.

Opposing this view, the concept of relative explainability introduces the idea that the standards of explainability are to be understood in a contextualized form. For example, the standards of machine-explainability could reflect what kind of depth of explanations we would expect from human decision-makers. The key difference lies in the normative standards that are used to assess new, big-data-trained technology: absolute explainability suggests that these technologies must be explainable independent of their use, while relative explainability seeks to find comparable, customary, and justified norms of explanations and applies them to this technology.

In medical explanations, this difference between absolute and relative explainability comes into force for diagnostics and treatments alike. For the former, we may put less normative pressure on explainable diagnoses if comparable diagnoses proposed by doctors would also only cover a limited level of depth. Similarly, especially in a fast-paced diagnostic environment, heuristics and other explanatory "shortcuts" are considered acceptable, suggesting that this could apply to technology as well. For the latter, we have established norms for dealing with not fully explainable treatments, e.g., most drugs are only partially based on a causal understanding of their workings, and their responsible use is secured through proper certifications and validations and individual risk assessment and consent.

Thus, for both diagnostics and treatment, there are outweighing reasons for explainability, often in forms of practical need and theoretical limitations of understanding, requiring a contextualized approach, i.e. relative explainability.

2.2 Explainability: Global and Local Standards

In understanding explainability as a relative feature of medical AI, the problem of locating the relata to which the standards of machine explainability should be set arises: what is the object of comparison to which the explainability of medical AI is relative? For the sake of generalizability of normative claims, contemporary normative standards seem to be orientated on the global optimum of explainability standards, i.e., the best possible explanation of human physicians. Though this has already been shown to result in double standards for human physicians and AI alike (Zerilli et al., 2019), little attention has been given to the implications of existing differences in local standards of explainability in the provision of healthcare.

The demand for generalizable global standards of medical practice is a common feature of medical ethics: ethically justified standards ought to apply everywhere with the same force. The expectation for a certain quality of medical care as well as the ethical considerations for the implementation and application of such care is reflective of justified claims of patients and physicians indifferent to their material living situations.

However, from any contextualized and non-ideal perspective of justice, it becomes clear that not every medical context actually *can* provide the same healthcare. Considerations about healthcare are thus inevitably, for pragmatic reasons, divided into global and local ones. Global considerations reflecting somewhat

idealized medical circumstances, reflective of the highest currently available medical standards (most often found in affluent countries of the Global North), and local considerations reflecting the realities in already deprived areas while securing justifiable location- or situation-specific standards.

Take, for example, the discussion surrounding the need for a development of local explainability standards in Sub-Saharan Africa (Penu et al. 2021). While the authors do not concentrate on an analysis of medical AI, the thrust of the argument remains the same: enterprises in Sub-Saharan Africa face different challenges and may require different standards in their activities from those imposed on them from abroad.

2.3 A Four-Way Matrix of Standards of Explainability

The question must now be how we ought to assess the standards of explainability for AI-based medical decision support systems (AI-DSS) when including these varied justice considerations. Curiously, the topic of justice in the contextualization of explainability has, as of yet and to our knowledge, not been put forward or discussed in the literature. Explainability, even if problematized in terms of the sufficiency of explanations, is centered around specific clinical settings and informed consensus (see, e.g., Arbelaez Ossa et al., 2022).

If we understand standards of technologically mediated medical explanations to be both dependent on the standards of a physician's explanations (absolute vs. relative) and on the standards of real-life decision-making contexts (global vs. local), we encounter four options for organizing an AI's need to justify its decision-making processes (Table 1):

This matrix shows the different possible permutations of the introduced distinctions. Global absolute explainability, which we consider to be the standard approach when discussing ethically acceptable standards for AI-DSS, is concerned with the universal standards for medical AI indifferent to our current standards of explainability in medical decision-making and indifferent to local practical limits. This approach, however, has been accused of producing double standards for AI as we appear to demand higher standards of this technology in terms of explainability than we did and do of physicians in current medical practice. These higher standards do not seem to be put on reliable ethical footing once we consider other factors than the explainability of technologies.

On the opposite side, local relative explainability covers those decisions made cognizant of the limits of local requirements. When there is only one physician having to make a rushed decision with little to no knowledge about the patient, this decision is still not floating in the sea of indifference, but within a highly modified, situation-specific context, like immediate gut feelings, lack of relevant expertise, or fast and frugal heuristics (Gigerenzer & Todd, 1999). One can still make morally blameworthy bad decisions in these situations — yet, the blame is not measured against the highest possible standards of explainability, but rather measured against the baseline standards of the expected explainable decision-making (i.e., following a standardized procedure in case of limited deliberation times).

Table 1 Dimensions of explainability standards

	Local/partial	Global/universal
Relative	Local/partial standards relative to local experts and locally available technologies and established standards	Global/universal standards that apply relative established standards of human explanation
Absolute	Local/partial standards that apply to any decision of AI-DSS	Global/universal standards that apply to any decision of AI-DSS

Global relative explainability covers the explainability of decisions of technology compared to ethically justified and standardized human capacities. However, as these relative standards are compared to the highest possible level of human capacity (i.e., the explainability of decision-making processes available to a patient with access to a well-educated physician with sufficient time and capacities), they miss accounting for the needs of non-ideal contexts.

We understand global relative explainability to be a more helpful approach to the explainability issue of technologies than global absolute explainability (cf. Kempt et al., 2022), since an absolute standard may lead to double standards and an unjustified expectation towards performances of machine explanations.

The box for local absolute explainability standards, in this matrix, is somewhat empty, as nobody would seriously propose explainability standards for technology based on a very specific set of local limitations (Arbelaez Ossa et al., 2022, may go in this direction). This randomness of local principles demanding universal validity is not a tenable ethical position.

2.4 A Short Comment on Accuracy in Explainability Standards

We understand explainability to be an independent parameter for the evaluation of the normative permissibility of an AI-DSS. Another such parameter, often considered in the same evaluative context, is the AI-DSS's performance (see e.g., Ploug & Holms, 2020 discussing contestability). In medical AI contexts, performance most often refers to the machine's accuracy or reliability to deliver results. Especially accuracy is often discussed in contrast to explainability rather than as connected parameters: take as an example London's discussion favoring the former even at the expense of the latter (London, 2019), Kempt et al.'s introducing context-dependent explainability requirements (Kempt et al. 2022), or, as introduced above, Arbaelez Ossa et al. (2022). Often enough, it seems, are the ability to robustly explain an AI-DSS's process and the accuracy of the desired output two separate parameters to investigate. We do not deny that performance is usually considered the more relevant parameter for ethical permissibility of technologies (imagine one technology that is fully explainable but basically always wrong against one technology that is fully unexplainable but always correct in its output); however, our investigation is explicitly concerned with explainability standards as a topic of both ethical considerations as well as engineering challenges. Thus, we keep the parameters of accuracy and explainability distinct from each other.

3 The Core Problem: Global Inequalities in Healthcare

Despite the existing struggles and injustices in the distribution of resources, access to global healthcare services, and the possible implications of an increased role of AI-DSS on the same, considerations of justice are often underrepresented in the debate on AI-DSS.

In the light of global absolute explainability standards, many AI-based decision support systems fall short of these standards and thus are argued not to be morally admissible (e.g., Bjerring & Busch, 2021). Even when adjusted for global relative explainability standards, systems are only admissible if they surpass a physician's ability to explain their own decision-making (Kempt & Nagel, 2021). However, in areas in which expertise is scarce, one could argue that AI-DSS can be beneficially used, and thus be considered ethically acceptable, even if they fail to reach those standards. The need for at least *some* quality medical care might trump concern about absolute explainability, which might be considered by some as an unnecessary luxury. Thus, from the perspective of justice, we encounter a conundrum: should medical AI-DSS that can help people improve expert-scarce regions' healthcare infrastructure be used even though those AI-DSS do not satisfy global explainability standards?

For the following discussion, it is important to keep in mind that discussing the claim to possibly lower local standards of explainability is part of an attempt to provide more and better healthcare overall. The following analysis thus proceeds under the assumption that it would be possible, even with a lower explainability standard, to make quality healthcare available to those who formerly had *no* access to it — but at the price of a lower explainability-standard than the ones implemented in other contexts. We take expanding access to healthcare to be a prima facie duty of ours and remain non-committal about the normative theories that may explain this duty.

This problem displays features that are known from other cases of decision-making under conditions of systemic and structural background injustice (Heilinger, 2020; Young, 2011) where a fully morally satisfying option seems to be unavailable: What would ultimately be morally demanded is to abolish unjust structures that distribute options and resources so unequally. But what should be done under unjust real conditions? Which less than morally perfect steps can and should be undertaken in order to advance the lot of vulnerable or disadvantaged groups, while a more just distribution is not (yet) realized? And: Should not all efforts be directed towards overcoming unjust structures, instead of attempting to provide, with the help of profitable technology, some band-aid solutions that run the risk of perpetuating the problematic status quo (Heilinger, 2022)?

We thus face a hard moral conflict: on the one hand, justifying locally lower standards of explainability of AI-DSS will lead to a number of ethical problems in the explainability of clinical decision-making: such as the general competitive pressure to provide AI-DSS that are less explainable than required, potentially violating the democratic good of explainability (Kempt et al., 2022), or the global decrease of explainability standards caused by a downward-spiral of ever-lower explainability standards. On the other hand, preventing access to at least some urgently needed medical care where otherwise is no care at all seems also morally unacceptable. Thus, in the following, we will seek the morally preferable, rather than a morally ideal solution for this conflict.

While the general structure of this conflict can be found in other instances as well (cf. for example Mitra & Biller-Andorno, 2013), our following line of arguments is specific about explainability as a distinctive challenge. We like to stress that the following

does *not* necessarily hold in the same way for other standards, such as, e.g., accuracy considerations.

3.1 The Relevance of Justice in Healthcare

Health figures are among the essential conditions conducive to a flourishing life and access to adequate healthcare is essential for securing health. Given the importance of both health and healthcare, it has become a major concern for theories of justice to assess how health and access to healthcare is and should be distributed between individuals and groups of individuals (such as formed by countries, ethnicities, gender, and socio-economic status). The pioneering work of Norman Daniels applied John Rawls's *Theory of Justice* (1971) to the realm of healthcare (Daniels, 1985) and health (Daniels, 2008), contributing to a lively debate about justice in health. It is undeniable that from any theory of justice, health, and healthcare constitute key elements for a just society. For instance, proponents of a distributional approach to justice argue that health and consequently healthcare are crucial to equal opportunity and thereby, subject to concerns on *distributive* justice (Daniels, 2008; Rawls, 1971). Similarly, Elisabeth Anderson recognizes health to be one of those goods, necessary for citizens to function as equal members of society, and hence, the provision of healthcare is crucial to the *relational* approach of justice (Anderson, 1999, p.327). From a relational perspective, thus, we owe healthcare to all our fellow citizens. Consequently, healthcare is considered a public good (Anderson, 1999, pp. 330f; Voigt & Wester, 2015).

Different established standards in local healthcare practices often reflect or result from a variety of pre-existing inequalities and may generate new forms of inequalities. For example, the lower the economic means, the fewer there are testing kits for a certain disease and thus, the more stringent usually the requirements will be for selecting patients to get tested. Most areas of the world have achieved some minimal standards for medical practices, like sterilized equipment or the purity grade of medical drugs. When introducing the explainability of medical decision-making as a good provided in medical contexts, e.g., to support informed consensus for some medical procedures, we may also consider these different local resources and acceptability standards for certain features of new technologies.

Distributive inequalities in medical knowledge and resources cause ethically significant inequalities in the quality of care between expert-abundant and expert-scarce regions. As far as these inequalities cause avoidable harm, they count in most approaches to justice as indefensible and thus formulate a constant imperative to minimize these inequalities (Wolff, 2012). Moreover, in practice, the distribution of healthcare seems to be widely recognized to have special priority as equalized access to healthcare has been improved more widely than the equalizing of other material goods (Daniels 2008; UN 2000).

3.1.1 Approaching Justice in Healthcare Diagnostics

To tackle the issue of injustices of access to healthcare diagnostics, we see two options that would improve the situation of people in vulnerable or disadvantaged

regions. These options are not mutually exclusive: first, one could redistribute medical knowledge and resources generally to address the background inequalities; and second, one could reorganize the normative standards of explainability that would allow supplying disadvantaged patients with better access to technology such as AI-DSS.

Redistributing medical knowledge in a fair and just way, while generally more desirable, has some obvious constraints: Promoting expedited education of physicians to work in otherwise expert-scarce areas, or sending a portion of the group of experts from expert-abundant to expert-scarce areas (e.g., the program "Doctors without Borders"), would be an important means to equalize the standards in healthcare across regions. However, it is a time-consuming process depending on political will and coordinated effort. Thus, while remaining the ultimate goal, a full redistribution of medical knowledge, experts, and resources is not a feasible short-term solution for patients in need of immediate help. Consequently, we might need to examine immediate measures, supplementing the redistribution of medical expertise and resources, to improve the situation of expert-scarce regions.

Given the goal of improved care, our normative standards of explainability matter in the context of global health inequalities as there are technological resources that might lower inequalities in healthcare provision. AI-DSS could diminish the inequalities regarding quality of care, by offering (semi-)automated diagnosis and treatment recommendations without the need for much expert-supervision (Wahl et al., 2018). Yet, despite efforts in the field of explainable AI, the decisions of AI-DSS are usually not explainable to the degree human decisions are explainable (see Sect. 2.1). If we suppose a global absolute standard of explainability of AI-DSS, we must conclude that most of these AI-DSS are morally inadmissible even if they could provide some helpful healthcare services to regions without experts. If we suppose global relative standards (relative to the standards of the best human experts globally), we also must conclude that diagnostic AI-DSS are simply not exhibiting the features necessary for moral permissibility.

If someone suggested to lower global absolute standards of explainability for AI-DSS to make these machines permissible after all, we should acknowledge some strong arguments in favor of such suggestion: in the light of the global need for healthcare and without a feasible alternative to provide explanations for diagnoses, making AI-DSS more available has the ability to significantly reduce the unjust inequality to healthcare. Similarly, it applies for lowering global relative standards of explainability.

Such a proposal could still insist on not lowering standards of explainability in areas where experts are abundant, as this change of norms can be seen as limited to AI-DSS: in areas where physicians provide explanations for medical decision-making processes, we can expect them to remain to do so at current levels.

3.2 Rejecting Universal Explainability Standards in Healthcare

So far, we have emphasized the importance of healthcare for different approaches to justice and hence, for normative explainability standards as part of quality

healthcare. In this section, we discuss the different ways of justifying normative explainability standards of medical decision-making. Now, one could argue that these standards should apply generally, and thus reject attempts from expert-scarce regions to choose lower explainability standards to gain access to these AI-DSS which fail to reach the general standards. However, while the decisional discretion in which normative standards of explainability of AI-DSS are sufficient should lie with those communities using these AI-DSS, the difference in need and urge requires a delicate and careful analysis under which conditions these demands can be granted.

Rejecting universal explainability standards does not imply there should not be minimum requirements for explainability. Explainability of medical decision-making, as outlined above, generally contributes to the acceptability of such decisions. As other authors (Grote & Berens 2020; McDougall, 2019) have pointed out, the autonomy of patients is related to the ability to explain diagnoses and procedures to them. AI-decision support systems devoid of any explainable feature would likely violate the autonomy of patients to reasonably assess the risks associated with their diagnosis and the following treatment (see e.g., Kempt & Nagel, 2021 for an analysis of the connection between explainability and responsibility). These deliberations generally take place under the assumption of expert-abundance, in which patients have access to established procedures, reasonable expectations for explanations, and a selection from different options of treatment. Furthermore, lowering explainability standards for AI-DSS seems reasonable only if the corresponding AI-DSS (1) generates advantages that otherwise would remain inaccessible to those benefiting from it, and if (2) increasing the standards straight away is highly improbable, even if in a generally just world it would have to be considered a primary and realistic practical goal.

The moral importance of explainability and its implications are undeniable, as an egalitarian's perspective on normative explainability requires the distribution of knowledge as the ultimate goal in parallel to context-dependent explainability standards. This certainly includes the knowledge of the development, design, and implementation of AI-DSS as well as medical education.[1] Notably, in the following proposal, the redistribution of medical knowledge is not replaced by the provision of AI-DSS but supplemented for the time it takes to equalize both knowledge and resources, sufficiently in order to resolve global inequalities. That is, adapting normative explainability standards may merely diminish global inequalities in healthcare, while only equalizing the established standards seems to constitute a genuine means to global justice in healthcare.

Finally, while explainability standards ought to be context-dependent, we argue that only robust reasons such as the context of expert-scarcity and thus, the availability of care, should be permitted to influence the standards of explainability. We are aware that most often other factors, such as cost considerations, are influencing decisions in the medical sphere. If, e.g., for economic reasons, there are significant

[1] We also acknowledge that context-dependent explainability standards undermine the possibility to establish genuine relational equality, as long as the context itself is not equalized. Context, in this sense, denotes the local established standards of explainability.

inequalities in the explainability of the offered healthcare services among different income brackets, this surely threatens the idea of moral equals and thus undermines the current proposal.

3.2.1 Global Explainability Standards

One could argue to lower the global standards of explainability either to an absolute minimum, possibly met by all, or to the relative minimum of local explainability standards.

On the one hand, the generalization of the maximum relative standard of explainability as a requirement has shown to miss the potential of technologies for expert-scarce regions. On the other hand, setting low but globally absolute explainability standards would potentially unnecessarily threaten the autonomy of patients in expert-scarce regions (as seen in Bjerring & Busch, 2021), having existing local standards that are higher than the proposed global standards (cf. Sect. 3.2.2). Moreover, by setting global standards, the standards are risking to introduce further relational inequalities, as arguments from the perspective of expert-abundant regions may not hold for expert-scarce regions and vice versa. Therefore, global standards, in general, would risk introducing relational inequalities as one region would act on the behalf of the others (Voigt & Wester, 2015).

Additionally, any currently discussed normative explainability standards, either globally absolute or relative standards, are often not met by the executing physicians in expert-scarce regions and even less by the AI-DSS. However, it seems odd to insist that this should be a reason to stop providing healthcare services if local experts cannot live up to these standards and demand the AI-DSS to help their work. Thus, global standards would inherently introduce double standards for human explanations on the one hand and AI-DSS-explanation on the other within the same region. These double standards, however, require strong arguments to be justifiable (Zerilli et al., 2019; Kempt et al., 2022).

Given the likelihood of regional double standards and the implication for relational justice, we suggest exploring local approaches to explainability standards.

3.2.2 Local Explainability Standards

Explainability standards could be set locally. Local absolute standards, i.e., standards applied to the medical AI-based technology, independent of any further contextualization of local needs and abilities, do not seem to us as a reliable way to establish normatively justified procedures. The quality of healthcare is a dynamic feature that grows with political and economic stability and can fade with crises, instability, or catastrophes. Using local absolute explainability misses these dynamics as it determines the necessary explainability of the technologies deployed in those regions at a specific point in time. If a local healthcare administration decided to set a specific normative standard of explainability in decision-making processes and then lost a substantial amount of their workforce due to some political instability, it seems unlikely that they would insist on the previously set standards.

However, if a local healthcare administration decides on allowing lower explainability standards for their needs to administer their services more effectively, they might have some good arguments to introduce these decisions. This is due to local relative standards, as these can be used to argue for a differentiated approach to the local needs determined by the healthcare administration. It would allow for distributing healthcare support that may be urgently needed and that is able to improve the quality of healthcare provided to people.

This, however, may introduce worries about double standards between closely neighboring regions and within regions or administrative districts. If we argue that explainability standards should be allowed to be sensitive towards context-specific requirements, i.e., relative to pre-established local standards of explanations of medical decision-making, then it can happen that explainability standards can vary even within regions, e.g., within individual hospitals (those with better physicians against those with worse ones), or between richer and poorer neighboring regions.

As we will address this issue in our discussion of objections, we will also discuss how the idea can be introduced that allows regions that fail to meet the general ethically acceptable standards (i.e., global relative standards) to argue for local explainability standards.

4 Objections: Avoiding Slippery Slopes

Three objections can be raised against our proposal. They target the distribution of responsibility, the approach's handling of norms, and its potential incentivization and reinforcement of unbalanced power structures.

4.1 Responsibility

The first issue we encounter concerns the issue of responsibility distribution. This concern is focused on the trade-off between expert-scarcity and lower standards of explainability. For one, responsibility concerns for unexplainable technologies have been discussed in the context of informed consensus violations on the side of the physician (Bjerring & Busch, 2021). Thus, lowering explainability standards may be morally impermissible, as those standards would lead to bigger moral problems of operating unexplainable technologies.

We may reject this argument on at least two grounds. First, our distinction of absolute vs. relative explainability shows that explainability ought not to be seen as a mere black-box problem, but rather one of appropriate comparison. Considering further the justified need to differentiate within the relative explainability approach, local communities might as well deliberately decide on accepting these machines in light of their own healthcare needs. Thus, communities may bear the responsibility for operating machines their operators may not be able to explain, as long as the beneficial outcome is weighed against the needs of the particular community and a local and relative sufficiency threshold regarding explainability is respected.

The second rejection concerns those who may provide those technologies in the first place. The fact that some technology may not reach the levels of explainability needed or required for justified use in expert-dense areas while being "marketable" in other areas does not preclude the producers of these technologies from further improving their devices. Some other conditions, i.e., accuracy and reliability of these technologies (London, 2019), ought to be guaranteed, as otherwise the benefit of using these technologies at all may be in question, rendering investigations into their explainability pointless.

Thus, our approach can counter emerging responsibility concerns by distributing the responsibility, on the one hand, towards those deciding the healthcare standards of their respective communities and by not cutting those loose from responsibility that produce and distribute these technologies in the first place.

4.2 Mapping Standards on Abundance

Another argument may be formed from a theoretical objection to our reconstruction. Thus far, we understood expert-scarcity to be the functional absence of reliable human medical expertise, i.e., doctors and other medical professionals, and adjusted our concept of explainability accordingly. However, one may correctly point towards the fact that expertise is not a binary, but a scale. Only looking at expert-dense and expert-scarce areas may suggest that one either has access to human medical expertise or not, but there are vast differences in the quality and accessibility of medical expertise, requiring patients to make all sorts of decisions regarding their own healthcare. Some regions may have a lot of physicians with relatively little expertise, while others may have a relatively low density, but still easy access to medical knowledge through experts.

Moral norms, in contrast to expertise distribution, are usually difficult to reflect on such a scale. Most norms cover a set of cases based on their shared, norm-relevant features, which usually requires these features to be binary (or be present or not present). While there can be norms reflective of a scaled feature ("the more there is of *x*, the more there should be of *y*"), it appears problematic to simply assume this is the case here.

Adjusting explainability standards accordingly may lead to an absurd situation in which one area is deemed adequately staffed with high-level medical expertise, while a neighboring region with fewer experts may be "eligible" for lower explainability standards, simply because accessibility to explainable medical care is comparatively slightly lower.

However, justifying lower medical explainability standards ought to be measured against the accessibility of patients to higher standards. "Locality" as a measure for the justification of setting contextualized standards is not ignorant of other legal, political, and geographical limits. While we can condemn the stark differences in healthcare access and standards along national borders (i.e., between North and South Korea, Finland, and Russia, or Singapore and Malaysia), accessibility as a social reality ought to be accounted for in these considerations. That means,

mapping standards of local relative explainability should account for these borders and realities, as healthcare standards are set by discreetly defined institutions.

4.3 Exploitation Strategies and Post-Colonial Worries

As we have framed the core problem for adopting digital medical systems in lieu of expert-guided systems as one of need, the strongest objections to our proposal must problematize the underlying power structures making these decisions. Thus, our third objection regards the power imbalances and its potential for exploitative consequences of these communities with vulnerable economic and institutional structures. It is not difficult to anticipate that the point of our proposal — taking communities in their own goal setting seriously to allow for local relative explainability standards — can be turned on its head when facing real-life application. As most expert-scarce areas are, at least in the contexts of medical care, highly dependent on and in urgent need of external support, we can infer that these communities will have a weak negotiation position to receive a good outcome. Without a proper reflection on the material difference of those communities to communities with higher expert-density, we may remain ignorant of the power structures behind healthcare decisions and the different perspectives global agents can have on such decisions.

4.3.1 Privacy

The first currently applicable concern of allowing the provision of unfinished healthcare is the business model that may emerge here, as one dimension of big data-driven AI-DSS, is that they improve with increased data (generally speaking). This means that allowing companies to export and use their unexplained (and, thus, from a global standard's perspective unfinished) products may lead to privacy issues. Trading a nation's healthcare data for the promises of improved healthcare for its citizens may be too good an offer for governments to pass on (this debate has extensive precedence in other bioethical debates, e.g., drug studies).

In many cases this issue is related to the one of post-colonial power imbalances, but ought to be reflected upon separately, as it can also be framed of a conflict of interest between big corporations on the one side (to get access to as much health data as possible) and any community's interest in protecting its citizens' privacy. As they both share the same overall goal, i.e., the improvement of AI-DSS to diagnose and treat patients, the emerging conflict of interests is merely one of the means to get there.

However, we suggest aiming for contracts between nations and AI-companies that protect the privacy of the citizens while allowing corporations to train their AI-DSS. Several of these data and privacy protection frameworks exist and have been applied to other data collection schemes.

While health data is especially worthy of privacy protections, and we can assume that communities agreeing to use lower local and relative explainable AI-DSS due to expert-scarcity come from a position of need, we argue that these privacy protection frameworks should be interpreted as strictly as possible. This can be done by

granting use-rights to companies for the data only to improve the AI-DSS in question. This should be a cornerstone of any kind of agreement between companies and communities in need of AI-DSS lacking explainability of its decision-making processes, and not a merely preferred outcome of negotiations. This is needed to avoid the emergence of market logics for health data, in which one country or community could "sell out" its health data cheaper than the neighboring country, and thus offering companies the chance to pit nations against each other in a race to the bottom.

4.3.2 Dependencies

Next to the privacy-concerns encountered in other types of debates, such as drug- and other field studies, the concern of creating dependencies can also become a major issue in the future development of healthcare, both as a consequence of the aforementioned privacy issue but also as an independently emerging issue.

As AI-DSS-based healthcare technologies' progress is built on accumulating healthcare data to train the AI, those who harvest and process the data are the ones in the position to improve on their products. Thus, those who obtain the data are in the best position to improve healthcare services, while those who buy and use the AI-DSS will not be able to develop and improve upon their own services. The argument can be made that the earlier a community gains access to AI-DSS without the proper technological infrastructure, the faster the standards of healthcare access rise without the specific communities to sustain these standards on their own. In contexts of expert-scarcity, this may even increase the dependency of expert-scarce communities on AI-DSS and their development, ultimately replacing one problematic situation with another without offering a long-term solution.

As our approach does not claim that local relative explainability is preventing dependencies, it is worth considering what needs to be done to avoid these dependencies to emerge. However, this cannot be done here.

4.3.3 Post-colonial Power Structures Reinforced

Many of the inequalities in access to healthcare can be productively viewed through the socio-historical perspective of post-colonialism, in which the oppressive structure of inequality is perpetuated under the guise of aid and support from former colonizing communities to former colonized ones.

One argument against the proposal under discussion in this paper, namely to allow for local relative standards of explainability, could be that it is merely reflective of such post-colonial mindset. In lowering the standards of explainability to provide aid to expert-scarce areas (presumably mostly found in the global south), the worry must be that (a) that privileged or rich nations provide low quality, techno-solutionist help that is intended to replace actual, long-term, and subsidiary aid, like the reduction of expert-scarcity, and (b) that these richer nations, despite the insight in the necessity and utility of such aid, are not working nearly as diligently on developing more reliable, explainable, and precise AI-DSS. Instead, the concern may be that these technologies are merely unusable, insufficient approaches to technology

for the elevated moral standards of the rich nations, but good enough for the poorer ones, akin to discarded second rate items.

However, the proposal under discussion in this paper does not contend that the AI-DSS in question should be viewed as such second-rate quality technology, and thus richer nations cannot claim themselves sufficiently engaged in minimizing inequalities when they provide these technologies to expert-scarce areas. We acknowledge that "providing" AI-DSS with local standards can merely be an offer that always ought to be accompanied by the continued efforts to sustainably lower the expert-scarcity.

Furthermore, this worry of post-colonial power structure reinforcement has to assume that the improvement of AI-DSS to be more explainable is not pursued with the necessary means. However, as several studies have suggested, we may not be able to provide the aimed-for explainability standards necessary to suffice for relative explainability in the first place (Ghassemi et al., 2021). Thus, withholding this technology to expert-scarce areas may mean withholding this technology for the foreseeable future. We may not be discarding second-rate technology, but technology that is as good as it gets in explaining its decision-making process. Ironically, then, it is not the lowering of standards for expert-scarce areas, but having too high standards for expert-dense areas that prevents reducing this inequality.

5 Concluding Remarks

Despite considerations of justice being central to the provision and distribution of healthcare, the influence of these considerations on the ethical debates about medical AI has been marginal. Decontextualized debates surrounding the normative relevance of explainability of AI systems require a connection to the needs and demands of healthcare-deprived populations. As the redistribution of technologies is a feasible short-term solution to diminish inequalities in global healthcare, the duty to redistribute these technologies ought to be measured against the permissibility of global explainability standards. We thus propose to consider these normative explainability standards to be context-dependent, i.e., relative to the locally established explainability standards of human expert diagnoses. This way, the differences in availability of quality healthcare can be accounted for and adjusted locally according to the preferences and needs of vulnerable populations without imposing unready or denying sufficiently ready healthcare devices.

We are well aware that technological solutions alone do not constitute a sufficient remedy to the existing inequalities in healthcare. Technology alone will not fix healthcare dependencies and requires constant reflection of its just distribution. However, acknowledging the moral importance of explainability in healthcare decision-making as generally desirable provides at least a starting point in weighing the need for global explainability standards on the one hand, and the specific needs of populations relative to their access to quality healthcare on the other hand. Introducing local relative explainability standards allows for redistribution without erasing the normative imperative to do more, such as initiate educational programs to improve the general expert-scarcity.

The arguments and conclusions that have been made throughout this paper further point out the relevance of justice considerations in the ethical discourse on explainable AI and AI-DSS in healthcare and elsewhere.

Acknowledgements We would like to thank Jan-Christoph Heilinger, Niël Conradie, Peter Königs, William Jared Parmer, Chaewon Jun, Fabian Kießling, Horst Hahn, Bianca Hoffmann, Markus Wenzel, and Helen Heinrichs for discussions on some key concepts. Lastly, we would like to thank the reviewers and editors of the journal for their helpful comments and leading the process.

Author Contribution All the authors contributed to the main idea of the manuscript. HK and NF wrote the main draft of the manuscript. SKN commented extensively on the final draft.

Funding Open Access funding enabled and organized by Projekt DEAL. 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.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication All the authors have agreed to the publication of this draft.

Competing Interests The authors declare no competing interests.

References

Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., ... & Z-Inspection initiative. (2022). To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, *1*(2), e0000016.

Anderson, E. S. (1999). What is the point of equality? *Ethics, 109*(2), 287–337.

Andreotta, A. J., Kirkham, N., & Rizzi, M. (2021). AI, big data, and the future of consent. *Ai & Society*, 1–14.

Arbelaez Ossa, L., Starke, G., Lorenzini, G., Vogt, J. E., Shaw, D. M., & Elger, B. S. (2022). Re-focusing explainability in medicine. *Digital. Health, 8*, 20552076221074490.

Azzopardi-Muscat, N., & Sørensen, K. (2019). Towards an equitable digital public health era: Promoting equity through a health literacy perspective. *European journal of public health*, *29*(Supplement_3), 13–17.

Beauchamp, T., & Childress, J. F. (1989). *Principles of biomedical ethics*. Oxford University Press.

Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology, 34*(2), 349–371.

Brall, C., Schröder-Bäck, P., & Maeckelberghe, E. (2019). Ethical aspects of digital health from a justice point of view. *European journal of public health*, *29*(Supplement_3), 18–22.
Daniels, N. (1985). *Just health care*. Cambridge University Press.
Daniels, N. (2008). *Just health: Meeting health needs fairly*. Cambridge University Press.
Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics, 47*(5), 329–335.
Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1
Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750.
Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press.
Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*(3), 205–211.
Heilinger, J.-C. (2020). *Cosmopolitan responsibility*. De Gruyter.
Heilinger, J.-C. (2022). The ethics of AI ethics. A constructive critique. In: *Philosophy & Technology* (online first).
Kempt, H., Heilinger, J.-C., Nagel, S.K. (2022). Relative explainability and double standards in medical decision making. *Ethics and Information Technology.*
Kempt, H., & Nagel, S. K. (2021). Responsibility, second opinions and peer-disagreement: Ethical and epistemological challenges of using AI in clinical diagnostic contexts. *Journal of Medical Ethics, 48*, 222–229.
Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57.
London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report, 49*(1), 15–21.
McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics, 45*(3), 156–160.
Mitra, A. G., & Biller-Andorno, N. (2013). Vulnerability and exploitation in a globalized world. *IJFAB: International Journal of Feminist Approaches to Bioethics*, *6*(1), 91–102.
Penu, Obed Kwame Adzaku; Boateng, Richard; and Owusu, Acheampong. (2021). Towards explainable AI(xAI): Determining the factors for firms' adoption and use of xAI in Sub-Saharan Africa. AMCIS 2021 TREOs. 35.
Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics – A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine, 107*, 101901. https://doi.org/10.1016/j.artmed.2020.101901
Rawls, J. (1971). *A theory of justice*. Harvard University Press.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
United Nations Committee on Economic, Social, and Cultural Rights (UN-CESCR) (2000). General Comment No. 14: The right to the highest attainable standard of health. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G00/439/34/PDF/G0043934.pdf?OpenElement. Accessed 21 Nov 2022
Voigt, K., & Wester, G. (2015). Relational equality and health. *Social Philosophy and Policy, 31*(2), 204–229.
Wadden, J. J. (2022). Defining the undefinable: The black box problem in healthcare artificial intelligence. *Journal of Medical Ethics, 48*(10), 764–768.
Wahl, B., Cossy-Gantner, A., Germann, S., & Schwalbe, N. R. (2018). Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings? *BMJ Global Health, 3*(4), e000798.
Wolff, J. (2012). The demands of the human right to health. *Aristotelian Society Supplementary, 86*(1), 217–237. https://doi.org/10.1111/j.1467-8349.2012.00215.x
Young, I. M. (2011). *Responsibility for justice*. Oxford University Press.
Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology, 34,* 265–288. https://doi.org/10.1007/s13347-019-00382-7
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.