Original research article

BIG DATA & SOCIETY

Big Data & Society
October–December: 1–13

DOI: 10.1177/20539517251386037
journals.sagepub.com/home/bds

Sage

When is black-box AI justifiable to use in healthcare?

Sinead Prince[1] **and Julian Savulescu**[1,2,3]

Abstract

Although it is reasonable and valuable to seek explanations for decisions made by artificial intelligence (AI), it is simply not possible with black-box AI algorithms. However, these algorithms can produce highly beneficial and efficient outputs that could be extremely useful to patients, treating teams, hospitals, and funding bodies. This poses a dilemma: is black-box AI justifiable to use in healthcare? This article analyses the normative reasons that can defend and justify the use of black-box AI in healthcare; this analysis includes, but does not give lexical priority to, explainability. This is pertinent given the current prohibitions of black-box AI in healthcare, such as in Australia. This article defines justifiability as decisions based on robust reasons and thus identifies reasons that can justify the use of black-box AI in healthcare. These include the algorithms' explainability and accuracy, the seriousness of the decision's consequences, any relevant bias, the context of the decision, and the level of human intervention. We argue that whilst each of these separate considerations is important, only accuracy and reliability are necessary, and to be sufficient, it is likely that some further reasons arising from the nature and context of the decision will be required.

Keywords

Artificial intelligence, explainability, bioethics, justifiability, transparency

Introduction

The prohibition or restriction of the use of black-box artificial intelligence ('AI') in healthcare, namely AI that lacks causal explanations, through regulation is not uncommon. For example, the Australian Therapeutics Good Administration (TGA) (2024) defines black-box AI as insufficiently transparent, and therefore prohibited from receiving regulatory approval for use in healthcare. The *European Union Artificial Intelligence Act 2024*, Chapter III Article 13 requires high-risk AI systems (defined to include AI systems that pose a significant risk to health) to be 'sufficiently transparent' as opposed to explainable, but Chapter IX Article 86 provides persons subject to a black-box AI decision with the right to an explanation of the role of the AI system in the decision- making procedure. The literature is also unresolved on the issue. Freyer et al. (2024) conducted a systematic review of the reasons and motivations for and against the use of explainable AI in medical decision making, and found that the debate is still unsettled – 27 papers argued against requiring explainability but 17 argued for explainability requirements. Many authors have claimed that explainability is necessary for AI use in healthcare (Babic and Cohen, 2023; Bjerring and Busch, 2021; Rueda et al., 2024). The regulation of black-box AI in healthcare is thus in tension with the benefits that black-box AI could provide to patients and healthcare providers, and the debate on resolving this tension is unsettled.

We believe that blanket prohibitions on black-box AI in healthcare are excessive, and the need for and type of explanation required for ethical use of black-box AI in healthcare is contextual and relative to other reasons for using it in healthcare. That is, the justifiability of black-box AI ought not to be reduced to mere explainability; the moral justifiability of medical black-box AI is contingent on many factors. Whilst explainability importantly contributes to transparency, accountability, and trustworthiness (Hamon et al., 2020; Rudin, 2019; Rueda et al., 2024), black-box AI algorithms are valuable *for the very reason* they aren't explainable: they can process exponentially more data than humans and can make significantly more accurate

[1]Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[2]The Royal Children's Hospital Melbourne, Murdoch Children's Research Institute, Melbourne, Australia
[3]Uehiro Oxford Institute, University of Oxford, Oxford, UK

Corresponding author:
Julian Savulescu, Uehiro Oxford Institute, University of Oxford, 16-17 Saint Ebbe's St, Oxford OX1 1PT, United Kingdom.
Email: julian.savulescu@uehiro.ox.ac.uk

predictions (Ding et al., 2022). For example, in diagnostics such as medical imaging analysis and COVID-19 chest CT scans (Ding et al., 2022; Kriza et al., 2021), predicting future hospitalisations using patient's electronic health record, essential hypertension, and lower respiratory disease (Zhang et al., 2018), predicting patient mortality during an ICU admission (Shickel et al., 2019), and speed diagnosing diabetic retinopathy (Senapati et al., 2024). Some black-box AI is as equally accurate as clinicians, such as in mammogram screening (McKinney et al., 2020), classifying skin cancers (Esteva et al., 2017), and ocular imaging for eye diseases (Ting et al., 2019). Some models even have enhanced diagnostic accuracy (Mota et al., 2024), see for example, Microsoft's AI Diagnostic Orchestrator (MAI-DxO), which is four times more accurate than experienced physicians (who lacked access to colleagues, textbooks or chatbots) (Nori et al., 2025).

Following the recent and consistently improving research on the beneficial nature of healthcare black-box AI, we will claim that the lexical priority afforded to causal explainability as necessary for the ethical and lawful uses of black-box AI warrants investigation. This article seeks to draw this literature together and make an original contribution by analysing how the various object-given reasons and considerations of using black-box AI in healthcare interplay. We want to understand what considerations are at play that make explainability necessary or unnecessary. This article is therefore not a literature review of current leading considerations of the use of black-box AI in healthcare, but a normative analysis for understanding how such considerations intersect and potentially extenuate each other according to the contextual use of an algorithm. In Section 'What is a justification?', we will therefore define and defend an account of justifiability by which to measure whether black-box AI has met a sufficient ethical threshold. Section 'What reasons are relevant in justifying black-box AI?' will highlight varying reasons that can be used to justify black-box AI, namely, the role of AI explainability and accuracy, the seriousness of the patient's condition, the role of human intervention, the effect of bias, and the nature of the context. In Section 'Balance and justifiability', we will provide a summary of how we can foreseeably use these casuistic lessons drawn from case studies to explain when and how to justify the use of black-box AI in healthcare.

What is a justification?

Determining whether it is justifiable to use black-box AI in healthcare requires us to define justifiability. Broadly defined, justifiable decisions are those with robust *reasons* to pursue them: they are decisions that are 'just, right, desirable or reasonable' (Malgieri and Pasquale, 2022). We want to know whether a decision is rooted in object-given reasons (Parfit, 2011), and whether these object-given reasons are the strongest reasons at play, that is, whether they are not outweighed or extenuated by other reasons. For example, simply because a person can *explain* the reason for which they came to a decision, doesn't mean the decision was a good one.

However, justifications are contextual depending on the person or situation for *whom* the justification is required. For example, some philosophers may defend a decision on whether there are sufficient utilitarian reasons for it, whereas some technicians may justify the use of black-box AI based on reliability and safety in achieving a particular healthcare outcome. Similarly, patients may justify choosing a healthcare decision as against their own values and healthcare goals. The justifiability of black-box AI in healthcare therefore ought not to ignore the various definitions, perspectives, reasons, and values of those involved in developing, implementing, and using AI decisions.

In our attempt, therefore, to assess the justifiability of black-box AI in healthcare, we will adopt a *non-principled approach* for determining the threshold of justifiability. Also known as a pluralistic approach, we will methodically analyse the use of black-box AI in a variety of healthcare decisions using commonsense morality (Sidgwick, 1874). We consider the reasons purported to justify the use of black-box AI in healthcare, and bring them into reflective equilibrium with our understandings of healthcare and bioethics by modifying the lexical priority of principles and developing coherence between the relevant considerations (Daniels, 2003; Rawls, 1971). As such, we are concerned with identifying normative reasons to use black-box AI in healthcare. We will consider when normative reasons have greater lexical weight in certain situations, which factors may be necessary or sufficient for justification, and when a particular factor may have an extenuating impact on justifiability. We do not endorse a single approach or normative framework such as utilitarianism or deontology, rather, given the sheer diversity of medical decisions, we argue that the ethics of AI in healthcare ought to be managed through a bottom-up approach, with justifiability drawn from the context of the decision, rather than abstractedly from a top-down approach. As such, much of our analysis and normative argument is casuistic, drawing on real world problems in medical decision making and the leading concerns or benefits drawn from the literature. That is, the justifiability of the use of black-box medical AI cannot be normatively dominated by any singular approach but must be sensitive to a variety of reasons and contexts.

What reasons are relevant in justifying black-box AI?

Explainability

Given the substantial literature on the importance of explanations in justifying black-box AI, we should begin with

understanding the nature, role, and limits of explanations in justifications. Explanations are generally viewed as both intrinsically valuable as an epistemic good and instrumentally valuable for achieving transparency (Bjerring and Busch, 2021; Ordish et al., 2020). This is first because if we can understand why a prediction was made, the model will be seen as trustworthy and thus used more widely in healthcare (Balasubramaniam et al., 2023; Bjerring and Busch, 2021; Hois et al., 2019). Second, transparency protects the rule of law by ensuring accountability for decisions that are not consistent with public norms of justice (e.g. bias) or for wrongful harm (Binns, 2018; Kempt et al., 2022a; Maclure, 2021; Steging et al., 2021). C*eteris paribus,* if we could choose between two AI algorithms, we should choose the one that delivers results which can be explained. For this reason, leading ethical guidelines require explainability and transparency in AI system deployment (Jobin et al., 2019).

However, providing *any* explanation for a decision does not necessarily make the decision justified. This is for two reasons. First, people can make bad choices despite being able to explain them (Alvarez, 2017). In the same way, explanations themselves are not intrinsic to the justification of a *good* decision; explanations are merely the expression of the reasons used to make a decision (whether good or bad). The value of an explanation instead lies in helping to illuminate the intentions of the decision-maker, identify when bad decisions are contributing to poor healthcare outcomes, and pinpoint who is accountable for bad outcomes.

This leads us, however, to a more important point: there are various definitions and standards of what is required by an explanation. This is because explainability is a broad concept – it is not solely whether we can explain the *causal* or *motivating* reasoning for a decision (Alvarez, 2017; Raz, 2011). For example, when regulators expect explainability, they are broadly asking for AI algorithms to explain 'the reason, rationale, or predictive factor in that particular outcome' (London, 2019). Conversely, scientific AI literature defines explainability according to what can be reasonably or legitimately expected of it (Hamon et al., 2020), its purpose (Balasubramaniam et al., 2023), or whether it can be rationalised or simulated (Amann et al., 2020). Developers might seek explanations of algorithmic decision making through understanding pixel activation (Mersha et al., 2024).

But within healthcare there are different stakeholders seeking explanations of different tools and phenomena for different reasons (Durán, 2021; Kempt et al., 2022a; Malgieri and Pasquale, 2022; Ordish et al., 2020). For example, within a single context involving a patient determining which treatment option to follow, the patient may want explanations of success rates, side-effects, and standards versus alternative treatments; the doctor may want an explanation of which symptoms are indicative of a specific diagnosis; and, regulatory approval bodies may want explanations of the risks, costs, benefits, and side effects of treatments for patients populations.

Evidently, the definition and threshold of a satisfactory explanation is not a one-size-fits-all. The users and beneficiaries of medical AI are a distinct community with distinct needs and as such, various explanations may be justifiable for various stakeholders. The threshold explanation required to justify an algorithmic decision *varies* according to the intended recipient. The *causal* or motivating explanation is merely one kind of explanation, and is not always equal in value to other explanations (Balasubramaniam et al., 2023; Hamon et al., 2020). Let us consider a case study: *Chronic Pain.*

A black-box AI device suggests a pain relief treatment plan for a woman who has been suffering chronic pelvic pain for 10 years. The algorithm proposes treatment plans with medication that reduce patient pain and has been tested through clinical trials with an accuracy rate in successfully and substantially reducing the pain of 70% of patients. The device cannot explain why the medication is successful. However, the doctor disagrees with the recommendation and provides a different treatment option that has moderate certainty (50% chance) but has an explanation. The doctor asks the patient which treatment she would like to try.

In this case study, there is no explanation for the patient's chronic pain, but the black-box AI model is more likely to reduce patient pain than the doctor's professional opinion. There are several relevant facts to highlight from this model that give us reasons to justify its use in these circumstances.

First, the doctor provides the patient the option to decide *for herself* whether to follow their advice, or the AI advice. This is important because the patient has been enduring the condition for 10 years – her narrative and lived experience of evidently lacking answers and effective treatment options might make her value an explanation over more successful treatment, but it also might give her reason to value effective pain relief over an explainable option. Second, the decision is particularly clinical, in that patients often are not involved in the clinical diagnostic problem solving or identifying which treatments are clinically viable: patients generally only become involved in determining which of the recommended treatments to pursue. Finally, presenting the option between using AI-derived or human-produced options to the patient involves the patient in her own decision making, and is consistent with professional guidelines and laws around informed consent. In this case study, the patient has been *informed* of the missing explanation, and this allows her to weigh the benefits and risks of treatments against her own values to provide *informed* consent. Thus, the lack of a causal or motivating reason by the AI does not alone render the use of the AI in this context unjustified.

An objection might be made here by pointing to the regulatory or accountability requirement for an explanation in medical decision making, particularly for ensuring the decision was just and fair. For example, a patient might be subject to a black-box diagnosis and mistakenly *not* diagnosed with breast cancer. Causal explanations help to justify decisions by ruling out the possibility that unjust

reasons were used, such as failure by the AI developers to train the AI on racially diverse samples. This is a legitimate concern; however, it is not so significant as to make *causal* explainability necessary for justifiability as opposed to other possible explanations. For example, AI developers have begun producing Explainable AI ('XAI') or 'white-box' models that use secondary algorithms to reverse engineer the original black-box model decision and provide rationalisations for why this might have happened (Babic and Cohen, 2023). Alternatively, doctors can take black-box AI decisions and make their own assessment as to whether the decision is justifiable and consistent with their own medical expertise (Babic and Cohen, 2023; Hadfield, 2021). Consider the following case study, *Heart Attack*:

A black-box AI device predicts that a patient has a 70% chance of suffering a heart attack in the next 3 months and to immediately begin taking beta blockers. The doctor recommends this advice to the patient and explains that such a diagnosis is justifiable because in recent months the patient is a middle-aged woman who has complained of neck and jaw pain, upper abdomen discomfort, ankle swelling, and wheezing. The patient then evaluates this advice against her values and priorities and agrees to take the medication.

In the presence of this '"good" chain of reasoning', in which the doctor considers how the factors might have been weighed and measures this against professional standards of care, the decision is not only justifiable to the patient, but if adopted, will also have be justified *by* the patient against their own values (Muralidharan et al., 2024). This satisfies reasonable expectations that doctors exercise their own judgment and so the lack of *causal* explainability does not render the decision unjustified. We can therefore still have justifiable explanations in two ways: first, the doctors justify the decision against medical expertise; and second, the patient can assess it against their own values and justify it to themselves.

Objections have been made to the use of XAI or post-hoc rationalisations of black-box AI on the grounds that they are 'not faithful' (Afnan et al., 2021), 'misleading' (Rudin, 2019), 'inherently insincere' (Babic and Cohen, 2023), and 'fool's gold' for giving false trust, false impressions, and false confidence (Babic et al., 2021; Peters, 2023; Rudin, 2019). Some of these concerns can be alleviated with modifying the terminology, such as by labelling these as 'interpretative' rather than 'explanatory'. A better response, however, is to highlight the unique nature and purpose of medical knowledge and care that renders uncertainties in knowledge secondary in importance to reliable and beneficial outcomes. Modern medical knowledge is wide, complex, and deep – no one person holds, or could ever possibly hold, all that is known (Ferreira, 2021). Cohen (2020) argues that doctors are 'likely quite ignorant of the underlying trial design or results that led the FDA to believe that the drug was safe and effective, but her knowledge that it has been FDA-approved supplies the necessary epistemic warrant' to justify recommending the treatment. Similarly, professional standards do not require physicians to explain technologies such as MRI machines despite relying on these devices for significant decisions (Sand et al., 2021). Inscrutability of some expert decision making is inherent in such justifiable divisions of labour, and we cannot expect all practitioners to equally scrutinise all medical decisions.

One difference, however, between AI and the division of labour in medical decision making, is that the causal reason can be known by the latter but not in the former (Bjerring and Busch, 2021; Ferreira, 2021). Kempt et al. (2022b) argue that whilst patients generally do not care about technology specifics, such as in diagnostics or details, 'the point of the conditional good of explainability is its potentiality.' If a doctor cannot explain how exactly a new machine works, the researcher, technician, or developer can. In comparison, this is not similarly possible with black-box AI.

However, not all explanations are known for medical phenomena. The opacity in medical decision making is far more routine in medicine than critics realise (Kawamleh, 2023; London, 2019). Medical research and patient treatment planning are often heuristic and medical certification processes are primarily justified through evidence of safety and efficacy rather than causal explainability (Kempt et al., 2022b). For example, paracetamol is causally unexplainable but is available over the counter to the public because of how low-risk, beneficial, and useful it is for a variety of conditions. Shavhasi (2016) argues that what justifies unexplainable but useful treatments is that we can make hypotheses as to their causality that are consistent with accepted bodies of scientific knowledge. London (2019) thus argues that:

'As counterintuitive and unappealing as it may be, the opacity, independence from an explicit domain model, and lack of causal insight associated with some of the most powerful machine learning approaches are not radically different from routine aspects of medical decision-making.'

This is likely because, when all else is equal, we care more about saving a life, treating disease, and reducing pain than providing *causal* explanations for our practices, particularly when other types of explanations or justifications are available. So whilst understanding causal explanations *improves* medical practices through research, we *justify* medical decisions because we know that it works (London, 2019). Arguably, therefore, we can similarly justify accurate and reliable black-box AI through *post-hoc* explanations rather than through causal explanations. This is because post-hoc rationalisations and justifications enable us to hold someone accountable for the final decision, as well as ensuring that the AI decision is consistent with good reasoning rather than unfair or biased decision making. Casual explainability is therefore neither sufficient nor necessary for justifying the use of black-box AI in medical decision making when post-hoc rationalisations are made, shared decision making is practiced, and the AI model is accurate and beneficial.

Accuracy

A widely accepted compromise across AI, legal, and ethical literature is that if transparency must be sacrificed, it should at least be for improved performance (London, 2019; Maclure, 2021). Some argue that in the absence of explanations, the success rate expected of medical AI ought to be much higher than professional standards (Sand et al., 2021). That is, medical AI ought to be clinically proven through reliable testing to produce results as well as, or better than, physicians. The accuracy or reliability of the black-box model entirely depends on the rigour of the clinical trial process, as well as the kind of data collected (e.g. according to race, sex, and age).

Fortunately, medical AI is proving to be highly accurate in certain contexts (Mota et al., 2024; Rodriguez-Ruiz et al., 2019; Ting et al., 2019; Zeltzer et al., 2023). This is due to the increased availability of data from which AI can make decisions, and the increased capacity of AI to process mass data, and to learn and adapt in real time. This increased accuracy in comparison to standard diagnostic testing and treatment predictions is intrinsically valuable in medical decision making. The overarching goal of medical care is after all, to improve patient well-being specifically by addressing their needs with medical care.[1] That black-box AI can be a highly reliable and accurate decision-making tool renders its use important and perhaps required in certain contexts. We are willing to go so far as to say that being as accurate in diagnostic testing or reliable in reducing patient symptoms as doctors is *necessary* to justify the use of black-box AI in healthcare decision making. Nevertheless, the accuracy of black-box AI is not itself sufficient for justifiability. There are several reasons for which increased accuracy may not be sufficient to justify the use of black-box AI.

One reason is if an explanation would increase a patient's well-being due to the nature of the condition or treatment history. For example, the average diagnostic delay for endometriosis is 6.6 years (but can be up to 27 years in the UK) (Fryer et al., 2024). Women also report general bias and prejudice against them from the medical community, specifically, they experience epistemic distrust against them as knowledge-holders of their chronic pain and symptoms (Jackson, 2019). In such cases of systematic opacity that undermines their autonomy, some patients may prefer explanations to opacity even at the cost of slightly or even moderately improved accuracy.[2] For example, in our above *Chronic Pain* case, it is less clear which treatment is most justifiable without recourse to the patient's values: the value of explainability may provide relief to the uncertainty of her past and future, and this psychological relief may be more valuable than that of increased accuracy of treatment. As such, it is not so clear that the superior accuracy of black-box AI is *sufficient* to justify its use: sometimes explainability may be of more value to the patient.

We could make similar arguments about some diseases like acne that are persistent, ongoing, non-life threatening, and in which patients have often tried a range of methods to treat it, such as prescriptions and self-treatment through facial products. Indeed, acne treatments are commodified and arguably exploited by pharmaceutical companies with patients spending hundreds on products with no evidence or information whether it is targeted toward their specific kind of acne or suited to their skin type, mainly because patients generally do not understand the causes and corresponding treatments of acne (Tan et al., 2001). A patient may have been promised treatments by practitioners, and none of them have worked. Even if black-box AI provides a highly accurate answer, if we cannot explain it the patient may lack trust in the unexplainable decision. As such, a patient may prefer a causal explanation for a recommended treatment of their acne with slightly or moderately lower accuracy, over an unexplainable but slightly higher accurate treatment. Research also indicates that patients are *more* likely to follow treatment plans if the doctor answers all of the patient questions (Chan, 2023). As such, the benefits of patient commitment to treatment plans (particularly long treatment plans such as acne medication) may outweigh the benefits of some increases to accuracy.[3] Whilst, the equivalent or increased accuracy of a black-box model to achieve its intended outcome is *necessary* to justifying its use, accuracy alone is insufficient for justification. We shall now consider a variety of medical circumstances that, together with conditions of accuracy and reliability, can be sufficient to justify the use of black-box AI in medical decision making.

Seriousness and urgency

An important consideration in determining whether there is a robust reason to justify the use of black-box AI is the urgency of a medical situation (London, 2019). Babic et al. (2021) acknowledge that prediction explanations might not be so necessary if the algorithm is consistently shown to be more accurate in situations with higher and serious risks. However, some scholars make claims that the more serious the implication, the more reason to ensure supervisory efforts and human involvement (Ding et al., 2022). This might be because the more serious the risk of harm, the greater the need for an explanation if it goes wrong, or because increased seriousness requires greater trust from the patient and such trust requires transparency (Hois et al., 2019).

A better understanding of this condition, therefore, is to discern between *seriousness* and *urgency,* and how these sub-conditions give us reasons to justify using black-box AI in healthcare. Seriousness refers to the risk and gradient of harm arising from the condition. Urgency refers to the time sensitive nature of the decision being made. Not all serious conditions are urgent, and not all urgent conditions

are serious. For example, Huntington's disease is an adult-onset condition in which patients live 15–20 years after the onset of the disease: it is serious but not necessarily urgent. However, there is no cure, and so the potential recommendation by a black-box AI model to curtail the symptoms, although unexplainable, may be a good reason to pursue AI-recommended decisions.

However, this 'last resort' option does not apply to all contexts. For example, a person recently diagnosed with Huntington's disease may prefer to try standard or recommended treatment options as suggested by their doctor. Whether a condition is serious does not necessitate object-given reasons to use black-box AI; instead, the seriousness of a condition impacts the justifiability of black-box AI in two ways. First, if black-box AI is being used to *diagnose* a more serious condition, doctors are more justified in using black-box AI insofar as they are justified in using other highly reliable machines that provide ways to assist in diagnoses incapable of being achieved by humans. But (and second), if black-box AI is being used to *propose a treatment plan*, the patient ought to be engaged through shared decision making to choose their treatment plan. That is, the patient ought to be aware of the use of black-box AI, consent to its use, and still engage with the doctor to (as best they can) self-determine their treatment plan. The seriousness of a patient's condition renders the need for increased accuracy (to prevent serious consequences), and this may in turn justify or even require the use of more accurate AI devices. However, more serious conditions carry more serious risks, and so decisions that are normative in nature, such as treatment planning, will require patient involvement to respect the significant impact of healthcare decisions on the patient's life.

Urgency may also strengthen the reason to use black-box AI, particularly in serious cases. In life-threatening emergency cases, getting informed consent from patients, and by extension, spending time explaining the details, or even the basics, of treatment options, is generally not required (Boyle and Stepanov, 2021). The justification for not requiring consent from patients or spending time explaining and informing the patients of the relevant options or recommended treatments is to allow practitioners to provide the best medical care without being constrained by regulations that would cause medical harm due to delays. Arguably, in the context of emergency healthcare, the concerns about inexplicability or not informing patients about the use of AI would therefore not be as concerning or different from current ethical practice in healthcare. Indeed, it would be inconsistent with current legal and practice standards to exclude the use of highly accurate medical aids in the emergency room because they cannot provide causal explanations.[4] We are not excluding the possibility of justifying the use of medical black-box AI in non-urgent situations, but arguing that other reasons may be necessary to justify its use. Insofar as a situation is both urgent *and* serious (and accurate), the use of accurate and reliable black-box AI is likely generally justifiable.

Context of the decision

The nature or context of the decision can also impact the justifiability of using black-box AI in medical decision making (Babic and Cohen, 2023; Kempt et al., 2022a; Ordish et al., 2020). Some medical decisions are particularly well suited to black-box AI (not all decisions are made inherently better by an increase in data). For example, diagnostic and prognostic decisions must be rooted in empirical data and patterns, and so long as the initial algorithm is trained and approved by relevant regulatory authorities, these models can meet and even supersede human decision-makers due to their capacity to (efficiently) process significantly more data. As such, the more data computable by the decision-maker, the more accurate the decision becomes (up to a point). However, not all types of healthcare decisions should be made by unexplainable decision-makers. To illustrate this claim, we will consider how the context of a black-box AI decision can vary the nature of justifiability despite all involving distributive justice issues.

Liver transplantation allocations are difficult decisions to make due to the complexity of calculating success rates, the limited number of donors, and the difficulty of fairly distributing organs according to normative values. If a black-box AI model is used to determine which patients should receive an organ, a patient who is rejected or ranked lower than others may wish to know why to ensure that it was done fairly (Rueda et al., 2024). In this context, in which interpersonal comparisons and highly normative valuations are being made to determine who will live according to ethical theories of justice, we have strong reason to be concerned about irrelevant factors that may impact decision making. Unlike medical diagnosis and prognosis predictions which are based in scientific data and patient history, organ distribution is a question based on predictive accuracy *and* normative values of how to fairly distribute organs. It is these kinds of *inherently normative* decisions that demand explanations to ensure such norms are just. It is also likely that a patient who is not provided a liver transplant will want a review of the decision-making process and be informed of the causal and motivating reason for their rejection (Robbins, 2019) – studies show people are generally concerned about the role of black-box AI in liver donation decision making (Drezga-Kleiminger et al., 2023). We might compare this to ordinary treatment predictions, in which it would be odd or at least uncommon for a patient to question whether they were recommended a particular medication for normative, rather than clinical, reasons. Our claim, therefore, is that decisions that are *inherently* normative may not always be justified in being made through black-box AI. That is, even if a black-box AI tool is highly accurate at predicting who is more likely to survive, insofar as this is not the only basis on which decisions will always be made, distributive justice issues can extenuate reasons to use black-box AI in medical decision making.

Emergency triage situations are similarly fraught with distributive justice issues but are contextually distinct because how quickly a person is treated is itself a determinative factor in survival and quality of life outcomes. The under resourced, urgent, and serious nature of emergency triage compounds a need for quick, accurate, and efficient healthcare. Given reduced time for deliberation, standards of care are also different: emergency doctors deal with 'variable degrees of uncertainty, with less-than-ideal information, and under severe time constraints' (Iserson, 2006). Physicians make 'efficiency-thoroughness trade-offs': not by choice, but by necessity (Baartmans et al., 2022), and regulatory and health law reduces standard requirements, such as the need for informed consent (Boyle and Stepanov, 2021). The priority of triage decision making is improving as many patient outcomes as quickly as possible (Savulescu et al., 2020). This utilitarian framework is generally consistently applied in times of resource shortages and health crises (Vearrier and Henderson, 2021). Furthermore, given the time and resource pressures, emergency decisions are often justified through post-hoc rationalisations, and as such, are arguably indistinguishable from the post-hoc rationalisations of black-box AI.

Nevertheless, we might distinguish between standard or consistent resource rationing (such as in emergency departments), and extraordinary rationing (such as during natural disasters or pandemics). In the former, AI might be considered more reliable given the consistency of data available in emergency departments. In extraordinary circumstances, however, the nature of the emergency is unpredictable and often impacts disadvantaged populations disproportionately more heavily than privileged ones. We might argue that worries about distribution of resources in extraordinary circumstances might be more akin to organ donation cases, in which following utilitarian principles might further perpetrate bias. At the same time, however, the *urgent* context of extraordinary rationing contexts renders the efficient and accurate nature of black-box AI significantly *more beneficial* and important despite its normative complexity, and such factors may be sufficient for justifying its use.

A final example of the importance of context in justifiability is the role of black-box AI in addressing the inequality in access to healthcare experienced by regional and remote patients. Regional and remote patients experience both inequality in *access* and *quality* of healthcare resulting in the 'rural mortality penalty' (higher rates of morbidity and mortality) (Cosby et al., 2019). This is due to a variety of factors, including the unequal distribution of resources and personnel to rural areas. As such, these patients would greatly benefit from earlier diagnoses, timely expert advice on when to seek treatment, and earlier access to prescription-only treatments (Cao et al., 2024; Cohen, 2020; Dubey and Tiwari, 2023; Tyler et al., 2024). Accurate black-box AI devices might assist with regional and remote patients accessing expert quality care they otherwise would not access. For example, AI-enhanced remote patient monitoring devices have been shown to improve physician capacities to check patient vital signs and physiological parameters remotely through telecommunication devices (Jeddi and Bohr, 2020). These devices have also been used in detecting falls (Pan et al., 2020), cardiac arrhythmia (Devi and Kalaivani, 2020; Neto et al., 2017), and cardiovascular diseases (Bekiri et al., 2020); predicting short-term heart rates and respiratory patterns (Yang et al., 2020); monitoring mental health and predicting suicide (Bernert et al., 2020); and, predicting chronic diseases such as diabetes (Mujumdar and Vaidehi, 2019). If, however, we demand causal explainability, completely non-biased algorithms, or impose extremely high accuracy requirements, we might be denying some patients access to healthcare altogether given their lack of available options. For many people who live in remote and regional areas, there is simply no accessible healthcare at all, and as such, black-box AI might provide sufficiently accurate healthcare.

We ought to be careful, however, of lowering standards of care in ways that lead to relativism between neighbouring communities and even within individual hospitals (Kempt et al., 2022a).[5] Nevertheless, balancing the value of increasing access to healthcare with the value for in-person medical care or explainability is an (unfortunately) necessary compromise. Addressing healthcare inequalities with sufficiently accurate, accessible, and efficient healthcare can outweigh the moral concern of a lack of causal explainability or unequally accurate black-box AI in comparison to metro areas (although not unacceptable standards). Arguably, therefore, accurate black-box AI may be justifiable in circumstances when it addresses or improves inequalities in access to healthcare and healthcare outcomes.

Shared decision making

Another reason to justify certain uses of black-box AI in medical decision making is when these decisions are reviewed, interpreted, tested, and confirmed by a human, and thus, decision making is still shared between doctor and patient (Amann et al., 2020; Holzinger et al., 2019; Maclure, 2021; Robbins, 2019). Just because black-box AI is inscrutable, does not mean that all aspects of AI development, implementation, and use are inscrutable (Ferreira, 2021). As Ehsan et al. (2021) argue, 'the "ability" in explainability does not lie exclusively in the guts of the AI system'. Black-box AI need not simply apply through a copy-and-paste exercise, but how and when human involvement in black-box AI use is *necessary* differs on the kind of involvement and the specific context. There are several aspects of what we term shared decision making with medical AI.

First, justifiably using black-box AI *necessarily* requires that there are regulatory bodies with technical knowledge of

black-box AI security, safety, efficiency, and accuracy that can regulate and provide guidance on testing and use of black-box AI in various conditions and systems. The role of humans in black-box AI decision making is not exclusively limited to the doctor; we might even go so far as to argue that responsibility for scrutinising the reliability, cost-effectiveness, and usefulness of black-box AI in healthcare ought to be proactively addressed through research, regulatory bodies, and institutional administrators. This includes ensuring that the ethical surveillance and audit required on certain black-box AI decisions is proactively determined. For example, if black-box AI is used in ethically sensitive areas such as organ donation, each decision ought to be ethically audited by humans before implementing such decisions. Similarly, in certain populations known to experience bias, regular auditing and reviewing of AI decisions ought to be proactively regulated. The responsibility for determining the extent and nature of such regulations sits firmly outside of the scope of this article.

Second, doctors ought to be involved in matters that require medical judgment and provide post-hoc justifications of AI decisions through shared decision making with the patient (Coeckelbergh, 2020). Physicians are trained according to medical knowledge and should explain the reasons why the AI decision is consistent with medical knowledge in relevant and meaningful ways to the patient. For example, they can explain why the treatment is recommended for the patient's condition, what the side-effects and benefits of treatment are, and why the diagnosis could rationally explain the patient's symptoms. Furthermore, in the same way technicians are relied upon for their technological expertise, a professional (not necessarily the physician) can also be trained to know some technological details of the black-box AI they used, such as the accuracy rate, the data it was trained on, the range of output values, how to monitor the device for decline, and the benefits and limits of the algorithm (Durán, 2021; Sand et al., 2021).

Shared decision making between doctors and patients regarding the use of black-box AI also protects patient autonomy in two ways (Prince and Lim, 2025). First, ensuring that the black-box AI decision is rationally consistent with medical science, the patient's capacity for rational decision making is respected. Second, doctor oversight can provide the epistemic warrant for the patient's trust in the decision-making process and empower the patient with more accurate diagnostic and prognostic information. The patient may then provide reasons and justifications for adopting the recommended decision or indicate why they would prefer an alternative. Such shared decision making would also ensure the patient justifies the decision to themselves (Muralidharan et al., 2024).

Arguably, shared decision-making and thus doctor involvement is necessary to the justifiability of black-box AI, but we believe there are some limited exceptions to the rule. The first is that doctor involvement does not always have to be immediate or concurrent. For example, black-box AI could be used to provide some forms of healthcare to remote and regional patients who would otherwise completely lack, or have extremely limited access to, healthcare (Kantipudi et al., 2021; Kothamali et al., 2023). A more contextual and useful position would be ensuring these patients have the option for review and post-hoc rationalisation by doctors upon request. We ought to be careful of justifying relativist and discriminatory healthcare and safeguard against the worsening of healthcare quality and forcing already discriminated groups from sacrificing explainable medical care out of mere necessity. Nevertheless, this concern can be proactively addressed by ensuring all patients accessing black-box AI algorithms have access to a human (e.g. doctor) to review the AI decisions and by only implementing algorithms that have been sufficiently regulated, trialled, and tested for bias and accuracy.

Some fears, however, have arisen that involving AI in medical decision making necessitates 'machine paternalism' in which physicians simply copy-paste AI decisions and inflict the decision on the patient (Afnan et al., 2021). Bjerring and Busch (2021) argue that in deploying reliable and accurate AI, practitioners will have an epistemic obligation to align their medical verdicts with those of AI systems in a way akin to general practitioners relying on the opinions of experts. Doctors might feel that they must accept the AI decision making for fear of litigation. However, if black-box AI is sufficiently accurate, reliable, *and* the medical practitioner has no reason to disagree with the decision, we ought not to criticise medical practitioners for relying on the prediction. After all, doctors should use reliable and regulated tools that benefits their analysis. Nevertheless, in the complete absence of any justifiable reasons for adopting the AI prediction, particularly in the case where the AI prediction conflicts with known medical theory, we might accept that using such AI is not likely to be justifiable (unless perhaps, it is highly accurate, the patient is terminally ill with no other alternative, and provides informed consent) (Muralidharan et al., 2024).

Bias

We might consider a possible extenuating factor: that is, the worry that black-box AI can hide harmful biases against already marginalised and disadvantaged populations. The justifiability of black-box AI indeed depends in part on how biased the algorithm is, and how harmful the consequences of any bias (Amann et al., 2020). The most obvious harm is that black-box AI will be less accurate and therefore reliable for such groups owing to poor data (Klugman, 2021; Robbins, 2019). Some have argued that bias is only relevant insofar as it reduces accuracy (Akinrinmade et al., 2023) and some argue that black-box AI bias is not morally different to human bias in medical decision making (Kawamleh, 2023; Kempt et al., 2022a; London, 2019). However, *is* does not infer *ought*: scholars argue that simply because

Table 1. A summary of the principle normative factors relevant when determining whether the use of black-box AI in medical decision-making is justifiable, and examples of how they can be balanced and used in potential ethical guidelines in medical decision-making.

Reason	Definition	Examples of how factor contributes to justifiable decision	Potential Guideline Recommendations
Explainability	The ability to provide understandable explanations for AI decisions, not limited to causal explanations	• Post-hoc rationalisations by doctors or explainable AI (XAI) methods that are consistent with medical knowledge • Technical explanations by developers on how the AI was trained, what data it was trained on, what its purpose is, and its limitations	When the decision-making process requires patient input or involves complex ethical considerations, prioritise explainability even if it means sacrificing some minor degree of accuracy.
Accuracy	The degree of correctness in AI predictions or diagnoses	• AI outperforming human doctors in diagnostic or prognostic testing • AI enables early detection of diseases (e.g. breast cancer) in comparison to standard testing	When the AI system demonstrates significantly higher accuracy than current clinical standards, and the primary goal is improved patient outcomes, prioritise accuracy over full explainability.
Context of the Decision	The urgency, seriousness, or extraordinary nature of the medical situation	• Using AI for urgent triage decisions in emergency rooms • AI-assisted diagnosis for life-threatening conditions like pulmonary embolism	In life-threatening or time-critical situations, prioritise the use of highly accurate black-box AI if it can provide faster and more reliable decisions than traditional methods.
Nature of the decision	The type and context of the medical decision being made	• Using AI for clinical diagnostics, prognostics, or treatment recommendations • AI-assisted decision-making in areas with limited healthcare access	For decisions involving complex ethical considerations or resource allocation, ensure explainable AI or human oversight and informed consent. For routine clinical decisions or in resource-limited settings, accurate black-box AI may be justifiable so long as an option for review is available.
Human involvement	The degree of human oversight, shared-decision making, and human interpretation in AI-assisted decisions	• Doctors reviewing and explaining AI recommendations to patients • Regulatory bodies overseeing AI implementation and use • Informed consent to use of AI in normative decisions • Human-only decision-making in ethically sensitive cases	Ensure meaningful human oversight in all AI-assisted medical decisions, with the level of involvement scaling with the complexity and ethical sensitivity of the decision.

medical practitioners and systems are already biased, does not mean that black-box AI is justified in being similarly biased (Maclure, 2021). In the case of black-box AI, the lack of explainability may allow biased algorithms to continue to perpetrate injustice and prevent accountability for such bias.

However, there are several considerations that are relevant when discussing bias. First, in cases of serious bias that results in inaccurate decision making in certain demographics, there will be corresponding decreases in accuracy and reliability. Given this lowered accuracy and thus increased harm, it may not be justifiable to use such algorithms. After all, if an algorithm intended to be used in a healthcare context fails to adequately work for approximately 50% of the population (e.g., women), then it cannot really be similarly accurate to existing standards. Some scholars accept this consequence and justify the use of healthcare that only works for a certain demographic or population (Vandersluis and Savulescu, 2024). Indeed, it is generally difficult to produce therapeutics that work for all patients 100% of the time. However, we cannot resort to defending oppression and discrimination as a reason for a reduction in the quality of medical care; black-box AI is not justifiable if it is only accurate for a privileged demographic.

Second, we can proactively address bias through regulatory standards in data testing and establish pathways for legal redress as and when mistakes inevitably occur. Some scholars and studies show that black-box AI, although causally unexplainable, is, or at least perceived to be, more easily screened for biases than humans (Drezga-Kleiminger et al., 2023; Kawamleh, 2023). After all, black-box AI has no

internal reason or unconscious bias against certain people; any inherent bias is due to missing data or learned patterns, and such, can be systematically tested for bias and modified following such testing. Although we ought to be cautious of overstating the ease of regulating and identifying bias, the mere *presence* of bias is not itself extenuating when justifying the use of black-box AI in medical decision making.

Furthermore, if medical AI devices become regulated through clinical trials or other regulatory processes, any criticisms of the role of bias and error should be redirected to the regulatory process, rather than black-box AI alone. If black-box AI is permitted to be biased, this is not inherent to the algorithm so much as it is in the clinical trial and regulatory process. However, clinical trials do not demand perfection, and as a society, we have long accepted the limits of medical treatments and diagnostics. As such, we have, or at least ought to have, regulatory systems that have clear principles for enforcing accountability when developers intentionally or negligently fail to proactively address bias, and redress for persons who are subject to bias or discriminatory treatment. Nevertheless, whether black-box AI is being introduced in such regulatory systems with laws and policies contextualised to AI will impact whether using such black-box AI is justified.

Balance and justifiability

We have thus far argued that we are unjustified in affording lexical priority to the explainability of medical decision making. The complexity of medical decision making means we cannot recommend a uniform normative framework for recommending the most justifiable means of using black-box AI, but we can identify the variety of factors at play and use casuistic reasoning to understand how these factors can be weighed to make its use justifiable. As we have argued, there are many good reasons to justify the use of black-box AI – and these reasons can interact and justify the absence of other factors. We have argued that accuracy is a necessary, although not sufficient, condition in any justification of its use. We have also considered how black-box AI can be more easily justified in urgent and serious cases because the need to make quick decisions directly impacts the patient's outcome, and that black-box AI is more likely to be justifiable in clinical decision making rather than normative decisions given that need humans to be accountable for inherently ethical decisions. Furthermore, we have argued that *unless absolutely necessary* to improve equitable healthcare outcomes, ensuring that decisions made by black-box AI is used in conjunction with a human decision-maker can provide the normative value of accountability and interpretability.

In contrast, sometimes a patient's unique medical history and experience of uncertainty, inexplicability, and poor treatment by the medical system may create reason to use explainable over accurate decision-making processes. The use of black-box AI may be unjustified to make a decision that is necessarily subjective, such as in accordance with the patient's values, and if it is implemented without informed consent and shared decision making with the patient. Finally, the presence of bias does not provide a blanket prohibition of black-box AI insofar as such bias has been proactively regulated and mitigated through clinical trials, testing, and regulation.

Justifiability is about reasonability and defensibility, and we can balance a variety of reasons in different contexts to determine when it is justifiable to use black-box AI. Importantly, however, it is not necessary to have all factors, but as we have shown, it is likely that AI must be accurate in addition to at least one other factor.

Conclusion

We have argued that accurate and reliable black-box AI can be justifiably used in medical practice under certain conditions (see Table 1 for a summary). The proposition that black-box AI is a 'trade off' of explainability for accuracy is thus arguably misconceived. Nevertheless, guidelines and policies must be sensitive to the interplay of considerations that changes according to each new medical context. The use of accurate and reliable black-box AI can therefore be justified in healthcare, and we must be careful of ousting it for the reason that regulators could more easily attribute accountability if we knew the reasons for which decisions were made.

Acknowledgements

Any use of generative AI in this manuscript adheres to ethical guidelines for use and acknowledgement of generative AI in academic research (Porsdam Mann et al. 2024). Each author has made a substantial contribution to the work, which has been thoroughly vetted for accuracy, and assumes responsibility for the integrity of their contributions. Porsdam Mann, S., Vazirani, A.A., Aboy, M. et al. Guidelines for ethical use and acknowledgement of large language models in academic writing. *Nat Mach Intell* 6, 1272–1274 (2024). https://doi.org/10.1038/s42256-024-00922-7

ORCID iDs

Sinead Prince https://orcid.org/0000-0002-7370-9526
Julian Savulescu https://orcid.org/0000-0003-1691-6403

Ethical approval

Not applicable.

Informed consent

Not applicable.

Author contributions

SP conceived and drafted the article. JS then drafted and edited the article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded in part by the Wellcome Trust [Grant number 226801/Z/22/Z] and supported by the National University of Singapore under its NUS Start-Up grant (NUHSRO/2022/078/Startup/13). Julian Savulescu, through his involvement with the Murdoch Childrens Research Institute, received funding from the Victorian State Government through the Operational Infrastructure Support (OIS) Program.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability statement

Not applicable.

Notes

1. The exact goals of medicine differ widely in philosophical and bioethical literature, but for the sake of this paper, we assume the shared broad goal of treating patients with available medical research and therapeutics (Boorse, 2016; Kingma, 2007).
2. Qualitative research may be required to determine how women with lived experience of these conditions would justify these decisions (Savulescu et al., 2021).
3. Although in this example, the doctor could answer questions about the benefits, side effects, risks of treatment, and the impact on lifestyle.
4. That doctors may neither have time nor medical expertise to reject the black-box AI decision is discussed below in the 'Shared Decision making' section.
5. We acknowledge that governments may not prioritize novel technologies for rural and remote patients. However, our analysis only claims that addressing healthcare inequality might justify the use of black-box AI despite the inherent inexplicability.

References

Afnan MAM, Liu Y, Conitzer V, et al. (2021) Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Human Reproduction Open* 2021(4): 1021.

Akinrinmade AO, Adebile TM, Ezuma-Ebong C, et al. (2023) Artificial intelligence in healthcare: Perception and reality. *Cureus* 15(9): e45594.

Alvarez M (2017) Reasons for action: Justification, motivation, explanation. In: Zalta EN (ed) *The Stanford Encyclopedia of Philosophy*. Winter 2017. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/ (accessed 20 August 2024).

Amann J, Blasimme A, Vayena E, et al. (2020) Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making* 20(1): 310.

Baartmans MC, Hooftman J, Zwaan L, et al. (2022) What can we learn from in-depth analysis of human errors resulting in diagnostic errors in the emergency department: An analysis of serious adverse event reports. *Journal of Patient Safety* 18(8): e1135–e1141.

Babic B and Cohen IG (2023) The algorithmic explainability "bait and switch". *Minnesota Law Review* 108: 857–909.

Babic B, Gerke S, Evgeniou T, et al. (2021) Beware explanations from AI in health care. *Science* 373(6552): 284–286.

Balasubramaniam N, Kauppinen M, Rannisto A, et al. (2023) Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology* 159: 107197.

Bekiri R, Djeffal A and Hettiri M (2020) A remote medical monitoring system based on data mining. In: 2020 *1st International conference on communications, control systems and signal processing (CCSSP)*, May 2020, pp. 282–286. Available at: https://ieeexplore.ieee.org/document/9151713 (accessed 16 October 2024).

Bernert RA, Hilberg AM, Melia R, et al. (2020) Artificial intelligence and suicide prevention: A systematic review of machine learning investigations. *International Journal of Environmental Research and Public Health* 17(16): 5929.

Binns R (2018) Algorithmic accountability and public reason. *Philosophy & Technology* 31(4): 543–556.

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

Boorse C (2016) Goals of Medicine. In: Giroux É (ed) *Naturalism in the Philosophy of Health. History, Philosophy and Theory of the Life Sciences*. Cham: Springer International Publishing, 145–177.

Boyle S and Stepanov N (2021) Providing emergency medical care without consent: How the 'emergency principle' in Australian law protects against claims of trespass. *Emergency Medicine Australasia: EMA* 33(3): 575–579.

Cao B, Huang S and Tang W (2024) AI Triage or manual triage? Exploring medical staffs' preference for AI triage in China. *Patient Education and Counseling* 119: 108076.

Chan B (2023) Black-box assisted medical decisions: AI power vs. Ethical physician care. *Medicine, Health Care and Philosophy* 26(3): 285–292.

Coeckelbergh M (2020) Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics* 26(4): 2051–2068.

Cohen IG (2020) Informed consent and medical artificial intelligence: What to tell the patient? *Georgetown Law Journal* 108(6): 1425–1460.

Cosby AG, McDoom-Echebiri MM, James W, et al. (2019) Growth and persistence of place-based mortality in the United States: The rural mortality penalty. *American Journal of Public Health* 109(1): 155–162.

Daniels N (2003) Reflective equilibrium. *Stanford Encyclopedia of Philosophy Archive*. Epub ahead of print 28 April 2003.

Devi RL and Kalaivani V (2020) Machine learning and IoT-based cardiac arrhythmia diagnosis using statistical and dynamic features of ECG. *The Journal of Supercomputing* 76(9): 6533–6544.

Ding W, Abdel-Basset M, Hawash H, et al. (2022) Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences* 615: 238–292.

Drezga-Kleiminger M, Demaree-Cotton J, Koplin J, et al. (2023) Should AI allocate livers for transplant? Public attitudes and ethical considerations. *BMC Medical Ethics* 24(1): 102.

Dubey A and Tiwari A (2023) Artificial intelligence and remote patient monitoring in US healthcare market: A literature review. *Journal of Market Access & Health Policy* 11(1): 2205618.

Durán JM (2021) Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artificial Intelligence* 297: 103498.

Ehsan U, Liao QV, Muller M, et al. (2021) Expanding explainability: Towards social transparency in AI systems. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*, Yokohama Japan, 6 May 2021, pp. 1–19. ACM. Available at: https://dl.acm.org/doi/10.1145/3411764.3445188 (accessed 4 July 2024).

Esteva A, Kuprel B, Novoa RA, et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639): 115–118.

Ferreira M (2021) Inscrutable processes: Algorithms, agency, and divisions of deliberative labour. *Journal of Applied Philosophy* 38(4): 646–661.

Freyer N, Groß D and Lipprandt M (2024) The ethical requirement of explainability for AI-DSS in healthcare: A systematic review of reasons. *BMC Medical Ethics* 25(1): 104.

Fryer J, Mason-Jones AJ and Woodward AL (2024) Understanding diagnostic delay for endometriosis: A scoping review. *Understanding diagnostic delay for endometriosis*. medRxiv. Epub ahead of print 9 January 2024. DOI: 10.1101/2024.01.08.24300988.

Hadfield G (2021) Explanation and justification: AI decision-making, law, and the rights of citizens. Available at: https://srinstitute.utoronto.ca/news/hadfield-justifiable-ai (accessed 3 July 2024).

Hamon R, Junklewitz H and Sanchez I (2020) *Robustness and explainability of artificial intelligence: From technical to policy solutions*. LU: European Commission. Available at: https://data.europa.eu/doi/10.2760/57493 (accessed 4 July 2024).

Hois J, Theofanou-Fuelbier D and Junk AJ (2019) How to achieve explainability and transparency in Human AI interaction. In: Stephanidis C (ed) *HCI International 2019 - Posters*. Communications in Computer and Information Science. Cham: Springer International Publishing, pp. 177–183. Available at: http://link.springer.com/10.1007/978-3-030-23528-4_25 (accessed 4 July 2024).

Holzinger A, Langs G, Denk H, et al. (2019) Causability and explainability of artificial intelligence in medicine. *WIRES Data Mining and Knowledge Discovery* 9(4): e1312.

Iserson KV (2006) Ethical principles—emergency medicine. *Emergency Medicine Clinics of North America* 24(3): 513–545.

Jackson G (2019) *Pain and Prejudice: How the Medical System Ignores Women and What We Can Do About It*. United Kingdom: Allen & Unwin.

Jeddi Z and Bohr A (2020) Chapter 9 - Remote patient monitoring using artificial intelligence. In: Bohr A and Memarzadeh K (eds) *Artificial Intelligence in Healthcare*. Academic Press, pp. 203–234. Available at: https://www.sciencedirect.com/science/article/pii/B9780128184387000095 (accessed 15 October 2024).

Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389–399.

Kantipudi MVVP, Moses CJ, Aluvalu R, et al. (2021) Remote patient monitoring using IoT, cloud computing and AI. In: Kumar Bhoi A, Mallick PK, Narayana Mohanty M, et al. (eds) *Hybrid Artificial Intelligence and IoT in Healthcare*. Singapore: Springer, pp. 51–74. Available at: https://doi.org/10.1007/978-981-16-2972-3_3 (accessed 15 October 2024).

Kawamleh S (2023) Against explainability requirements for ethical artificial intelligence in health care. *AI and Ethics* 3(3): 901–916.

Kempt H, Freyer N and Nagel SK (2022a) Justice and the normative standards of explainability in healthcare. *Philosophy & Technology* 35(4): 100.

Kempt H, Heilinger J-C and Nagel SK (2022b) Relative explainability and double standards in medical decision-making. *Ethics and Information Technology* 24(2): 20.

Kingma E (2007) What is it to be healthy? *Analysis* 67(294): 128–133.

Klugman CM (2021) Black boxes and bias in AI challenge autonomy. *The American Journal of Bioethics* 21(7): 33–35.

Kothamali PR, Srinivas N, Mandaloju N, et al. (2023) Smart healthcare: Enhancing remote patient monitoring with AI and IoT. *Revista de Inteligencia Artificial en Medicina* 14(1): 113–146.

Kriza C, Amenta V, Zenié A, et al. (2021) Artificial intelligence for imaging-based COVID-19 detection: Systematic review comparing added value of AI versus human readers. *European Journal of Radiology* 145: 110028.

London AJ (2019) Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *The Hastings Center Report* 49(1): 15–21.

Maclure J (2021) AI, explainability and public reason: The argument from the limitations of the human mind. *Minds and Machines* 31(3): 421–438.

Malgieri G and Pasquale FA (2022) From transparency to justification: Toward ex ante accountability for AI. *SSRN Electronic Journal*. Epub ahead of print 2022. DOI: 10.2139/ssrn.4099657.

McKinney SM, Sieniek M, Godbole V, et al. (2020) International evaluation of an AI system for breast cancer screening. *Nature* 577(7788): 89–94.

Mersha M, Lam K, Wood J, et al. (2024) Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing* 599: 128111.

Mota SM, Priester A, Shubert J, et al. (2024) Artificial intelligence improves the ability of physicians to identify prostate cancer extent. *Journal of Urology*. Philadelphia, PA: Wolters Kluwer. Epub ahead of print July 2024. DOI: 10.1097/JU.0000000000003960.

Mujumdar A and Vaidehi V (2019) Diabetes prediction using machine learning algorithms. In: *Procedia computer science 165. 2nd International conference on recent trends in advanced computing ICRTAC -DISRUP - TIV INNOVATION*, 2019 November 11-12, 2019, pp. 292–299.

Muralidharan A, Savulescu J and Schaefer GO (2024) AI And the need for justification (to the patient). *Ethics and Information Technology* 26(1): 16.

Neto LASM, Pequeno R, Almeida C, et al. (2017) A method for intelligent support to medical diagnosis in emergency cardiac care. In: 2017 *International joint conference on neural networks (IJCNN)*, May 2017, pp. 4587–4593. Available at: https://ieeexplore.ieee.org/document/7966438 (accessed 16 October 2024).

Nori H, Daswani M, Kelly C, et al. (2025) Sequential diagnosis with language models. arXiv:2506.22405. arXiv. Available at: http://arxiv.org/abs/2506.22405 (accessed 4 July 2025).

Ordish J, Bridgen T and Hall A (2020) *Black box medicine and transparency: Ethics of transparency and explanation*. Report. University of Cambridge: PHG Foundation. Available at: https://www.phgfoundation.org/wp-content/uploads/2023/10/black-box-ethics-transparency-explanation.pdf (accessed 15 July 2024).

Pan D, Liu H, Qu D, et al. (2020) Human falling detection algorithm based on multisensor data fusion with SVM. *Mobile Information Systems* 2020(1): 8826088.

Parfit D (2011) *On What Matters*. vol. 1. United Kingdom: Oxford University Press.

Peters U (2023) Explainable AI lacks regulative reasons: Why AI and human decision-making are not equally opaque. *AI and Ethics* 3(3): 963–974.

Prince S and Lim JE (2025) Black-Box AI and patient autonomy. *Minds and Machines* 35(2): 24.

Rawls J (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Raz J (2011) *From Normativity to Responsibility*. Oxford University Press. Available at: https://scholarship.law.columbia.edu/faculty_scholarship/1488.

Robbins S (2019) A misdirected principle with a catch: Explicability for AI. *Minds and Machines* 29(4): 495–514.

Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. (2019) Stand-Alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists. *Journal of the National Cancer Institute* 111(9): 916–922.

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.

Rueda J, Rodríguez JD, Jounou IP, et al. (2024) "Just" accuracy? Procedural fairness demands explainability in AI-based medical resource allocations. *AI & SOCIETY* 39(3): 1411–1422.

Sand M, Durán JM and Jongsma KR (2021) Responsibility beyond design: Physicians' requirements for ethical medical AI. *Bioethics* 36(2): 162–169.

Savulescu J, Gyngell C and Kahane G (2021) Collective reflective equilibrium in practice (CREP) and controversial novel technologies. *Bioethics* 35(7): 652–663.

Savulescu J, Vergano M, Craxì L, et al. (2020) An ethical algorithm for rationing life-sustaining treatment during the COVID-19 pandemic. *BJA: British Journal of Anaesthesia* 125(3): 253–258.

Senapati A, Tripathy HK, Sharma V, et al. (2024) Artificial intelligence for diabetic retinopathy detection: A systematic review. *Informatics in Medicine Unlocked* 45: 101445.

Shahvisi A (2016) No understanding, No consent: The case against alternative medicine. *Bioethics* 30(2): 69–76.

Shickel B, Loftus TJ, Adhikari L, et al. (2019) DeepSOFA: A continuous acuity score for critically ill patients using clinically interpretable deep learning. *Scientific Reports* 9(1): 1879.

Sidgwick H (1874) *The Methods of Ethics*. London: Macmillan.

Steging C, Renooij S and Verheij B (2021) Rationale discovery and explainable AI. In: Schweighofer E (ed) *Frontiers in Artificial Intelligence and Applications*. IOS Press. Available at: https://ebooks.iospress.nl/doi/10.3233/FAIA210341 (accessed 8 July 2024).

Tan JKL, Vasey K and Fung KY (2001) Beliefs and perceptions of patients with acne. *Journal of the American Academy of Dermatology* 44(3): 439–445.

Therapeutic Goods Administration (TGA). (2024) Artificial Intelligence (AI) and medical device software. Therapeutic Goods Administration (TGA). Available at: https://www.tga.gov.au/how-we-regulate/manufacturing/manufacture-medical-device/manufacture-specific-types-medical-devices/artificial-intelligence-ai-and-medical-device-software (accessed 16 December 2024).

Ting DSW, Pasquale LR, Peng L, et al. (2019) Artificial intelligence and deep learning in ophthalmology. *The British Journal of Ophthalmology* 103(2): 167–175.

Tyler S, Olis M, Aust N, et al. (2024) Use of artificial intelligence in triage in hospital emergency departments: A scoping review. *Cureus* 16(5): e59906.

Vandersluis R and Savulescu J (2024) The selective deployment of AI in healthcare: An ethical algorithm for algorithms. *Bioethics* 38(5): 391–400.

Vearrier L and Henderson CM (2021) Utilitarian principlism as a framework for crisis healthcare ethics. *Hec Forum* 33(1–2): 45–60.

Yang J, Xiao W, Lu H, et al. (2020) Wireless high-frequency NLOS monitoring system for heart disease combined with hospital and home. *Future Generation Computer Systems* 110: 772–780.

Zeltzer D, Herzog L, Pickman Y, et al. (2023) Diagnostic accuracy of artificial intelligence in virtual primary care. *Mayo Clinic Proceedings: Digital Health* 1(4): 480–489.

Zhang J, Kowsari K, Harrison JH, et al. (2018) Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access* 6: 65333–65346.