Two Reasons for Subjecting Medical AI Systems to Lower Standards than Humans

Jakob Mainz
Novo Nordisk
jakob-mainz@hotmail.com

Lauritz Aastrup Munch
Philosophy and History of Ideas
Aarhus University, Denmark
lauritzmunch@gmail.com

Jens Christian Bjerring
Philosophy and History of Ideas
Aarhus University, Denmark
filjcb@cas.au.dk

ABSTRACT

This paper concerns the double standard debate in the ethics of AI literature. This debate revolves around the question of whether we should subject AI systems to different normative standards than humans. So far, the debate has centered around transparency. That is, the debate has focused on whether AI systems must be more transparent than humans in their decision-making processes in order for it to be morally permissible to use such systems. Some have argued that the same standards of transparency should be applied to AI systems and humans. Others have argued that we should hold AI systems to higher standards than humans in terms of transparency. In this paper, we first highlight that debates concerning double standards, which have a similar structure to those related to transparency, exist in relation to other values such as predictive accuracy. Second, we argue that when we focus on predictive accuracy, there are at least two reasons for holding AI systems to a lower standard than humans.

KEYWORDS

Double Standard, Predictive Accuracy, Opacity, Cost-effectiveness, Speed

ACM Reference Format:
Jakob Mainz, Lauritz Aastrup Munch, and Jens Christian Bjerring. 2023. Two Reasons for Subjecting Medical AI Systems to Lower Standards than Humans. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3593013.3593975

1 INTRODUCTION

There is an ongoing debate about whether we should hold AI systems to different standards than humans when it comes to decisional transparency [8, 12, 24, 27, 33]. Some believe that we should hold AI systems to the same standard as humans [33]. Others believe that we should hold AI systems to a higher standard than humans [12]. The crux of the debate is whether AI systems used for decision-making should exhibit a greater level of transparency—or a lower degree of opacity—in their decision-making processes compared to humans. So far, no apparent consensus has emerged in the literature. Although the discussion on double standards regarding algorithmic transparency is well-known, we want to demonstrate that similar debates on double standards ought to exist in regards to other important criteria beyond transparency such as predictive accuracy, fairness, and trustworthiness. We believe these other desiderata deserve the same sort of double standard debate, and this paper is a first stab at extending the debate to include them.

For current purposes, we focus on the desideratum of predictive accuracy. For ease of exposition, we shall speak of predictive accuracy as the AI system's ability to make correct binary predictions. In turn, we can define predictive accuracy as the percentage of the system's predictions that are correct. More precisely, we can calculate the aggregate accuracy score of an AI system by adding its true positive predictions to its true negative predictions, and then divide this number by the total number of predictions:

$$Accuracy_{AI} = \frac{(TruePositives_{AI} + TrueNegatives_{AI})}{\begin{pmatrix} TruePositives_{AI} + TrueNegatives_{AI} + \\ FalsePositives_{AI} + FalseNegatives_{AI} \end{pmatrix}}$$

Similarly, we can calculate the accuracy score of humans as follows:

$$Accuracy_{Human} = \frac{(TruePositives_{Human} + TrueNegatives_{Human})}{\begin{pmatrix} TruePositives_{Human} + TrueNegatives_{Human} + \\ FalsePositives_{Human} + FalseNegatives_{Human} \end{pmatrix}}$$

So what we will be focusing on is whether our normative standards require that $\text{Accuracy}_{\text{AI}}$ be different from $\text{Accuracy}_{\text{Human}}$. We focus on accuracy for two reasons. First, accuracy is quantifiable and commensurable across human and artificial decision-makers. This sets accuracy apart from at least some other desiderata worth caring about in the context of algorithmic decision-making. For instance, no equally unambiguous metrics obviously exist for desiderata such as transparency and trustworthiness. Second, in many domains of critical decision-making, accuracy is often regarded as one of the most, if not the most important desideratum. This holds true in domains such as healthcare, criminal justice, and banking.

In this paper, we take a position in the double standard debate that has so far been vacant. Specifically, we argue that there are reasons for thinking that it can sometimes be morally permissible—if not morally obligatory—to hold AI systems to lower standards than humans. That is, we want to claim that it can sometimes be morally permissible to rely on an AI system even if $\text{Accuracy}_{\text{AI}}$ is lower than $\text{Accuracy}_{\text{Human}}$. We offer two reasons in favor of this conclusion. The first reason is that certain AI systems are—or supposedly will be—highly cost-effective: they can be expected to perform their designated tasks with efficiency comparable to humans but at a lower cost of usage than human labor. When AI systems are sufficiently cost-effective in this sense, we want to

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3593975

argue that we can have reason to rely on them instead of humans, even if they exhibit slightly less accuracy than humans in their medical predictive abilities. The second reason is that the predictive speed of many AI systems is much higher than that of humans: many AI systems arrive at their predictions significantly faster than humans do. When we refer to predictive speed, we are not only referring to the computational speed at which an AI system can crunch numbers. Instead, we are referring to the time it takes for an AI system to arrive at a decision in a given decision-making process. To use an example we will return to later, think of an AI system that detects deadly diseases. While it is obviously important that the AI system can crunch numbers fast, there are types of deadly diseases—as we shall see in detail later—for which it is even more important that it can detect these diseases earlier in the process than any human can. Indeed, as we shall argue later, in medical situations where diagnostic speed and early treatment is of vital importance, there can be excellent reasons to rely on AI systems rather than humans even if $\text{Accuracy}_{\text{AI}}$ is lower than $\text{Accuracy}_{\text{Human}}$. Although we focus on detection time here, it is worth noting that AI systems could increase speed through other mechanisms. While a human practitioner can feasibly only tend to one or a few patients at a time due to cognitive limitations, an AI system can plausibly diagnose multiple patients simultaneously due to its increased computational capacities. Insofar as the marginal cost of algorithmic computational power is lower than the marginal cost of human reasoning power, a contribution to predictive speed along these dimensions would also be cost-effective.

Note: we shall not argue that reasons concerning cost-effectiveness and predictive speed always trump competing reasons for holding AI systems to higher standards than humans when it comes to accuracy. We only argue that the values of cost-effectiveness and speed reflect genuine reasons that we ought to take into consideration when deciding questions about algorithmic decision aids. But we allow that these reasons might sometimes be overruled by stronger countervailing reasons. Moreover, considerations pertaining to cost-effectiveness and predictive speed present us with reasons that have so far gone largely unnoticed in the double standard debates. But, we argue, they ought to be considered when we balance reasons in favor of and against the idea that $\text{Accuracy}_{\text{AI}}$ must match $\text{Accuracy}_{\text{Human}}$.

Before proceeding, let us make a few further clarifications regarding the scope of our argument. There are several issues that might influence what a plausible stance in the double standard debate should look like. First, there are issues concerning the domain of operation. So far, the double standard debate has occurred at a fairly abstract level. However, given the highly domain-specific nature of the key arguments in the debate, it is helpful to concentrate on a specific domain of operation. We shall concentrate on the medical domain. The medical domain is a natural choice because considerations pertaining to algorithmic accuracy, cost-effectiveness, and predictive speed are particularly salient in the medical domain. But again, the arguments we give are likely to apply in some form in other central domains as well.

Second, there are issues concerning the level of autonomous decision power that we allocate to the relevant AI systems. For instance, it matters whether AI systems are used as decision support tools, or whether they are capable of making decisions on their own. For using an AI system as a decision support tool is typically easier to justify than relying on it as a decision-maker on its own. After all, there is a 'human in the loop' in the former but not the latter case. But if we can demonstrate that our argument holds true for AI systems that make unilateral decisions, we will also have shown that it holds true for systems that are merely used for decision support. As a result, we have chosen to concentrate on AI systems that make unilateral decisions. Such systems raise other worries than those that will concern us here, but we refer to other work for discussion of some of these [19].

Third, there are issues concerning the choice of comparison class. Which group of human clinicians should we compare the relevant AI system to? The answer here is far from obvious. Should we consider the best human clinicians in the field? The worst human clinicians? The average human clinician? Or the reasonably available human clinician or group of clinicians? Given that the accuracy levels of human clinicians vary considerably—and that the accuracy of a particular individual can also fluctuate over time—the selection of the comparison class from within the group of human clinicians carries significant weight in determining whether the AI system in question is more or less accurate than the human clinicians to which it is compared. Although little hangs on this issue for purposes of stating our argument, we shall assume that the AI system is compared, counterfactually, to the best, reasonably available human that could have made the decision, had the AI system not been in charge of the decision-making process. In practice, this would mean that we would often be comparing the AI system to whoever was replaced by the system when it was put in charge of the decision-making process. We include both 'best' and 'reasonably available' in our comparison class for the following reasons. First, practical circumstances could mean that the 'best' decision-maker (judged by some standard) that could make the decision is not 'reasonably available' for making a given decision. Suppose for instance that one human clinician by far exceeds all other clinicians in their cohort in terms of accuracy for some diagnostic procedure and in this sense counts as 'best'. Comparing the accuracy of the relevant AI system to the accuracy of that individual seems problematic because it is unfeasible to expect that this individual would be making all the decisions that must be made; after all, even the best human clinician needs some time off. Second, we include 'best' because some 'reasonably available' human clinicians may not be very good in the sense of not being very accurate. Although situations may arise where only such suboptimal human clinicians are reasonably available, this group appears to be a less suitable point of comparison when considering the value of substituting them with an AI system.

Finally, there are issues concerning the stakes associated with the relevant decision. Generally, it seems easier to justify replacing a human decision-maker with an AI system in a low stakes decision context. As such, one might think, it is also easier to justify the inclusion of AI systems that are less accurate than humans in low stakes contexts. To make our life as hard as possible, we shall thus focus on high stakes decision contexts involving AI systems in use for detecting and diagnosing deadly diseases.

2 COST-EFFECTIVENESS

As mentioned, the first reason that counts in favor of holding AI systems to a lower accuracy standard than humans concerns the cost-effectiveness of these systems. Replacing human clinicians with more cost-effective AI systems means that scarce healthcare resources can be spent elsewhere to produce more overall health or even save more lives. Cost-effectiveness is of course not the only thing that matters morally when we make medical decisions. But it is undeniable that it matters. If the maintenance of high accuracy levels comes with substantial opportunity costs—implying that the resources allocated to attain these levels could be utilized more effectively elsewhere—we should pause and reflect. So not only does cost-effectiveness matter, it is also uncontroversially considered a weighty consideration that can only be trumped by very weighty counter-considerations [22, 28, 29]. And this is all we have in mind when we say that the cost-effectiveness of a medical AI system gives us reason to use that system even if $\text{Accuracy}_{\text{AI}}$ is lower than $\text{Accuracy}_{\text{Human}}$. We may decide that other considerations trump cost-effectiveness in specific situations, but since medical resources are realistically always scarce, cost-effectiveness should always play some role in moral deliberations concerning the proper accuracy standards for medical AI systems. To the extent that AI systems are more cost-effective than humans, then, this fact will count in favor of using these systems even though $\text{Accuracy}_{\text{AI}}$ is lower than $\text{Accuracy}_{\text{Human}}$.

As it turns out, many expect AI systems to become very cost-effective in the future. These systems are expected by many to become cheap to implement and upkeep, and they are expected by many to perform at least as well as humans in terms of accuracy in the future [30]. To illustrate the potential magnitude of cost-effectiveness that we are dealing with, consider a recent study about diabetic screenings in Singapore [31]. By replacing human practitioners with AI systems, the study found that yearly savings of roughly $15 millions could be achieved by 2050 with respect to diabetic screening in Singapore alone, and that these financial reductions could be achieved without reducing the accuracy levels of the screening process. Similar studies show that novel AI systems significantly improve the accuracy levels of cardiovascular risk prediction, and that the systems needed for achieving this feat will be cheap to upkeep [3, 5, 9, 11]. To be sure, while studies like these show promising results, it is still too early to paint a general picture of how cost-effective medical AI systems will eventually be. Some existing systems tend not to perform very well once we move beyond the testing stage, and some systems tend to be very expensive to develop. So we cannot unequivocally say that all relevant future AI systems will be cost-effective. Luckily, for our argument to go through, we do not need to establish that most or all existing AI systems are very cost-effective, nor that they will be so in the future. We only claim that to the extent that they will be cost-effective it gives us reason to use them, even if $\text{Accuracy}_{\text{AI}}$ is lower than $\text{Accuracy}_{\text{Human}}$.

Of course, it matters morally how much lower AccuracyAI is compared to $\text{Accuracy}_{\text{Human}}$. If $\text{Accuracy}_{\text{AI}}$ is not higher than the accuracy levels achieved by tossing a coin, then there is not much point in using the AI system in question—regardless of how much cheaper it is compared to humans. Moreover, some may believe that patients are owed a certain minimal threshold of predictive accuracy from medical algorithms, arising perhaps from a duty of care in healthcare or from some professional norm dictating how much a medical prediction can vary from the profession's state of the art. If one were to adopt a strict absolutist stance on the required level of precision for a diagnostic procedure—insisting that only a specific level of precision is morally acceptable and any deviation from it is problematic—this would represent a limiting case. However, such a position is unlikely to be compelling precisely because it would conflict with the underlying objective of pursuing cost-effectiveness. But aside from this, the points we make below are compatible with the existence of such thresholds imposing certain limits on the variability of accuracy standards in light of other values such as cost-effectiveness or speed. Luckily, for our purposes, there is ample empirical evidence that even existing medical AI systems display levels of accuracy that come close to those of humans—indeed, there is evidence that some of them even surpass humans. Even when existing medical AI systems display levels of accuracy that are lower than those displayed by humans, the difference is normally only a few percentages. For instance, a recent systematic review found—among other things—that many medical AI systems perform very well in terms of accuracy in the context of musculoskeletal radiology [10, 14]. Across all performance measures—not only accuracy—the review found that the AI systems performed better than human clinicians in 38% of the studies included in the review, worse than human clinicians in 3.8% of the studies, while no difference was found in 58% of the studies. Importantly, in the few studies where the AI system did perform worse than human clinicians, the review showed that on average the difference was only approximately 5% [10].

Also, when discussing the relevance of cost-effectiveness, it is important to keep in mind that not all forms of cost-effectiveness may permissibly be pursued. As we pointed out above, the value of cost-effectiveness lies in how it allows for more healthcare resources to be allocated towards other beneficial endeavors. It is clear, however, that the permissibility of this maneuver is partially dependent upon what these other beneficial endeavors are. For instance, it may never be morally permissible to trade off even a few patients' lives with a slight reduction of headache pains for millions of other patients. While these considerations do not amount to an objection to our central argument, they add two important nuances to our argument.

First, not all increases in cost-effectiveness will warrant accepting lower accuracy thresholds. We should only compromise on accuracy to achieve cost-effectiveness if the resources saved can be utilized to create more benefits elsewhere, resulting in a net gain in overall medical benefit. If the compromise on accuracy results in a slight increase in lost lives, for instance, we may never end up with a net gain in overall medical benefit if the benefits generated elsewhere amount only to headache relieves. As such, we cannot—at least not in isolation from the broader medical context—specify a general rule that tells us how much a specific increase in cost-effectiveness will warrant a specific reduction of predictive accuracy. Rather, it all depends on the aims that specific levels of accuracy enable us to pursue and on the benefits that we can generate with the saved costs. A second nuance to consider is that the pursuit of cost-effectiveness may sometimes amount to trading welfare

between people. To make this concrete, suppose we implement an AI system for diagnosing gastric cancer that is slightly less accurate than the human alternatives but much more cost-effective. Suppose the resources liberated are then used to improve the treatment for people who suffer from a broken ankle. On one way of describing this situation, resources are shifted from individuals with gastric cancer, who may experience a slightly higher risk of not receiving an accurate diagnosis and subsequent treatment, to individuals with broken ankles who may benefit from better treatment to achieve a slightly better recovery. Suppose also that a much larger number of people experience broken ankles than gastric cancer. While it might seem obvious that this transfer of welfare from cancer patients to fracture patients is unjust, it is important to remember that different theories will take different views on the moral status of such transfers of welfare—even if all should recognize the value of providing the most health given the available resources. Act utilitarians, for instance, should prefer transfers like the one above provided they increase total utility. In contrast, many non-consequentialists hold that these sorts of transfers are morally unacceptable: it is not justifiable to cause a significantly greater harm to some individuals (for instance by increasing their risk of having undetected gastric cancer) to provide only a slight benefit to many others (such as improving their recovery from broken bones), even if the total sum of welfare is thereby increased**.** In making this point we are making the simplifying assumption that withholding a more accurate diagnostic tool from patients amounts to harming them. But the case is perhaps more accurately described as one of withholding a benefit. Or, to take a stock case from the relevant literature, if we have the option to either save one person from certain death or save a billion people from a slight headache, we should intuitively prefer saving the one person from certain death; indeed, we should continue to do so even if the number of slight headaches we could prevent increase tremendously. Such views are often referred to as "partially aggregative" because they say that we only sometimes should prefer to act in ways that maximize aggregate moral value, see [13]. If you share this intuition, you should want to say that pursuing slightly less accurate but cost-effective AI systems meant to mitigate severe harms is not justifiable if the plan is to allocate the saved resources towards mitigating much lesser harms.

We need not take a stand on these issues. Even if we adopt the view that some forms of trading welfare between people are morally impermissible—such as the trade between gastric cancer patients and broken ankle patients above—our argument remains relevant for many implementations of AI systems. Specifically, our argument will apply whenever we are contemplating lowering the accuracy of detecting or treating less severe diseases in order to enhance the detection or treatment of more severe ones.

So, as we have seen, while it remains hard in full generality to compare cost-effectiveness levels between medical AI systems and humans, it is clear enough from the literature that some AI systems are already more cost-effective than humans, and that we can reasonably speculate that we will see more of these AI systems in the future. At any rate, we have motivated the thought that to the extent that AI systems will be more cost-effective than humans, then we have a reason to use them instead of humans, even if $\text{Accuracy}_{\text{AI}}$ is lower than $\text{Accuracy}_{\text{Human}}$ and not too low on its own. And this, in conjunction with the nuances just mentioned, is all we need going forward.

3 PREDICTIVE SPEED

The second reason that counts in favor of holding medical AI systems to a lower accuracy standard than humans has to do with the predictive speed that some of them display. Elsewhere, we have argued that predictive speed is sometimes more important than predictive accuracy. Below we rehearse this argument but expand upon its significance to the double standards debate. In Mainz et al. (2022), we argue that predictive speed can be more important than predictive accuracy when the following two criteria are satisfied [18]:

(1) the opportunity costs of a slow decision-making process are severe, and

(2) the consequences of unnecessary treatment are acceptable compared to a late or even absent treatment**.**

Criteria (1) and (2) are often satisfied when we deal with AI systems that can detect and diagnose deadly diseases. Consider an example involving the AI system called Sepsis Watch. Sepsis Watch is a machine learning model that ". . . can predict onset sepsis a median 5 hours before clinical presentation, and it generally performs very well both in terms of triaging and monitoring of patients." [18] A median 5-hour earlier detection time might not seem like a big advancement, but it is. Detecting sepsis just a few hours earlier can make the difference between life and death. Sepsis is an extreme reaction to an infection, and it is one of the most common causes of death globally [15]. The global mortality rate of sepsis is roughly 30%, and every hour of treatment delay increases the risk of death by roughly 8% [32]. So, even if Sepsis Watch can 'only' detect sepsis a few hours earlier than humans, it can potentially mean that thousands, if not millions of people who would otherwise die from sepsis will survive.

The timely detection of sepsis is thus crucial, and as such, the speed by which a sepsis detection model can reach its verdict is an extremely important feature of the AI system. But how can it be more important than the overall accuracy level of the system? The answer lies in the low risks associated with treating individuals for sepsis. The primary treatment options are antibiotics and intravenous fluids. At a population level, an excessive use of antibiotics can of course lead to antibiotics resistance, but when we focus on individual patients, treatments involving antibiotics and intravenous fluids are very safe and have no significant side effects. In other words, even if the sepsis detection model incorrectly diagnoses some patients as having sepsis when they actually do not, the consequences of such mistakes are not very critical for individual patients. Accordingly, even if the false positive rate increases in situations where $\text{Accuracy}_{\text{AI}}$ is lower than $\text{Accuracy}_{\text{Human}}$, the speed of the AI system can still result in a net increase in the amount of lives saved. Interestingly, the same holds true even if the AI model—where $\text{Accuracy}_{\text{AI}}$ is lower than $\text{Accuracy}_{\text{Human}}$—has a higher rate of false negatives compared to human clinicians. This is because the speed at which the AI system can detect sepsis can still lead to saving more patients' lives, despite the increased likelihood of missing some patients who actually have sepsis. Following Mainz et al., we can illustrate the point with the following figure [18]:

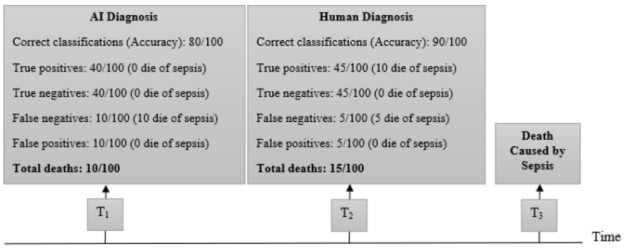

Figure 1: Comparison between an AI diagnosing 100 patients at time T1, and a human doctor diagnosing 100 patients later at T2, Even though the AI performs worse in terms of both accuracy, false negative rates, and false positive rates, the total number of deaths caused by sepsis are still lower for the AI because some of the human doctor's true positives end up dying of sepsis at T3 due to delayed diagnosis and treatment.

The figure should be interpreted as follows. Suppose that 100 patients develop sepsis while they are hospitalized. Normally, human clinicians detect 90 out of 100 cases of sepsis, so $\text{Accuracy}_{\text{Human}} = 0.9$. But because the clinicians can only detect sepsis at time T_2, 15 patients end up dying from sepsis at T_3. 10 of these 15 are among the clinicians' true positives. That is, they do not die because sepsis was not detected, but because it was detected too late. Suppose now that instead of human clinicians, an AI system like Sepsis Watch is introduced to detect sepsis. The AI system is less accurate than the human clinicians: it only detects 80 out of 100 cases of sepsis, so $\text{Accuracy}_{\text{AI}} = 0.8$. The algorithm makes 10 false positive predictions as well as 10 false negative predictions. This is twice as many for each error type compared to the human clinicians. But because the algorithm can detect sepsis already at T_1, the overall survival rate is higher for patients diagnosed by the algorithm than for patients diagnosed by humans. While 10 of the human clinicians' true positives died from sepsis at T_3 because they were diagnosed too late, none of the AI system's true positives end up dying simply because they receive their diagnosis in time.

When we focus on the health of individual patients, the case above shows that algorithmic predictive speed can be more important than algorithmic predictive accuracy. And, of course, this is not only true for sepsis detection models, but in most cases where conditions 1) and 2) are satisfied. In such cases, it will often be justified to hold AI systems to a lower standard of predictive accuracy than humans simply because their ability to deliver faster diagnoses means that more patients will live. Again, it may be that other considerations occasionally trump these considerations of predictive speed, but it remains true that algorithmic predictive speed is one important reason that counts in favor of relying on AI systems for purposes of crucial decision-making even when $\text{Accuracy}_{\text{AI}}$ is lower than $\text{Accuracy}_{\text{Human}}$.

4 CONCLUDING REMARKS

So far, the double standard debate has centered around the issue of transparency. In this paper, we have highlighted that there are other desiderata besides transparency that are subject to similar debates. We have focused on predictive accuracy. In the double standard debate on transparency, some have argued that we should hold AI systems to the same standard as humans, while others have argued that we should hold AI systems to higher standards. When we focus on predictive accuracy in a medical context, we have argued that there are at least two reasons for holding AI systems to lower standards than humans. The first reason is that AI systems can be more cost-effective than humans, and the second reason is that the predictive speed of AI systems can be much higher than that of humans. If we primarily care about producing as much overall health as possible, and saving as many lives as possible, then we have at least two reasons for using AI systems for medical decision-making—even when they are less accurate than human clinicians.

This conclusion is significant for two reasons. First, we claim, it opens a much needed double standard debate about the desideratum of algorithmic predictive accuracy. Second, it takes a new and surprising position in the double standard debate that has so far been vacant: namely that there can be reasons for holding AI systems to lower standards than humans in decision-making. Future research on double standards could with benefit explore further reasons for holding AI systems to lower standards than humans. But it should also explore potential double standards in relation to other desiderata such as fairness and trustworthiness. While we have not said much about these other desiderata in this paper, we suspect that the considerations of cost-effectiveness and predictive speed also count in favor of holding AI systems to lower standards than humans in light of these other desiderata. However, we leave it to another occasion to explore these issues further.

ACKNOWLEDGMENTS

We thank three anonymous reviewers for insightful comments and suggestions. This work was supported by the Carlsberg Foundation (grant number: CF20-0257)

REFERENCES

[1] Alvarado, R. (2021). Should we replace radiologists with deep learning? Pigeons, error and trust in medical algorithm. Bioethics, 36(2), 121-133.
[2] Aristidou, A., Rajesh, J. Topol, E. (2022). Bridging the chasm between algorithm and clinical implementation. The Lancet. 399(10325), P620.
[3] Bennett, C. C., & Hauser, K. (2013). Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. Artificial intelligence in medicine, 57(1), 9-19.
[4] Bjerring, J., Busch, J. (2021). Artificial Intelligence and Patient-Centered Decision-Making. Philosophy &. Technology, 34, 349–371.
[5] Dismuke, C. (2020). Progress in examining cost-effectiveness of algorithm in diabetic retinopathy screening. The Lancet Digital Health, 2(5), e212-e213.
[6] Durán, J., Jongsma, K. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. Journal of Medical Ethics, 47, 329-335.
[7] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., & Dean, J. (2019). A guide to deep learning in Healthcare. Nature Medicine, 25(1), 24-29.
[8] de Fine Licht, K. de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making. algorithm & Society, 1–10.
[9] Gönel, A. (2018). Clinical biochemistry test eliminator providing cost-effectiveness with five algorithms. Acta Clinica Belgica, 75(2):123-127.
[10] Groot, O., Bongers, M., Ogink, P., Senders, J., Karhade, A., Bramer, J., Verlaan, J., Schwab, J. (2020). Does Artificial Intelligence Outperform Natural Intelligence in Interpreting Musculoskeletal Radiological Studies? A Systematic Review. Clinical Orthopaedics and Related Research, 478(12), 2751-2764.
[11] Gruson, D., Bernardini, S., Dabla, P. K., Gouget, B., & Stankovic, S. (2020). Collaborative algorithm and Laboratory Medicine integration in precision cardiovascular medicine. Clinica Chimica Acta, 509, 67-71.

[12] Günther, M., Kasirzadeh, A. (2021). Algorithmic and human decision-making: for a double standard of transparency. AI & Society. Online first.
[13] Horton, J. (2021), Partial aggregation in ethics. Philosophy Compass
[14] Jayakumar, S., Sounderajah, V., Normahani, P. *et al.* (2022). Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. NPJ Digital Medicine, 5(11).
[15] Komorowski M, Celi L, Badawi O, *et al.* (2018). The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine. 24: 1716–1720.
[16] Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., & Novoa, E. M. (2019). Accurate detection of m6A RNA modifications in native RNA sequences. Nature communications, 10(1), 1-9.
[17] London, A. J. (2019). Artificial Intelligence and black-box medical decisions: Accuracy versus explainability. Hastings Center Report, 49(1), 15-21.
[18] Mainz J, Munch L, Bjerring JC, Godtfredsen S. Why algorithmic speed can be more important than algorithmic accuracy. Clinical Ethics. 2022;0(0). doi:10.1177/14777509221138750
[19] Munch, L., Mainz, J. & Bjerring, J.C. The value of responsibility gaps in algorithmic decision-making. Ethics Inf Technol 25, 21 (2023)
[20] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the Electronic Health Records. Scientific Reports, 6 (1).
[21] Mori, T., & Uchihira, N. (2019). Balancing the trade-off between accuracy and interpretability in software defect prediction. Empirical Software Engineering, 24, 2, 779-825.
[22] Nord, E., Daniels, N., & Kamlet, M. (2009). QALYs: some challenges. Value in health, 12, 10-15.
[23] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine, 375(13), 1216.
[24] Ross, A. (2022). AI and the Expert; a blueprint for the ethical use of opaque AI. AI & Society. Online first.
[25] Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harvard Data Science Review, 1(2).
[26] Topol, E. J. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York: Basic Books.
[27] Walmsley, J. (2020). Artificial intelligence and the value of transparency. algorithm & Society, 1–11.
[28] Wilkinson, D. J., & Savulescu, J. (2011). Knowing when to stop: futility in the ICU. Current opinion in anaesthesiology, 24(2), 160–165.
[29] Wilkinson, D., Petrou, S. & Savulescu, J. (2018). Expensive care? Resource-based thresholds for potentially inappropriate treatment in intensive care. Monash Bioeth. Rev, 35, 2–23.
[30] Wolff, J.Pauling, J. Keck, A., Baumbach, J. (2020). The Economic Impact of Artificial Intelligence in Health Care: Systematic Review. Journal of Medical Internet Research, 22(2):e16866.
[31] Xie, Y. *et al.* (2020). Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modeling study. The Lancet, (2), e240– 49.
[32] Zhao, X., Shen, W., Wang, G. (2021). Early prediction of sepsis based on machine learning algorithm. Computational Intelligence and Neuroscience. 6522633.
[33] Zerilli, J., Knott, A., Maclaurin, J. *et al.* (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? Philosophy & Technology, 32, 661–683.