Canadian Journal of Bioethics
Revue canadienne de bioéthique

Commentaire critique / Critical commentary

What Kind of Artificial Intelligence Should We Want for Use in Healthcare Decision-Making Applications?

Jordan Joseph Wadden

Volume 4, numéro 1, 2021

URI : https://id.erudit.org/iderudit/1077636ar
DOI : https://doi.org/10.7202/1077636ar

Aller au sommaire du numéro

Éditeur(s)

Programmes de bioéthique, École de santé publique de l'Université de Montréal

ISSN

2561-4665 (numérique)

Découvrir la revue

Citer ce document

Wadden, J. J. (2021). What Kind of Artificial Intelligence Should We Want for Use in Healthcare Decision-Making Applications? *Canadian Journal of Bioethics / Revue canadienne de bioéthique*, *4*(1), 94–100. https://doi.org/10.7202/1077636ar

Résumé de l'article

La perspective d'inclure l'intelligence artificielle (IA) dans la prise de décision clinique est une prochaine étape passionnante pour certains secteurs des soins de santé. Cet article propose une analyse des types de systèmes d'IA disponibles, en se concentrant sur les caractéristiques de niveau macro. Il examine notamment les forces et les faiblesses des systèmes opaques et des systèmes entièrement explicables. En fin de compte, l'article soutient que les systèmes de type « boîte grise », qui combinent opacité et transparence, devraient être utilisés dans le domaine des soins de santé.

Canadian Journal of Bioethics
Revue Canadienne de Bioéthique

JJ Wadden. *Can J Bioeth / Rev Can Bioeth*. 2021;4(1):94-100

COMMENTAIRE CRITIQUE / CRITICAL COMMENTARY (ÉVALUÉ PAR LES PAIRS / PEER-REVIEWED)

Canadian Bioethics Society / Société canadienne de bioéthique
Concours de rédaction / Writing Contest

What Kind of Artificial Intelligence Should We Want for Use in Healthcare Decision-Making Applications?

Jordan Joseph Wadden[a,b]

Résumé
La perspective d'inclure l'intelligence artificielle (IA) dans la prise de décision clinique est une prochaine étape passionnante pour certains secteurs des soins de santé. Cet article propose une analyse des types de systèmes d'IA disponibles, en se concentrant sur les caractéristiques de niveau macro. Il examine notamment les forces et les faiblesses des systèmes opaques et des systèmes entièrement explicables. En fin de compte, l'article soutient que les systèmes de type « boîte grise », qui combinent opacité et transparence, devraient être utilisés dans le domaine des soins de santé.

Abstract
The prospect of including artificial intelligence (AI) in clinical decision-making is an exciting next step for some areas of healthcare. This article provides an analysis of the available kinds of AI systems, focusing on macro-level characteristics. This includes examining the strengths and weaknesses of opaque systems and fully explainable systems. Ultimately, the article argues that "grey box" systems, which include some combination of opacity and transparency, ought to be used in healthcare settings.

Mots-clés
intelligence artificielle, prise de décision clinique, boîte grise, explicabilité, systèmes opaques

Keywords
artificial intelligence, clinical decision-making, grey box, explainability, opaque systems

Affiliations
[a] Department of Philosophy, University of British Columbia, Vancouver, Canada
[b] Ethics Services, Providence Health Care, British Columbia, Canada
Correspondance / Correspondence: Jordan Joseph Wadden, waddenjordan@gmail.com

BACKGROUND: ARTIFICIAL INTELLIGENCE IN HEALTHCARE

The idea of implementing artificial intelligence (AI) in healthcare is increasingly popular, especially in the areas of decision-making and diagnosis. This is because AI could outperform humans in both speed and accuracy. For example, Scott Mayer McKinney and colleagues demonstrated an AI system that outperformed six doctors in predicting breast cancer and that this system could reduce the workload of a second reader by 88% (1). If this performance is indicative of AI's potential in healthcare, widespread application could drastically and positively change diagnosis and decision-making.

No uniform definition exists for AI on which everyone reliably agrees, but there are generally two or three high-level distinctions to understand these types of technology. The first are reactive systems that are built for a specific purpose, sometimes called "narrow" or "weak" AI. The second are "general" systems, which are able to train on data sets and learn on their own (sometimes these are subsumed into the "narrow" category). The final type of system, called Artificial General Intelligence or "strong" AI, is currently entirely theoretical. These are systems that can replicate autonomous human intelligence (2). The following are some examples of these different kinds of systems, with which the public may be familiar: Stockfish (the chess playing system), IBM's Watson (which was built for Jeopardy but has now been applied in medicine), and HAL (the rogue computer assistant from *2001: A Space Odyssey*).

In this paper I focus on "general" AI. However, despite its potential, "general" AI has not been widely implemented in healthcare decision-making, at least outside of experimental settings or in innovative hospital environments. Rather, most AI in the field falls more or less into the "narrow" category because they are used as diagnostic tools, rather than as decision-makers.

I intend to examine three high-level categories of "general" AI that could be employed in healthcare: opaque systems (sometimes colloquially called "black boxes"), explainable AI (sometimes colloquially called "white boxes"), and semi-transparent systems ("grey boxes"). Opaque systems are those in which the user does not have access to the underlying processes used by the system to reach an output. These are typically considered highly accurate, but at the expense of accountability (3). Explainable AI is a category assigned to those systems that allow a user to clearly explain behaviour, predictions, and influencing variables. These are transparent and accountable, but often not powerful enough to do more than prediction or pattern matching. Finally, the semi-transparent "grey box" is a less discussed category that captures systems lying in between opaque and fully transparent. Despite this in-between category, the debate oftentimes leaves semi-transparent systems out of the discussion and presents a dichotomous choice between opaque or transparent systems. The introduction of grey systems changes the discussion from a dichotomy into a whole spectrum of potential tools.

ISSN 2561-4665

In the sections that follow I highlight the weaknesses with both opaque and transparent categories, ultimately arguing that some form of grey box is the most ethically justified type of AI for healthcare decision-making applications.

DECISION-MAKING OR JUST DIAGNOSTIC TOOL?

To illustrate this project's analysis, it is helpful to draw an analogy to the application of AI in areas where technologies have been implemented, at least to some extent. For example, Eric Topol compares AI in healthcare to AI in self-driving cars (4). Topol draws on an analysis of self-driving cars by Steven E. Shladover in which self-driving cars can be thought of as having six levels of automation: Levels 0 to 5 (5). These range from no automation at all, Level 0, to full automation, Level 5. Topol claims that the automotive industry has their sights set on Level 4, which is high automation with human interaction only in limited circumstances (4). Level 5 is unlikely to be implemented given our current understanding of, and limitations on, AI. Topol suggests that we should aim for Level 3 in healthcare, which is conditional automation with human involvement instead of total machine automation. He believes that we are unlikely to get to Level 4 in healthcare and that we will never get to Level 5; the public would never tolerate a complete lack of oversight by human agents in healthcare situations.

The comparison with self-driving cars seems a good starting point for an analysis of healthcare AI, for several reasons. For starters, it helps to delineate that different applications of AI will have different tolerable levels of automation. Distinguishing between the different levels helps to situate what theories can be applied in healthcare settings and the acceptable and appropriate "gaps" for research. Additionally, this also illustrates that there will likely be different tolerable levels of AI in different aspects of healthcare – perhaps we *can* tolerate Level 4 in electronic health records while in surgery we only ever want Level 2.

Some of the following examples will clarify this "levels" analysis. Currently, radiology departments are a potential area for the implementation of healthcare AI (6,7). Zeynettin Akkus and colleagues, for example, successfully demonstrated that the texture of brain magnetic resonance imaging (MRI) could predict the genomic anomaly of a co-deletion of chromosomal arms 1p/19q, which is crucial information for effective treatment of low-grade gliomas (a type of brain cancer) (8). Typically, assessment for such deletions requires a biopsy, but this team used a convolutional neural network to accurately predict this co-deletion, non-invasively. Here, AI acts in the capacity of a diagnostic tool – the "decision" returned is whether or not the patient suffers the co-deletion. Accordingly, this AI system would be judged lower on the automation scale; the human agent still evaluates the system's verdict and decides the next steps to take for the patient.

Manisha Bahl and colleagues, in another example, have demonstrated a proof of concept machine learning model that could predict which high-risk breast lesions need surgical excision and which lesions are a low risk for upgrade to cancer (9). Using such a system we could avoid removing benign lesions that would otherwise only be diagnosed as benign after removal. The authors claim that such an AI could allow, if implemented on a wide scale, for a reduction in unnecessary surgery (substituted instead for surveillance). Accordingly, this AI acts in more of a decision-making capacity. The "decision" here is not just whether the lesion will upgrade to cancer – rather, the "decision" includes whether the lesion should be surgically excised. This would be judged higher on the automation scale, as the human agent could choose to take the system's decision at face value and only operate where the AI indicates.

Both of the above examples are interesting, each raising their own ethical questions. For the purpose of this paper, I want to focus on healthcare AI systems with decision-making capacity, rather than just diagnostic capacity.

THE FIRST OPTION: OPAQUE SYSTEMS IN HEALTHCARE

David S. Watson and colleagues raise two important questions that come from the introduction of opaque systems in healthcare decision-making. First, they ask "If doctors do not understand why the algorithm made a diagnosis, then why should patients trust the recommended course of treatment?" (10). This question hinges on the deference to authority that needs to occur for many medical decisions to succeed. What it means for someone not to understand an opaque system is not always clear. It seems rather straight forward to think that if a system was created by a programmer, then that programmer should be able to understand it. Yet, opaque systems are more complex. For example, in machine learning systems, the developers decide on basic architectural principles and appropriate learning algorithms. However, developers do not decide which values, parameters, or connections the system applies, nor do they directly determine how a verdict is rendered (11). Disconcertingly, these opaque systems may identify patterns and connections that users have no ability to access. Consequently, it may be possible to explain how the system made these connections while simultaneously being unable to understand why the system made these connections.

Accordingly, there is an epistemic gap present at the time of application; an opaque system poses problems because the individual employing the systems lacks adequate understanding of the inner workings of the system. This poses a potential problem for healthcare, because we lack information for decision-making purposes that may be relevant to the consult, patient, team, etc. Without this information, a patient or clinical team may not be able to bring themselves to trust the process.

Trust in healthcare is a major factor in effective healthcare. According to Mark A. Hall and colleagues, trust matters for both intrinsic and instrumental reasons. Intrinsically, trust is one of the key elements that give the patient-professional relationship its meaning, importance, and substance. Whereas, instrumentally, they claim that trust has been hypothesized or shown to affect the willingness of a patient, for the following acts: seek care, reveal sensitive information, submit to treatment, participate in research, and recommend physicians to friends (12). Other reasons that trust can be negatively affected by opaque systems include a lack of explainability and the potential for bias that comes from their use (13).

So, if opaque systems present the risk of a loss of trust and can make decisions without a method to check for bias, then there is a significant reason to be cautious of their role in medicine. This is especially true when we consider that there are various populations who are already wary of the healthcare system as a whole (e.g., historically oppressed and targeted groups, such as First Nations or Black individuals).

One might counter this argument by affirming that the patient can trust the healthcare professional, which by extension allows for trust in the AI. The objection could argue along the lines that surely *the professional* would not be employing the system if *they* did not trust it to deliver an accurate diagnosis. Accordingly, a patient ought to be able to trust the system by proxy. Unfortunately, trust is not something that is so easily transferred. We can easily imagine a patient who trusts the professional in most circumstances but fails to trust them whenever they outsource part of the decision-making process to an AI system. The fact that an AI system was involved in the decision-making process might be non-negotiable for some or a cause for concern for others. In other words, these patients may accept AI as a diagnostic tool but refuse to consent to treatment if an AI was making the decision.

The second question that Watson and colleagues ask is whether "informed consent is even possible without some grasp of how the model reached its conclusion?" (10). Would this consent meet the ethical or legal standards we expect in healthcare? (14,15). The authors explain that they are concerned about consent because the nature of opaque systems means we necessarily cannot know how the systems are learning or drawing connections. Without this knowledge, the healthcare professional using an AI system cannot explain to a patient why the decision has been made. According to this argument, the "why" is a necessary component of *informed* consent.

Informed consent is a particularly important issue given some of the foundations of medicine. Informed consent is best described as an individual's autonomous authorization of a medical intervention or participation in some study. This brief description illustrates that informed consent is more than just basic 'consent'. Tom L. Beauchamp and James F. Childress articulate that informed consent has the following elements: competence (to understand and decide), voluntariness (in deciding), disclosure (of material information), recommendation (of a plan), understanding (of disclosure and recommendation), decision (in favour of/against a plan), authorization (of the chosen plan) (16).

Opaque systems, accordingly, pose a particular problem for disclosure and understanding, both of which are part of the *informational* components of informed consent. Indeed, opacity makes it difficult for a healthcare professional to disclose why a system makes a decision, and it stands in the way of a patient understanding why the system makes its decision. A patient can reasonably be deemed competent enough to make a decision and able to voluntarily make a choice, but without proper information they cannot provide informed consent (decide and authorize). At best, in these circumstances, we might be able to claim the patient is assenting, but this does not hold the same legal or procedural weight as informed consent.

Thus, opaque systems fail to meet the two desirable features outlined above: trust and informed consent. It might be the case that there are additional features they fail as well; however, I believe demonstrating these two failures are sufficient for my analysis. This is not to say that all applications of opaque systems are inappropriate. It is quite possible that opaque style systems may be completely appropriate, or that these two criteria do not hold the same weight, in other domains such as in automotive applications. But, given these concerns, it appears that opaque systems would be inappropriate for clinical decision-making.

THE SECOND OPTION: FULLY EXPLAINABLE ARTIFICIAL INTELLIGENCE

Since opaque systems are likely inappropriate for healthcare decision-making, let us turn to the next option: complete explainability. Fully explainable artificial intelligence (XAI) systems, or white box systems, provide a stark alternative to opaque systems (17). There are two main criteria for XAI: these systems must be understandable (18,19), and these systems must be transparent (11,20). While these criteria might sound similar, perhaps even two ways of making the same claim, they are in fact separate considerations. First, regarding understandability, we need to know how to interpret all pieces of the system that are accessible to the user and how to interpret the verdicts the system develops. This criterion alone is reasonably shared with opaque systems. Regarding transparency, however, the claims are that all pieces involved in the system need to be accessible to the user. Therefore, combining these two criteria, all components of a fully realized XAI system are understandable.

The positive argument for XAI in healthcare decision-making follows from these two criteria. If all accessible pieces of the system are transparent, then we can locate and isolate how the system learns, makes connections, and interprets data. If all accessible pieces of the system are understandable, then we are able to interpret how and why each connection is made in the learning process. This transparency could thus give users the ability to reason through and justify the AI recommendations – an ability that is not present in opaque systems.

Transparency is difficult to define in AI systems, comprising as it does many different considerations. Justin B. Biddle explains that transparency might mean "discoverable" – another ill-defined term – for some, while others intend it to just mean meeting particular benchmarks. Others may even submit that transparency only requires information about training metrics to be made available to users (21). Another layer of complexity that I think needs to be considered is the question of who a system needs to be transparent to – the user? the developer? etc. – and whether this transparency needs to reflect only explainability (the "how" it works) or reflect understandability (the "why" it delivers a given answer).

Additionally, full XAI might just be demanding too much for medical applications, and therefore go beyond what is reasonable or even desirable. Alex John London asserts that "decisions that are atheoretic, associationist, and opaque are commonplace in medicine" (22). Medicine mostly focuses on reliably arriving at a treatment outcome rather than the processes or mechanisms explaining why a particular outcome works. London further argues that the over-reliance on theories that explain why something might be the case can sometimes result in more difficulty treating an individual patient. In a way, this line of reasoning mirrors the shift toward evidence-based medicine in the 1990s.

It might not even be possible – at present or in the future – to fully explain how a system makes decisions in certain contexts in a way we can understand or learn from. When discussing radiology, for example, the images radiologists must decipher can sometimes be very difficult to read for even the most experienced professional. In these cases, the professional needs to determine which interpretation is most likely based on history, symptomology, and possibly other tests. An AI system might be able to decipher images with high levels of accuracy but be unable to explain them in a way our human eyes can perceive.

For this reason alone, I think that we should be wary of demanding full XAI in healthcare applications. As I mentioned above, there may be other contexts where this might be the most appropriate, such as if we want to use AI in public policy decision-making. Additionally, when we prioritize explainability we have to make trade-offs between power and accuracy, which could affect clinical usefulness. Power in this case refers to the system's ability to handle more complex processes while returning usable outputs. For example, random forest classifiers are able to perform more complex processes but are much less explainable when compared to, say, regression algorithms.

For the professional, full XAI may turn out to be non-beneficial. If two systems are designed under the same complexity, data, and constraints, and the professional *also has to explain the system's processes*, this system will lose some of its power. Indeed, the system has to use energy and time to interpret and explain the process it used to reach its decision, rather than only running processes and developing connections. But the problems may extend further – it is possible that there are some decision-making procedures that only opaque systems can make but which are incomprehensible for humans. A full XAI system would be much weaker in this regard as well because there is nothing it could do to make these essentially incomprehensible processes transparent.

THE CRITERIA SO FAR

The main criteria that have been identified so far are as follows:

A. AI systems in healthcare must not jeopardize informed consent.
B. AI systems in healthcare must not jeopardize patient trust.
C. AI systems in healthcare must not demand more explainability than is tolerable by the healthcare system.

Opaque systems fail (A) and (B) because, by design, we cannot make sense of the connections it draws between input data and the nodes it has developed using training data. Full XAI systems fail (C) because, by design, everything is explainable and this affects the resources that can be used for powerful computing. This leaves us with a question: is there anything that can satisfy all three criteria?

THE THIRD OPTION: SEMI-TRANSPARENT "GREY BOXES" TO THE RESCUE?

The last option I want to analyze are semi-transparent systems, sometimes also called grey box systems. These are AI systems that have reduced explainability (to some extent), and accordingly allow for some power-explainability trade-offs. James C. Christensen and Joseph B. Lyons define a grey box as a system that provides: "[S]ufficient information about the learning technology to establish trust wherein, much like with humans, we trust based on the synthesis of predictability, feasibility, and inference of intent based on one's knowledge of the goals, values, and interaction with the system." (23) They claim these systems are much like humans insofar as their reasoning is never fully knowable, but there can be trusted processes which reduce uncertainty and increase the understanding of rationales between agents.

Comparing grey boxes to human colleagues addresses an interesting objection against the view I presented above regarding opaque systems. I submitted that opaque systems are identical to (strong stance) or significantly similar to (weak stance) human minds because neither are transparent to a third party. Therefore, any arguments lodged against opaque systems should be lodged against human explanations as well. And this would mean that no procedure currently employed leads to informed consent or sufficient trust. Therefore, this objection continues, if we accept the above arguments against opaque systems, then we must also hold that human explanations are inappropriate for healthcare.

Instead, comparing human colleagues to grey boxes demonstrates a decent level of transparency in our interactions with colleagues. The introduction of semi-transparent systems effectively responds to this opacity objection. It is true that I can never know the full details or justification of why my colleague presents me with a recommendation. They may not even have full access to why they made a decision. However, and unlike opaque systems, there is *incomplete* detail and justification available – a colleague can explain why they reached their conclusion without knowing what pieces of their background led them to make the relevant connections. So, the objection against opaque systems fails because, upon closer inspection, human minds are semi-transparent rather than opaque.

The comparison here is also interesting because it allows for the possibility that radiologists do not 'know' what they are doing from an epistemological point of view. Instead, this may be one of the many jobs where you 'learn' the theory and then *learn* on the job through doing. Additionally, comparing grey boxes to humans indicates another interesting possibility – namely, that down the road, a so-called black box may be able to shift into a grey box. For example, if a radiologist has developed tacit knowledge from examining thousands of images, they might not be able to explain this to another professional through verbalizing or writing it down as instructions, because it is *tacit* knowledge. However, if they are able to reflect and codify this information somehow, then the tacit knowledge loses its 'black box' characteristic. Perhaps the same could be true about some of the less complex opaque systems if we develop new techniques for interpreting data.

We believe many things without fully knowing the justification for or having a complete understanding of the reasoning that underpin a phenomenon. For example, we often make decisions about what the best course of treatment is for a patient despite knowing that we do not have perfect information on their condition. In these situations, we rely on our best interpretation of the data we can collect.

A thought experiment might be helpful to appreciate how semi-transparent decision-making systems seem, I think, intuitively better than opaque or full XAI systems. Suppose that we are presented with a particularly difficult case and suppose that we can get the assistance of three colleagues to help us come to a decision. Unfortunately, after asking the first colleague for help, the other two will have to respond to another consult – so, we have to choose wisely who to ask.

- Colleague B is known for her astounding accuracy and ability to identify problems. However, she always walks away before explaining herself, and many patients refuse to accept her suggestions.
- Colleague W is much slower in his decision-making. He often will take minutes to hours thinking over a single question to make sure he's considered everything. However, his meticulousness ensures that you and the patient understand his reasoning. Patients like him and happily accept his recommendations… when he finally makes them.
- Colleague G is pretty quick in her responses, but slower than B. She's also pretty good at explaining her recommendations, like W, but she sometimes uses jargon too difficult even for you to understand. Most patients like her and will accept her advice.

These three colleagues are analogues for the AI systems that I discussed above. If my arguments have been sufficiently persuasive, then the best colleague/system to engage for help is Colleague G/grey boxes. Indeed, they meet all three criteria: they do not jeopardize informed consent or trust, and their explanations are understandable without reflecting a burden on time or other resources, at least to some extent.

AN OBJECTION

London notes that ceding to the 'unknown' can create problems for healthcare (22). He explains that "some commentators regard ceding medical decision-making to black box systems as contravening the profound moral responsibilities of clinicians" (15). Following from his argument regarding black boxes, the conclusion as to how this problem extends to grey boxes is straightforward as well. Indeed, their nature is *partially opaque*. The reason that this problem rests on moral responsibility is because an AI that operates on anything less than full explainability cannot give complete reasons for the verdicts and decisions it recommends.

The problem here can be shown through an analogy. When a healthcare professional relies on a colleague for help in coming to a decision, they can question their interlocutor to get more information. But, the objection goes, when a healthcare professional relies on an AI system that is not *fully* explainable, they cannot ask clarificatory questions (or, if they can, they are significantly limited based on what components are transparent and which are opaque). Accordingly, the objection is that my analogy between Colleague G and grey boxes breaks down. Therefore, the responsible approach is to sacrifice the predictive power and enhanced accuracy in both opaque and semi-transparent systems in favour of the simplicity and full access to knowledge present in XAI – effectively mimicking the 'actual' inter-personal relationship that can occur between colleagues.

I think that this kind of argument is overly cynical. Grey box systems can be explainable or opaque to varying degrees, depending on what we decide needs to be transparent and understandable. Broadly claiming that they suffer from the same problem as do opaque systems minimizes this apparent customizability. I think that grey box systems are the ideal we should strive for in healthcare AI because they have some of the power of opaque systems while retaining some of the explainability of XAI. As mentioned above by Christensen and Lyons, this mirrors how human minds communicate decisions (23). When we ask questions of our colleague in hopes for more information, we are not accessing their mind in a fully explainable or direct way. Instead, we are interpreting their speech which has resulted from some series of thoughts we have no ability to access. Asking questions of our colleagues does not turn them into XAI analogs; instead, employing grey box systems would most closely mirror current practices used for seeking assistance in decision-making. As such, and because the greyness could be altered depending on application, I think this approach avoids much of the trustworthiness, consent, and demand worries that plague opaque and fully explainable AI systems.

CONCLUSION

I believe I have shown that for healthcare we ought to aim for grey box AI systems because they are capable of addressing the main criticisms of both opaque and full XAI systems, all while best mirroring how clinical decision-making currently functions. An additional benefit is that grey boxes can come in a range of shades, rather than the stark all or nothing options presented in opaque boxes or full XAI. This customizability is a unique feature that could allow for greater application of AI in healthcare, as tailored decision-making systems could be incorporated into more nuanced areas of clinical practice.

Reçu/Received: 29/04/2020

Publié/Published: 01/06/2021

Remerciements
Je tiens à remercier tout particulièrement la Société canadienne de bioéthique d'avoir organisé le concours de rédaction pour étudiants qui a donné lieu à cette version de mon manuscrit. En outre, je tiens à remercier les rédacteurs et les réviseurs de la Revue canadienne de bioéthique pour leur temps, leur considération et leurs suggestions. Je tiens également à souligner les suggestions qui m'ont été faites par plusieurs membres du département de philosophie de l'UBC, en particulier celles de Daniel Steel, Christopher Mole et Jonathan Ichikawa. Cet article s'appuie sur des recherches financées par une bourse de doctorat du Conseil de recherches en sciences humaines.

Acknowledgements
I want to give special acknowledgement to the Canadian Bioethics Society for hosting the student writing competition that resulted in this version of my paper. In addition to this, I want to thank the editors and reviewers at the Canadian Journal of Bioethics for their time, consideration, and suggestions. I also want to recognize the suggestions given to me by several members of the UBC Department of Philosophy, especially those from Daniel Steel, Christopher Mole, and Jonathan Ichikawa. This paper draws on research supported by a Social Sciences and Humanities Research Council Doctoral Fellowship.

Conflits d'intérêts
Aucun à déclarer

Conflicts of Interest
None to declare

Édition/Editors: Erica Monteferrante & Aliya Affdal

Les éditeurs suivent les recommandations et les procédures décrites dans le Code of Conduct and Best Practice Guidelines for Journal Editors de COPE. Plus précisément, ils travaillent pour s'assurer des plus hautes normes éthiques de la publication, y compris l'identification et la gestion des conflits d'intérêts (pour les éditeurs et pour les auteurs), la juste évaluation des manuscrits et la publication de manuscrits qui répondent aux normes d'excellence de la revue.

The editors follow the recommendations and procedures outlined in the COPE Code of Conduct and Best Practice Guidelines for Journal Editors. Specifically, the editors will work to ensure the highest ethical standards of publication, including: the identification and management of conflicts of interest (for editors and for authors), the fair evaluation of manuscripts, and the publication of manuscripts that meet the journal's standards of excellence.

Évaluation/Peer-Review: Anonymous & Daniel Schönberger

Les recommandations des évaluateurs externes sont prises en considération de façon sérieuse par les éditeurs et les auteurs dans la préparation des manuscrits pour publication. Toutefois, être nommé comme évaluateurs n'indique pas nécessairement l'approbation de ce manuscrit. Les éditeurs de la Revue canadienne de bioéthique assument la responsabilité entière de l'acceptation finale et de la publication d'un article.

Reviewer evaluations are given serious consideration by the editors and authors in the preparation of manuscripts for publication. Nonetheless, being named as a reviewer does not necessarily denote approval of a manuscript; the editors of *Canadian Journal of Bioethics* take full responsibility for final acceptance and publication of an article.

REFERENCES

1. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.
2. Coleman F. A Human Algorithm: How Artificial Intelligence is Redefining Who We Are. Berkley, California: Counterpoint; 2019.
3. Tannam E. What are the benefits of white-box models in machine learning? Silicon Republic. 20 Feb 2019.
4. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York, NY: Basic Books; 2019.
5. Shladover SE. The truth about "self-driving" cars. Scientific American. Dec 2016.
6. Chockley K, Emanuel E. The end of radiology? Three threats to the future practice of radiology. Journal of the American College of Radiology. 2016;13(12):1415-1420.
7. Recht M, Bryan RN. Artificial intelligence: threat or boon to radiologists? Journal of the American College of Radiology. 2017;14(11):1476-1480.
8. Akkus Z, Ali I, Sedlář J, et al. Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence. Journal of Digital Imaging. 2017;30(4):469-476.
9. Bahl M, Barzilay R, Yedidia AB, et al. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. Radiology. 2017;286(3):810-818.
10. Watson DS, Krutzinna J, Bruce IN, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ. 2019;364:l886.
11. Zednik C. Solving the black box problem: a normative framework for explainable artificial intelligence. arXiv:1903.04361 [cs.GL]; 4 Jul 2019.
12. Hall MA, Dugan E, Zheng B, Mishra AK. Trust in physicians and medical institutions: what is it, can it be measured, and does it matter? The Milbank Quarterly. 2001;79(4):613-639.

13. Nundy S, Montgomery T, Wachter RM. Promoting trust between patients and physicians in the era of artificial intelligence. JAMA. 2019;322(6):497-498.
14. Schiff D and Borenstein J. How should clinicians communicate with patients about the roles of artificially intelligent team members? AMA Journal of Ethics. 2019;21(2):E138-145.
15. Cohen IG. Informed consent and medical artificial intelligence: what to tell the patient? Georgetown Law Journal. 2020; 108:1425-1469.
16. Beauchamp TL, Childress JF. Principles of Biomedical Ethics (7th ed.). New York, NY: Oxford University Press; 2013.
17. The Lancet Respiratory Medicine. Opening the black box of machine learning. The Lancet Respiratory Medicine. 2018;6(11):801.
18. Doran D, Schulz S, Besold TR. What does explainable AI really mean? a new conceptualization of perspectives. arXiv:1710.00794. 2 Oct 2017.
19. Hsu W, Elmore JG. Shining light into the black box of machine learning. Journal of the National Cancer Institute. 2019; 111(9):877-879.
20. Wellner G, Rothman T. Feminist AI: can we expect our AI systems to become feminist? Philosophy & Technology. 2019;33:191-205.
21. Biddle JB. On predicting recidivism: epistemic risk, tradeoffs, and values in machine learning. Canadian Journal of Philosophy. 2020; First view. 1-21.
22. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Center Report. 2019;49(1):15-21.
23. Christensen JC, Lyons JB. 2017. Trust between humans and learning machines: developing the gray box. Mechanical Engineering. 2017;139(6):S9-S13.