Document generated on 11/14/2026 6 a.m.

Canadian Journal of Bioethics
Revue canadienne de bioéthique

Naming and Diffusing the *Understanding Objection* in Healthcare Artificial Intelligence

Jordan Joseph Wadden

Volume 7, Number 4, 2024

URI: https://id.erudit.org/iderudit/1114958ar
DOI: https://doi.org/10.7202/1114958ar

See table of contents

Publisher(s)

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

ISSN

2561-4665 (digital)

Explore this journal

Cite this article

Wadden, J. J. (2024). Naming and Diffusing the *Understanding Objection* in Healthcare Artificial Intelligence. *Canadian Journal of Bioethics / Revue canadienne de bioéthique*, *7*(4), 57–63. https://doi.org/10.7202/1114958ar

Article abstract

Informed consent is often argued to be one of the more significant potential problems for the implementation and widespread onboarding of artificial intelligence (AI) and machine learning in healthcare decision-making. This is because of the concern revolving around whether, and to what degree, patients can understand what contributes to the decision-making process when an algorithm is involved. In this paper, I address what I call the *Understanding Objection*, which is the idea that AI systems will cause problems for the informational criteria involved in proper informed consent. I demonstrate that collaboration with clinicians in a human-in-the-loop partnership can alleviate these concerns around understanding, regardless how one conceptualizes the scope of understanding. Importantly, I argue that the human clinicians must be the second reader in the partnership to avoid institutional deference to the machine and best promote clinicians as the experts in the process.

© Jordan Joseph Wadden, 2024

This document is protected by copyright law. Use of the services of Érudit (including reproduction) is subject to its terms and conditions, which can be viewed online.

https://apropos.erudit.org/en/users/policy-on-use/

This article is disseminated and preserved by Érudit.

Érudit is a non-profit inter-university consortium of the Université de Montréal, Université Laval, and the Université du Québec à Montréal. Its mission is to promote and disseminate research.

https://www.erudit.org/en/

Canadian Journal of Bioethics
Revue Canadienne de Bioéthique

JJ Wadden. *Can J Bioeth / Rev Can Bioeth.* 2024;7(4):57-63

ARTICLE (ÉVALUÉ PAR LES PAIRS / PEER-REVIEWED)

Naming and Diffusing the *Understanding Objection* in Healthcare Artificial Intelligence

Jordan Joseph Wadden[a,b,c]

Résumé
Le consentement éclairé est souvent considéré comme l'un des problèmes potentiels les plus importants pour la mise en œuvre et l'intégration généralisée de l'intelligence artificielle (IA) et de l'apprentissage automatique dans la prise de décision en matière de soins de santé. En effet, on se demande si, et dans quelle mesure, les patients peuvent comprendre ce qui contribue au processus de prise de décision lorsqu'un algorithme est impliqué. Dans cet article, j'aborde ce que j'appelle *l'objection de la compréhension*, c'est-à-dire l'idée que les systèmes d'IA causeront des problèmes pour les critères d'information impliqués dans un consentement éclairé approprié. Je démontre que la collaboration avec les cliniciens dans le cadre d'un partenariat humain dans la boucle peut atténuer ces préoccupations concernant la compréhension, quelle que soit la manière dont on conceptualise la portée de la compréhension. Surtout, je soutiens que les cliniciens humains doivent être le deuxième lecteur dans le partenariat afin d'éviter la déférence institutionnelle envers la machine et de promouvoir au mieux les cliniciens en tant qu'experts dans le processus.

Mots-clés
intelligence artificielle, éthique clinique, consentement, prise de décision clinique, radiologie, compréhension, humain dans la boucle

Abstract
Informed consent is often argued to be one of the more significant potential problems for the implementation and widespread onboarding of artificial intelligence (AI) and machine learning in healthcare decision-making. This is because of the concern revolving around whether, and to what degree, patients can understand what contributes to the decision-making process when an algorithm is involved. In this paper, I address what I call the *Understanding Objection*, which is the idea that AI systems will cause problems for the informational criteria involved in proper informed consent. I demonstrate that collaboration with clinicians in a human-in-the-loop partnership can alleviate these concerns around understanding, regardless how one conceptualizes the scope of understanding. Importantly, I argue that the human clinicians must be the second reader in the partnership to avoid institutional deference to the machine and best promote clinicians as the experts in the process.

Keywords
artificial intelligence, clinical ethics, consent, clinical decision-making, radiology, understanding, human-in-the-loop

Affiliations
[a] Centre for Clinical Ethics, Unity Health Toronto, Toronto, Ontario, Canada
[b] Centre for Clinical Ethics, Scarborough Health Network, Scarborough, Ontario, Canada
[c] Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada

Correspondance / Correspondence: Jordan Joseph Wadden, waddenjordan@gmail.com

INTRODUCTION

Informed consent tends to be recognized as one of the more significant potential problems for the implementation and widespread onboarding of artificial intelligence (AI) in decisional capacities. Consent is both a legal and an ethical construction. As a legal construct, free and informed consent is a legal obligation for anyone – including healthcare professionals whose aim is to benefit a patient – who desires to infringe on another person's bodily rights. As it is most commonly understood as an ethical construct, informed consent comes from Beauchamp and Childress' list of criteria including: competence, voluntariness, disclosure, recommendation of a plan, understanding, expressing a decision, and authorization of the plan (1). Practically, these criteria are taken to involve things like the capacity to follow an explanation of one's condition, need for diagnosis, treatment plan, etc., or the ability to ask meaningful questions about the information that has been presented that are not simply answered by restating the facts themselves.

The issue of consent with advanced AI has been raised in the literature (2-6), as well as offline at many conferences I have attended and in discussions with colleagues. I seek to address these concerns by exploring whether the development and use of advanced AI jeopardizes informed consent. I argue that advanced AI systems need not jeopardize consent in any meaningful way. I start by defining the *Understanding Objection* that I believe is lobbied against advanced AI in healthcare, which I call an implicit objection because it has not been directly claimed yet instead exists as an amalgam of views presented in the literature and in lectures across various venues. This objection focuses on the idea that AI systems, no matter how they are constructed, will cause problems for the informational criteria involved in proper informed consent, namely around patient understanding.

Importantly, for the purpose of responding to concerns in the literature, I assume no specific scope of patient understanding. I do not tie my argument to whether understanding ought to be something general or whether it ought to be about nuanced or include technical details. Instead, I intend to show that however one uses the term "understanding" it is not inherently an impediment to the safe use of clinical AI.

2024 JJ Wadden. Creative Commons Attribution 4.0 International License
ISSN 2561-4665

My analysis focuses primarily on defeating this *Understanding Objection* through promoting a human-in-the-loop partnership with the clinician, rather than leaving the tool to examine data and recommend something for the clinician to consider while in isolation. I accomplish this through drawing on current practices in radiology and how they might be applicable to AI beyond radiological imaging algorithms. An important aspect of my analysis is the order in which I believe this human-in-the-loop partnership needs to occur. I will argue that if the human acts as second reader, then we can benefit from the power of AI algorithms in decision-making without risking deference to technology which cannot appreciate nuance and context of all other medical factors at play in a patient's case.

Before moving to the *Understanding Objection*, and my response, I first want to provide some explanations of terminology I will employ throughout. These terms are "Black Box AI" and "Explainable AI." Black boxes are systems which are functionally opaque – in other words, the user does not have access to the underlying processes used by the system to reach its output (7). Developers create these systems, but the machine then learns through complex means that ultimately cannot be clearly explained to the user. Due to this lack of explanation, one of the biggest problems is that the user is unable to understand the system – in healthcare, this would have the further implication that if the user (physician) does not understand, then there is no way to consistently ensure that the patient understands enough to provide informed consent (8).

Explainable AI, also sometimes referred to as XAI, is at the exact opposite end of the spectrum. These are systems which are open and transparent to the user – in other words, a user can see how the system has made its decisions. These come in degrees of explainability, sometimes referred to as "white" and "grey" boxes, a nod to the previously established black box category (7). Transparency and explainability are difficult terms to define, as everyone seems to use them a little differently. A good way to understand the difference is that transparency entails that the intricacies of the decision-making are visible, while explainable entails that whatever reasoning is visible can be understood by a user. One of the biggest problems with these XAI systems is the potential threat to privacy. When a significant amount, or all, of the data is transparent and explainable, there is a possibility for the misuse of data. This is particularly true of images, such as x-rays, which can include artefacts irrelevant to care but be exploitable by bad actors in the hospital (e.g., breast images, or a Throckmorton sign).[1]

THE UNDERSTANDING OBJECTION

A number of scholars have discussed the idea that medical AI, due to its opaqueness and non-understandability, would threaten informed consent. For instance, Watson and colleagues ask whether consent is even possible without some understanding of how the algorithm reaches its conclusions (5). Moreover, they suggest a dilemma between black boxes and robust XAI:

> [M]any popular machine learning algorithms are essentially black boxes—oracular inference engines that render verdicts without any accompanying justification. This problem as become especially pressing with passage of the European Union's latest General Data Protection Regulation (GDPR), which some scholars argue provides a "right to explanation" […] Some caution with regard to transparency[2] is advisable. A fully open source approach may enable misuse of the algorithm for harmful purposes outside the clinical context. […] Scrutiny of machine learning is important but should not expose people to disproportionate risks or privacy violations. (5)

They then suggest that the way to balance these concerns is to aim for a granular approach to transparency. This would entail designing a system in a way that allows some users to receive diagnoses explained by referencing basic or familiar biological concepts, while others using the same system may want and so receive a more detailed account in terms of mechanisms. Such a granular approach would be supported by legal recognition in many jurisdictions that informed consent requires a degree of modularity in what information is provided to the patient. This modularity is meant to ensure that patients receive the amount of information they individually require to ensure the criteria of informed consent are adequately met. However, Watson and colleagues also argue that we need to focus on the utility of a given explanation before we move down this path regarding individual user preferences. Unfortunately, their interpretation of utility it is focused on users – in other words, the clinicians – rather than the subjects (i.e., patients) of medical interventions who ultimately are the ones who need to provide informed consent and therefore need to demonstrate understanding.

Schiff and Borenstein discuss this argument by dilemma more directly. They explain that consent requires a patient understand relevant information, which to them includes the purpose of a procedure, potential benefits, potential risks, alternative treatment options, and so on (4). However, they state that with AI devices there are additional demands:

[1] The "Throckmorton sign", also known as a "John Thomas sign," is a slang term for an unscientific and humorous artefact on x-rays where a patient's penis points in the direction of unilateral disease. This could constitute a privacy breach since the patient's identifiers may be in the image, and the AI may therefore incorporate data about a specific patient's genitals into its learning algorithm. When in an XAI system, depending on the levels of transparency or explainability, this image could be available for anyone studying its decision-making reasoning at any later stage.

[2] Watson and colleagues use "transparency" and "explainability" interchangeably. I believe this is a mistake in how we discuss advanced AI systems as it enables clinicians, researchers, and policy-makers to talk past each other while believing they are using the same definitions.

> When an AI device is used, the presentation of information can be complicated by possible patient and physician fears, overconfidence, or confusion. Moreover, for an informed consent process to proceed appropriately, it requires physicians to be sufficiently knowledgeable to explain to patients how an AI device works, which is rendered difficult by the black-box problem. (4)

They conclude by stating that it may not be possible to consent if a physician does not "fully understand" how to explain an AI system's predictions and errors. Others, such as Lee and colleagues have echoed these strong calls for explainability by claiming that "[d]espite the clear trade-off between accuracy and explainability, explainable models are needed to ensure safety of the patients and establish trust in the AI models" (6). In their view, then, while black box systems are more accurate, they compromise understanding and therefore cannot be justified.

This full understanding is an enormous burden, though, and may itself be problematic for patient privacy. Others, such as London, have claimed that fully understandable systems are not acceptable because they are too demanding for medical applications (3). Some in the field have gone as far as to say that transparency is fundamentally at odds with the objectives of AI. Lipton writes:

> Some arguments against black-box algorithms appear to preclude any model that could match or surpass human abilities on complex tasks. As a concrete example, the short-term goal of building trust with doctors by developing transparent models might clash with the longer-term goal of improving health care. Be careful when giving up predictive power that the desire for transparency is justified and not simply a concession to institutional biases against new methods. (2)

Again, we see a dichotomy: fully understood systems, which cannot be implemented or which may lack predictive accuracy, versus black box systems, which cannot be explained. Taken together, this literature raises questions about whether ethical applications of medical AI are possible.

My brief exposition here demonstrates, first, that the threat to informed consent due to a lack of understandability is a concern discussed in the ethical literature regarding medical AI, and second, that discussions of this topic often present a dilemma between opaque black boxes and robust XAI systems. If black boxes are incompatible with informed consent, while XAI systems are infeasible or suffer from other ethical problems, then we are left in a state where it is unclear how healthcare AI systems can ethically proceed. This is the basis of the *Understanding Objection*. According to this argument, black boxes are precluded because a patient cannot give informed consent due to lack of understanding, while robust XAI is too onerous to be used in medical practice and may suffer from other ethical drawbacks relating to privacy or reduced accuracy. So, the *Understanding Objection* concludes that neither option provides a stable ethical basis for AI in diagnostics or treatments.

UNDERSTANDING AND A FORGOTTEN COLLABORATOR

In response, I believe we ought to recognize the role the clinician should still play as a "second reader" to the system's diagnostic decisions. The concept of first and second readers comes from radiological practice where there are two primary scenarios for the reading of a radiograph: a single-reading and a double-reading (9). In a single-reading scenario there is typically only one clinician who performs the diagnostic analysis. Sometimes there is one reader who is assisted by computer aided diagnostic software (CAD). If there is concern raised, perhaps by the ordering clinician or by the uncertainty of the radiologist themselves, then a second consensus reader is consulted. In some settings, this consensus read is purely for statistical and quality assurance purposes and does not impact the patient's care. For a double-reading, there are more options. The readers can be conducting their analysis simultaneously, serially but with the second unaware of the first's analysis, or serially where the second has the additional context of the first reader's notes and opinions. If there is disagreement, a third consensus or arbitration reader is consulted. Typically, clinical applications will focus on the single reader with CAD, the simultaneous double-read, or the serial with knowledge double-read. Other options are reserved for teaching, research, or quality assurance.

McKinney and colleagues highlighted in their recent study that their AI could outperform clinicians in breast cancer mammography and reduce the workload for a second reader by 88% (10). In this study the team sought to involve an AI in the standard of care process for reading mammograms in order to test the system in a real-world application scenario. They stress-tested this algorithm based on two country's standards of care: the United Kingdom and the United States. In the UK, each mammogram is read using the serial double-reading method. In the US, each mammogram is read by a single reader. In both cases, the AI outperformed the first reader, and in the UK the system showed non-inferiority to the second reader. Importantly, the improvement on standard error was greater in the US where a second reader is not the norm.

McKinney and colleagues found that the UK standard of practice is the more efficient and safer of the two when it comes to integrating AI, and this was determined to be *because there is a second reader* (10). Perhaps more importantly, I argue, *is that the second reader is a human clinician*. McKinney and colleagues, however, take the reduction in second reader workload due to their AI to be indicative of a potential automation of this second role. They explain that the AI system could be used to provide automated and immediate feedback in screening and diagnostics while maintaining the two-reader standard-of-care. Essentially, the argument is to use the AI to double-check and confirm the first reader's diagnosis as if it were the second reader since there is statistical non-inferiority in its use.

This human-AI collaboration approach is not necessarily unique. Kempt and Nagel developed a short taxonomy of second opinions where they present the possibility that AI could be used to provide a second opinion in clinical contexts (11). They outline that there are three main types of second opinions: 1) patient-initiated quality control, 2) physician-initiated assessment, and 3) legal or institutionalized checks and balances. Kempt and Nagel suggest that AI could alleviate some of the cognitive labour required to perform these second opinions while maintaining the same epistemic diligence we currently require. Importantly, however, they highlight that without processes to ensure understanding through explanatory reasoning, disagreements between physicians and algorithms may not be resolved responsibly. For this reason, they develop their rule of disagreement, which requires that if a machine contradicts the initial diagnosis of a clinician, a second opinion from another human clinician is required.

Grote and Berens also present a scenario where a clinician acts as a first reader and the AI acts as second reader (12). In a partnership between philosophy and computational neuroscience, they highlight two studies in which pathologists were shown whole-slide images and were asked to classify the image. One group in each study was left on their own to review their decision, while the second group received an analysis from an algorithm. The results demonstrated significant improvements in one study, generally speaking, while the other one took a more fine-grained look at the results. This narrower examination found that experts benefitted only when they were originally uncertain, while novices benefitted regardless of their original confidence. Grote and Berens take this to indicate that when the machine provides an accurate read, there are significant benefits to using AI as a second reader. However, they also highlight an important limitation – the collaboration between human and AI breaks down when the algorithm gets things wrong. They say this is because the algorithm is checking the work of the clinician, which can make the expert unnecessarily question their work, while also leading novices to be overly reliant on the system's verdict.

The biggest concern that Grote and Berens raise, though, is that this collaboration results in clinicians feeling compelled to defer to the machine (12). They argue this is the case because of the prevalence of defensive medicine. Defensive medicine refers to clinical actions during patient care where the aim is to protect the *clinician*, rather than the patient – this occurs frequently in situations where there is a real risk of being sued or having one's licence questioned. Defensive medicine regarding the algorithm as second reader entails deferring to the machine to protect the clinician from accusations of malpractice or of deviating from protocols set by hospital administrators. Indeed, Grote and Berens raise the very real possibility that clinicians may even be pressured by the hospital to abide by algorithmic decisions to protect the hospital from accusations of malpractice, even when they are not certain of the validity of the algorithm's recommendations.

There are some significant problems with a collaborative model of human-AI diagnostics. However, I believe they only exist when the relationship is ordered human-first, AI-second. Instead, if the AI is required to go first and the clinician is the second reader, the relationship changes. No longer is the AI verifying the clinician's diagnosis – now the clinician takes on the role of the verificatory agent. This means that the clinician has the final say, and if they disagree with the system they will be able to explain their justification for disagreement by referencing their own work as well as referencing the AI's notes. This framing avoids setting novices up for failure (though, admittedly, it does not eliminate the possibility) and it avoids the charge of malpractice because the system is not set up as the check in our checks-and-balances. Instead, the human is checking the AI to make sure the AI did not make an error.

Strangely, the proposal for human-second, or the related human as primary decision-maker, has not made its way to healthcare-specific discussions at the time of writing this article. That does not mean the argument does not exist elsewhere, however. In fact, this already exists in more generalized or high-level discussions of AI – such as in the Montreal Declaration (13) – or even in some more specific discussions such as those regarding human-machine interactions and interfaces in AI-enabled warfare (14). Where these arguments exist, they are often part of a Responsibility Principle, such as "In all areas where a decision that affects a person's life, quality of life, or reputation must be made, where time and circumstance permit, the final decisions must be taken by a human being and that decision should be free and informed" (11). In other instances, the imperative exists as a response to cognitive biases in human-machine interactions. These biases can stem from the illusion of control, the desire for predictability, the reliance on heuristic shortcuts, and the moralization of amoral machines (14). Thus, the reprioritization I have outlined above for the healthcare field is in fact aligned with how discussions of AI systems have been developing in other fields.

An example of a system that is currently available that could be structured with a human-second approach is the RapidAI software when used in the context of ischemic stroke care. Slater and colleagues recently compared this AI with human readers for the detection of intracranial vessel occlusion and found that the software was able to improve workflow by alerting neurologists and radiologists to possible abnormalities in the scans (15). However, they also found that reliance on RapidAI, or even a proposed sole reliance on RapidAI, could result in substantial incorrect classifications of eligibility for therapy. In fact, Slater and colleagues found that human readers with variable amounts of experience performed better than the AI. Eight human readers with experience with advanced stroke ranging from several years to less than 12 months examined the same images as a single version of RapidAI, totaling 500 individual sets of images. This study pit the AI against the clinicians, rather than having a true first- then second-reader, so these results can only bring us so far. However, the system as designed is meant to provide decision *support* rather than decision *replacement*.

Despite this directive from other fields, and the warnings that come with software such as RapidAI, it appears that some already want this to go a step further in the healthcare sphere, as signaled by McKinney and colleagues, wherein the AI also becomes the first reader in single-reading jurisdictions. In other words, there is *only an AI reader with no human second reader*. This consideration arises when we question the goals of introducing advanced AI to the healthcare system in the first place (16). It may seem at first like the proposals for advanced AI in healthcare are meant to eventually move past this human-in-the-loop stage of integration. Instead, it has been argued that AI technologies ought to be used in ways that streamline tasks by handling those where no human intervention is required, thereby freeing up the clinician's time for more meaningful interaction with their patients (6,16). On this view, the collaborative nature of the human-AI relationship is that AI takes the burdens from the clinician and the clinician leaves tasks for the AI to handle alone. This would mean that my proposal for a clinician second reviewer fails to reach the end-goal for implementing AI in healthcare. Instead, on this view, my proposal might be better understood as an intermediary step towards autonomous AI decision-makers in healthcare diagnostics settings.

A potential counter that AI optimists may lodge against arguments that emphasize the need for responsibility considerations is that the criterion for a human to make the final decision is met by the patient, not the clinician administering the AI system. Since it is the patient who must consent to the testing, or the intervention, the counter argument might state that they are the more important human agent in the scenario. In other words, the clinician is potentially replaceable so long that the AI remains adherent to the legal requirements for consent. This would allow the system to move beyond a human-in-the-loop.

I believe that the various arguments from those investigated above, and especially this potential counter, could all benefit from more attention to the difference between "can" and "should". Many developers are interested in building stronger and stronger systems (16). The problem with this is that it entices other developers and policymakers, but it does not consider the needs of clinicians or patients. Providing safe and appropriate technology for healthcare involves balancing the desire to be more efficient and more powerful with the rights, needs, and safety of users and patients. Systems that do not need any human input would be incredibly dangerous – this could lead to simple mistakes going unchecked, or to the acceptance of systems without requisite fidelity (17). What I add to this existing work is the direct integration with clinical ethics and the patient's ability to consent. A black box system that offers no justification and does not interact with a human clinician could never be consented to by a patient because it would not provide any meaningful justification they could understand. An XAI system, on the other hand, would provide too much information for the average patient to know how or be able to interpret the recommendations, without help.

An additional problem with the assumptions given by technological optimists like those I've referenced above is that they are assuming the extra time will be allocated to clinicians in a way that allows for the building of relationships and interactions. Instead, we have to acknowledge that current stressors on the healthcare system, patient backlogs, billing structures, and many other factors all actively work against this optimistic possibility. So, instead of giving more time to build relationships, we may instead be setting up an unchecked system while further overworking our limited specialists.

The novelty of my argument for clinical discussions is that by keeping the human clinician involved as the *second reader*, we ensure that a human is always in-the-loop *and* that a human clinician is always signing off on a decision. So long as this second reader can understand the reasons for a specific diagnosis, the *Understanding Objection* fails. In this situation, the clinician can either provide the necessary explanation for the patient to understand the relevance of the system, regardless of whether the system is itself a black box, grey box, or an XAI system, or they can directly refute the system because they are able to understand where the system made its mistakes. Notice that the human clinician does not render the AI system redundant, either. This human second reader model still allows for the use of complex diagnostic AI systems that can outperform human clinicians. Thus, medical AI systems in radiology may be justified on the grounds that they can diagnose based on information the human could not have seen, may have missed, or otherwise would not have registered for whatever reason. Moreover, the radiologists' workload may be reduced and tests processed more quickly, since one human reader is removed from the loop and the AI system is faster than a human radiologist.

BENEFITS OF THE SECOND READER

A benefit of the mandatory second reader, human-in-the-loop approach that I propose is that it allows the AI its own autonomy without sacrificing the clinician's or the patient's autonomy. The AI can be autonomous and still be subject to supervision. What I mean by this is that the system could be a black box and still protect consent so long as the clinician follows various obligations to ensure safety of the system – obligations such as questioning the recommendations even if they agree with them, for example. For XAI systems the human second reader removes the concerns over too much information or being too unwieldly because they are also evaluating the images. Second readers examine the image with the first reader's notes as a guide, rather than acting as an evaluator of the first reader's notes. So long as clinicians ensure the information presented to the patient is not overwhelming, consent can be maintained here as well. Even for systems that are not as easily classified as "either-or," the second reader would likely have an easier time performing their analysis because the system has already provided some justification for its recommendation without burying it under a mass of irrelevant information. So long as the clinician then communicates to the patient in a way they can understand, the ability to consent is preserved. This is how we currently handle consent for other technologies such as complex blood tests or genetic assays.

Additionally, this clinician second reader model provides a unique benefit to both AI readers and clinician readers. On the one hand, human expert clinicians are good at understanding the nuances of clinical images as they exist within a specific

context (18), while AI are good at quickly synthesizing information at levels humans cannot process unassisted (16). On the other hand, human radiologists make errors because they under-read the scans, are imperfect reasoners or may focus on only some salient aspects on an image and overlook others (19), while AI systems can make errors due to misclassifying an image, bugs in the code, or even miniscule and imperceptible modifications to the image (20). From a data ethics and privacy lens, Alvarado explains the main problem with AI systems in diagnostics is how easy it is for them to fail when confronted with variations and minor, random deviations from what they expect to see – and how catastrophic it can be when this happens (21). From a user's lens, Lee and colleagues' team of psychiatrists and cognitive software developers summarize the problems with AI making decisions alone quite well when they explain "it does not have the ability to make compassionate, fair, and equitable decisions. AI cannot self-regulate or self-correct or consider diversity among people and perspectives, ethics, and morality" (6). On their own, both have significant problems to overcome. However, I believe that the clinician second reader approach to AI diagnostics allows for a collaboration between human and machine that addresses each reader's shortcomings.

One final benefit of bringing the human-as-final-authority proposal into the clinical sphere is that it succinctly answers the concerns that AI will make radiologists and other specialists obsolete. One particularly strong phrasing of this comes from Hinton, a cognitive psychologist and computer scientist known for his work on artificial neural networks. At the 2016 Machine Learning and Market for Intelligence Conference he made the following claim during a recorded Q&A:

> Let me start by just saying a few things that seem obvious. I think if you work as a radiologist you're like the coyote that's already over the cliff but hasn't yet looked down so it doesn't realize there's no ground underneath him. People should stop training radiologists now. It's just completely obvious that within five years deep learning is going to do better than radiologists because this can be able to get a lot more experience. It might be ten years. But we've got plenty of radiologists already. (22)

With due respect to Hinton, and others who make similar claims, I think it is unlikely that medical AI will lead to a future without human radiologists. Such claims have not been borne out for previous technologies that were meant to replace whole functions or professions, such as the MRI replacing anatomical reporting. Moreover, the claim that radiologists will no longer have a role in radiology seems informed by abstract fancy rather than practical application. As I have discussed above, systems will need a human-in-the-loop to address practical concerns such as whether a patient can consent to a treatment. Without careful attention to informed consent in AI we risk setting up hospitals to revert to the rejected paternalistic ways of the past.

Reçu/Received: 27/11/2023
Publié/Published: 2/12/2024

Remerciements
Cette recherche a été financée en partie par la bourse de doctorat 752-2020-2225 du Conseil de recherches en sciences humaines du Canada. L'auteur souhaite également remercier Daniel Steel pour avoir revu et commenté un projet antérieur qui a donné lieu à cette analyse.

Acknowledgements
This research was funded in part by a Social Sciences and Humanities Research Council of Canada Doctoral Fellowship 752-2020-2225. The author would also like to acknowledge Daniel Steel for reviewing and commenting on an earlier project that evolved into this analysis.

Conflits d'intérêts
Aucun à déclarer

Conflicts of Interest
None to declare

Édition/Editors: 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 & Nathaniel Bendahan

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. Beauchamp TL, Childress JF. Principles of Biomedical Ethics 7th ed. Oxford University Press; 2013.
2. Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue. 2018;16(3):31-57.
3. London AJ. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. Hastings Center Report. 2019;49(1):15-21.

4. Schiff D, Borenstein J. How should clinicians communicate with patients about the roles of artificially intelligent team members? AMA Journal of Ethics. 2019;21(2):138-45.
5. Watson DS, Krutzinna J, Bruce IN, et al. Clinical applications of machine learning algorithms: Beyond the black box. BMJ. 2019;364:l886.
6. Lee EE, Torous J, De Choudhury M, et al. Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging. 2021;6(9):856-64.
7. Wadden JJ. What kind of artificial intelligence should we want for use in healthcare decision-making applications? Canadian Journal of Bioethics/Revue Canadienne de Bioéthique. 2021;4(1):94-100.
8. Wadden JJ. Defining the undefinable: the black box problem in healthcare artificial intelligence. Journal of Medical Ethics. 2021;48(10):764-68
9. Geijer H, Geijer M. Added value of double reading in diagnostic radiology, a systematic review. Insights into Imaging. 2018;9(3):287-301.
10. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.
11. Kempt H, Nagel SK. Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts. Journal of Medical Ethics. 2022;48(4):222-29.
12. Grote T, Berens P. How competitors become collaborators—Bridging the gap(s) between machine learning algorithms and clinicians. Bioethics. 2022;36(2):134-42.
13. Declaration Development Committee. Montreal Declaration on Responsible AI. Montreal, QC: Universite de Montreal.
14. Johnson J. The AI commander problem: Ethical, political, and psychological dilemmas of human-machine interactions in AI-enabled warfare. Journal of Military Ethics. 2023;21(3-4):246-71.
15. Slater L-A, Ravintharan N, Goergen S, et al. *RapidAI* compared with human readers of acute stroke imaging for detection of intracranial vessel occlusion. Stroke: Vascular and Interventional Neurology 2024;4(2):e001145.
16. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Toronto, Basic Books; 2019.
17. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1(5):206-15.
18. Lesgold A, Rubinson H, Feltovich PJ, Glaser R, Klopfer D, Wang Y. Expertise in a complex skill: Diagnosing x-ray pictures. In: Chi MTH, Glaser R, Farr MJ, editors. The Nature of Expertise. Mahwah, NJ: Lawrence Erlbaum Associates; 1988. p. 311-42.
19. Brady, AP. Error and discrepancy in radiology: Inevitable or avoidable? Insights into Imaging. 2017;8(1):171-82.
20. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. International Conference on Learning Representations 2014 Proceedings. arXiv:1312.6199
21. Alvarado R. Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI. Bioethics. 2022;36(2):121-33.
22. Creative Destruction Lab. Geoff Hinton: On Radiology. Filmed at the 2016 Machine Learning and Market for Intelligence Conference, Toronto: Ontario; November 2016.