



Perspective

The Ethics of Artificial Intelligence in Clinical Diagnostic Decision-Making

Laura Zielinska^{1*}, Bruno Solberg² and Johanna Colombo³

¹Center for Medical Ethics and Law, Delta University Center for Bioethics, Netherlands

²Department of Bioethics, Northmoor School of Public Health Ethics, United Kingdom

³Department of Bioethics, Northmoor School of Public Health Ethics, United Kingdom

ABSTRACT

Artificial intelligence is increasingly embedded in the diagnostic process, from autonomous screening tools to decision-support systems that flag findings for clinician review, raising ethical questions that go well beyond whether these systems are accurate enough to trust.

This perspective examines how the ethical stakes of diagnostic AI vary with its deployment model, and surveys recent literature on algorithmic bias and the limits of treating accuracy and fairness as interchangeable goals.

It then turns to the black-box problem and the unsettled question of who is accountable when an opaque but well-validated system contributes to a diagnostic error, and to informed consent in an AI-mediated diagnostic encounter that patients are rarely told about.

We also consider evidence on automation bias and the deskilling of clinicians who come to rely on diagnostic AI, and the regulatory apparatus -- centered on the FDA's evolving framework for adaptive software as a medical device -- that has been slow to keep pace with either concern.

We conclude that no single deployment model or regulatory fix resolves these tensions, and that the ethical questions AI raises for diagnosis are best addressed deployment model by deployment model rather than as a single undifferentiated debate about "AI in medicine."

Keywords: artificial intelligence; clinical diagnosis; algorithmic bias; informed consent; medical ethics

Citation: Zielinska L, Solberg B and Colombo J (2023) The Ethics of Artificial Intelligence in Clinical Diagnostic Decision-Making. *Bioethics*. 1(1):01-05.

Corresponding Author: Laura Zielinska

E-mail Address: l.zielinska@ducbl.nl

Article History: Received 11 December 2022 | Accepted 16 March 2023 | Published 22 June 2023

Copyright: © 2023 Laura Zielinska, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (<https://creativecommons.org/licenses/by/4.0/>).

INTRODUCTION

A radiologist reviewing a chest CT, a dermatologist assessing a suspicious lesion, and an emergency

physician triaging a patient with chest pain are all, increasingly, doing that work alongside a machine-learning system trained to recognize patterns in medical images or clinical data that no single

clinician could have seen enough cases to learn directly. The clinical promise is real: several trials of AI-based decision support have reported measurable gains in diagnostic accuracy when the tool is used well [1]. But diagnosis in medicine has never been a purely technical act -- it is the point at which a patient's symptoms are translated into a category that determines what happens to them next, and that translation carries ethical weight independent of how accurate it turns out to be [2].

Introducing an algorithm into that translation does not just add a new tool to the clinician's kit; it changes who or what is doing the interpreting, how much of the reasoning behind a conclusion can be inspected, and who is answerable when the conclusion is wrong. A recent systematic review of the ethics literature found that arguments both for and against AI-assisted clinical decision-making tend to cluster around the same handful of concerns -- accuracy, transparency, accountability, and the preservation of clinical judgment -- even when authors disagree sharply about how those concerns should be weighed [3]. This review takes up those concerns as they apply specifically to diagnosis: the moment in care where AI's pattern-matching strengths are most obviously useful and its ethical costs are least obviously priced in.

We proceed in five parts. We first sketch what diagnostic AI actually does and where it sits in the clinical workflow. We then examine bias and the limits of "accuracy" as an ethical benchmark, the black-box problem and what it does to explainability and accountability, informed consent in an AI-mediated diagnostic encounter, and finally deskilling, automation bias, and the regulatory apparatus meant to govern all of the above.

AI IN THE DIAGNOSTIC ENCOUNTER: PROMISE AND SCOPE

Diagnostic AI systems in current use range widely in both technical design and clinical role. Some function as autonomous readers -- for instance, screening tools cleared to interpret a retinal image for diabetic retinopathy without a clinician reviewing every case -- while most function as a "second reader" or clinical decision support system that offers a probability estimate or flags a finding for a physician who retains final authority [4]. This distinction matters ethically far more than it is usually treated as mattering in public discussion of "AI diagnosis," because the ethical questions that follow -- about oversight, liability, and consent -- depend heavily on how much independent judgment the system is actually exercising.

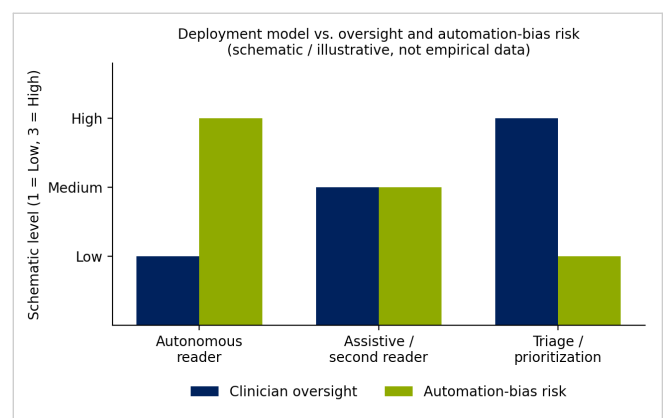
Table 1 lays out three broad deployment models along these lines and the ethical considerations that track each one. What the table is meant to show is that "AI in diagnosis" is not one thing ethically; a triage algorithm that flags a scan for priority radiologist review raises a

different set of concerns than a system whose output is treated as the diagnosis itself. Much of the ethics literature blurs this distinction, discussing "AI diagnosis" as a single category when the degree of clinician oversight built into a given deployment is doing most of the ethical work [5]. Figure 1 makes the same point schematically, plotting three illustrative deployment models against two dimensions -- how much clinician oversight the model assumes and how much risk of automation bias it correspondingly carries -- and showing that the two tend to move in opposite directions: the models that reduce clinician oversight also tend to be the ones where uncritical reliance on the algorithm's output is easiest to fall into.

Table 1. Illustrative comparison of three diagnostic AI deployment models and the ethical considerations that track each one.

Deployment model	Clinician oversight	Typical example	Primary ethical concern
Autonomous reader	Minimal; output may stand as the result without independent review	Retinal screening for diabetic retinopathy	Accountability for error falls almost entirely on the system and its regulatory clearance
Assistive / second reader	Clinician reviews and can override the recommendation	Radiology decision-support flagging a suspicious finding	Automation bias -- deferring to the flag even when independent judgment points elsewhere
Triage / prioritization	Clinician retains full diagnostic authority; AI only reorders workflow	Algorithmic prioritization of scans for urgent read	Lower per-decision stakes, but errors can still delay care for deprioritized patients

Figure 1. Schematic comparison of clinician oversight and automation-bias risk across three diagnostic AI deployment models (illustrative, not derived from a single empirical study).



BIAS, FAIRNESS, AND DIAGNOSTIC ACCURACY

Much of the early ethical concern about diagnostic AI centered on algorithmic bias -- the risk that a model trained predominantly on data from one demographic group performs worse, sometimes substantially worse, on patients outside that group. This is not a hypothetical risk; documented cases of diagnostic and risk-prediction algorithms performing unevenly across race, sex, and skin tone have been widely discussed as a call for more open, representative training data and more rigorous subgroup validation before deployment [6]. Health-equity-focused reviews have argued that unless bias auditing becomes a standard, mandatory part of the validation pipeline rather than an afterthought, AI risks quietly reproducing and scaling the same disparities that already exist in under-resourced diagnostic pathways [7].

More recently, some commentators have pushed back against framing this purely as a "fairness" problem to be solved by rebalancing datasets, arguing that the more fundamental ethical obligation is simply accuracy for every patient a system is used on -- and that a narrow focus on group-level fairness metrics can obscure cases where a system is inaccurate for everyone, or where fairness corrections trade accuracy in one group for accuracy in another without addressing the underlying data problem [4]. This is a genuine disagreement in the literature, not a settled point: whether "bias" in diagnostic AI is best understood as a distinct ethical wrong requiring its own remedy, or as one symptom of an accuracy problem that a sufficiently rigorous validation standard would catch regardless of demographic framing, remains open. What is not in serious dispute is that accuracy claims made in a vendor's validation study do not automatically transfer to a different hospital's patient population, equipment, or diagnostic conventions, which is one reason several reviews call for local validation before a system is trusted at the bedside [8].

EXPLAINABILITY, TRUST, AND ACCOUNTABILITY FOR ERROR

Many of the diagnostic models that perform best are also the least interpretable -- deep neural networks whose internal reasoning cannot be stated in terms a clinician can independently evaluate, only inspected after the fact through approximate techniques that highlight which parts of an image or record most influenced the output. This is the "black box" problem, and it has generated a genuine split in the bioethics and health-informatics literature. One position holds that explainability is a practical and legal necessity: a clinician cannot exercise independent professional judgment over a recommendation whose basis is opaque, and a patient cannot meaningfully consent to a decision no one can explain to them [9]. A more permissive position argues that black-box AI can be ethically justified in some circumstances provided the

system has been rigorously and transparently validated on relevant populations, even if the internal mechanism defies interpretation -- on the grounds that clinicians already rely on plenty of well-validated tests and rules of thumb they cannot fully explain mechanistically [10]. Framed around the principle of nonmaleficence, other authors have argued that the ethical weight of the black-box problem should scale with the stakes of the decision, tolerable for a low-risk triage flag but much harder to justify for a diagnosis that directly determines treatment [11].

The explainability debate feeds directly into a harder question: who is accountable when an AI-assisted diagnosis is wrong. Qualitative work with clinicians in the UK found considerable uncertainty among practicing physicians about where their own liability begins and ends when a diagnostic tool contributes to an error -- whether following an incorrect AI recommendation constitutes reasonable reliance on a validated clinical tool or a failure of independent judgment [8]. Legal and ethics scholarship examining trust and trustworthiness in clinical decision support has argued that existing liability frameworks, built around an individual clinician's standard of care, do not map cleanly onto a diagnostic process that is now distributed across a clinician, a software vendor, a health system's implementation choices, and the training data a model was built on [12]. Some commentators frame this bluntly as a question of who "owns the mistake" when a system contributes to harm, and note that current malpractice doctrine offers no settled answer [13]. Building genuine trust in these systems, several recent reviews argue, requires treating trustworthiness as an engineering and governance property that has to be designed in -- through validation, monitoring, and clear lines of responsibility -- rather than something that follows automatically from a system performing well in a validation study [14].

INFORMED CONSENT AND PATIENT AUTONOMY

Patients undergoing a diagnostic workup are rarely told, in any explicit way, that an algorithm contributed to their radiologist's reading or their triage priority. Scholarship on informed consent in AI-assisted care has begun arguing that this omission is ethically significant and not merely a matter of administrative convenience: if an algorithm materially shapes a diagnostic conclusion, disclosure of that fact plausibly belongs in the same category of information a reasonable patient would want before consenting to a workup, alongside who is reading their scan and what tests are being run [15]. The traditional one-time consent model, obtained at intake and rarely revisited, sits awkwardly with diagnostic AI systems that are continuously updated, retrained on new data, or swapped out by a health system without the patient ever being informed that the tool assisting their

diagnosis has changed [15].

Legal scholarship on the intersection of AI and consent doctrine has noted that most jurisdictions' informed consent law developed around discrete interventions -- a surgery, a specific test -- and offers little clear guidance on whether a hospital's routine use of a background diagnostic-support algorithm needs to be disclosed at all, as opposed to a novel intervention a patient might reasonably decline [16]. The stakes are not symmetric across specialties: work examining AI trained on especially sensitive data, such as mental health records, has argued that consent obligations should scale with sensitivity, and that models trained on patient data for one diagnostic purpose should not be silently repurposed for another without renewed authorization [17]. Underlying all of this is a harder question that mirrors the debate over decision-making capacity assessments more broadly: if a patient cannot meaningfully evaluate what an opaque algorithm is doing with their case, in what sense is their consent to its use actually informed, as opposed to merely obtained [18]?

DESKILLING, AUTOMATION BIAS, AND REGULATORY OVERSIGHT

A separate strand of concern has less to do with any single diagnostic error and more to do with what routine reliance on AI does to clinical skill over time. A scoping review of the deskilling literature found a consistent, if still largely observational, signal that clinicians who rely heavily on AI decision support for a diagnostic task show measurably reduced independent performance on that same task when the tool is removed, particularly among trainees who never fully develop the unassisted skill in the first place [19]. This connects to a well-documented phenomenon in human factors research called automation bias -- the tendency to defer to an automated recommendation even when a clinician's own judgment, or contradictory evidence in front of them, points the other way. Recent work specifically testing clinicians exposed to erroneous AI-generated advice found that the erroneous suggestion measurably degraded diagnostic reasoning relative to clinicians working without the AI input at all, a finding with obvious implications for any diagnostic workflow that assumes clinicians will simply "double check" an AI output.

Regulatory oversight has been slow to catch up with either concern. The FDA's traditional device-approval framework was built for products that do not change after clearance, which fits poorly with AI systems designed to keep learning from new data; the agency's more recent guidance for AI/ML-based software as a medical device attempts to address this through a predetermined change control plan that lets a manufacturer specify in advance how a model may be modified post-market without triggering a full new review [20]. Legal and health-policy analyses of this framework have generally welcomed the flexibility it

offers but note that it still leaves open how post-market monitoring for performance drift or emergent bias will actually be enforced in practice, as opposed to described in a submission document [21]. Until that enforcement question is answered, much of the burden for catching a diagnostic AI system that degrades, drifts, or performs unevenly across patient populations falls back on the same individual clinicians and health systems that the technology was partly meant to support.

CONCLUSION

The ethical questions surrounding AI in diagnosis do not reduce to a single dispute about whether the technology is accurate enough to trust. They span several distinct and only loosely related concerns: whether "fairness" and "accuracy" are the same target or competing ones, whether a system opaque enough to defy explanation can still be used responsibly, who bears liability when a diagnostic error involves a human-machine collaboration rather than a single clinician's judgment, whether patients are owed disclosure and a meaningful choice about a technology most of them do not know is involved in their care, and what happens to clinical skill and judgment when a tool designed to assist becomes a tool that is quietly relied upon. None of these questions has been resolved by the literature surveyed here, and several -- particularly the proper scope of disclosure to patients and the allocation of liability for AI-assisted error -- remain genuinely contested rather than merely under-discussed. What does seem reasonably clear is that treating diagnostic accuracy as the only metric that matters understates the ethical stakes considerably, and that the deployment model chosen for a given AI system -- autonomous, assistive, or triage -- determines much of which of these concerns is most pressing in a given clinical setting.

DECLARATIONS

Acknowledgements

The authors are grateful to colleagues in their respective bioethics departments for helpful conversations that sharpened the scope and argument of this review.

Funding

This review received no external funding.

Conflict of Interest

The authors declare no conflicts of interest relevant to this work.

Author Contributions

Laura Zielinska conceived the review, conducted the literature synthesis on diagnostic accuracy, bias, and regulatory oversight, and drafted the initial manuscript. Bruno Solberg contributed the sections on explainability, accountability, and liability and critically

revised the manuscript. Johanna Colombo contributed the analysis of informed consent, automation bias, and deskillling, and provided final approval of the submitted version. All authors read and approved the final manuscript.

Ethics Statement

As a literature review synthesizing previously published, publicly available research and policy analysis, this article did not involve new data collection from human participants and did not require institutional ethics committee approval.

Data Availability

No new data were generated or analyzed for this review. All sources discussed are cited in the reference list and are publicly available through their respective publishers.

REFERENCES

1. Impact of AI-Based Clinical Decision Support Systems on Diagnostic Accuracy Among Healthcare Professionals: A Systematic Review and Meta-Analysis of Randomized Controlled Trials (2026) Applied Sciences (MDPI).
2. Artificial Intelligence in Clinical Decision Making: Opportunities and Ethical Challenges: Clinical Applications and Ethical Considerations (2024) Developmental Medico-Life-Sciences.
3. Should Artificial Intelligence be used to support clinical ethical decision-making? A systematic review of reasons (2023) BMC Medical Ethics, DOI:[10.1186/s12910-023-00929-6](https://doi.org/10.1186/s12910-023-00929-6).
4. Ethical Use of Artificial Intelligence in Medical Diagnostics Demands a Focus on Accuracy, Not Fairness (2024) NEJM AI, DOI:[10.1056/AIp2400672](https://doi.org/10.1056/AIp2400672).
5. Ethical Considerations for AI in Clinical Decision-Making (2024) AJMC.
6. Addressing bias in big data and AI for health care: A call for open science (2021) PMC8515002.
7. Health Equity and Ethical Considerations in Using Artificial Intelligence in Public Health and Medicine (2024) Preventing Chronic Disease.
8. Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: a qualitative study of healthcare professionals' perspectives in the UK, PMC12235780.
9. On the practical, ethical, and legal necessity of clinical Artificial Intelligence explainability: an examination of key arguments (2025) BMC Medical Informatics and Decision Making.
10. Prince S, Savulescu J (2025) When is black-box AI justifiable to use in healthcare? DOI:[10.1177/20539517251386037](https://doi.org/10.1177/20539517251386037).
11. Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm" (2023) [journal article, ScienceDirect].
12. Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability (2023) Medical Law Review, 31(4):501.
13. The ethical challenge of AI in medicine: Who owns the mistakes? PMC13168724.
14. Ethics of trustworthy AI in healthcare: Challenges, principles, and practical pathways (2025) [journal article, ScienceDirect].
15. Informed Consent in the Age of Clinical AI: Beyond the One-Time Authorization Model (2026) DOI:[10.1080/15265161.2026.2675917](https://doi.org/10.1080/15265161.2026.2675917).
16. Artificial intelligence and the law of informed consent, in Research Handbook on Health, AI and the Law, NCBI Bookshelf.
17. Reclaiming informed consent to train mental health AI with patient data, PMC13230920.
18. Ethical considerations for the use of artificial intelligence in medical decision-making capacity assessments (2023) PMID:37717548.
19. Artificial intelligence in medicine: a scoping review of the risk of deskillling and loss of expertise among physicians (2026) PMID:41890350.
20. Artificial Intelligence in Software as a Medical Device (2024) U.S. Food and Drug Administration.
21. United States Food and Drug Administration Regulation of Clinical Software in the Era of Artificial Intelligence and Machine Learning, PMC12264609.