2026, Volume 6, Issue 1, Page No: 144-152
ISSN: 3108-5059

Society of Medical Education & Research

Asian Journal of Ethics in Health and Medicine

The Right to a Human Reason: Contestability in High-Stakes Clinical Decisions

Rafael Mendoza[1]*, Isabela Cruz[1], Carlos Lopez[2]

[1]Department of Clinical Contestability and Human Reasoning, Faculty of Medicine, National Autonomous University of Mexico, Mexico City, Mexico.
[2]Department of High-Stakes Decision Ethics, Faculty of Medicine, Monterrey Institute of Technology, Monterrey, Mexico.

***E-mail** ✉ rafael.mendoza@unam.mx

Abstract

High-stakes clinical decisions are increasingly shaped by algorithmic systems that may inform diagnosis, treatment, triage, discharge, risk classification, and access to scarce interventions. Ethical discussion has often emphasized notice, disclosure, transparency, and explanation. These protections are important but incomplete when an affected patient cannot ask why a consequential decision was made, obtain reconsideration by an empowered clinician, or secure revision when the decision cannot be adequately defended. This article develops a rights-based normative argument for a proposed right to a human reason. The right is not presented as an existing universal legal entitlement, nor does it presume that human judgment is inherently more accurate than artificial intelligence. Rather, it protects the patient's standing as someone to whom consequential clinical decisions should be answerable. A human reason is defined as an intelligible, patient-directed justification that an appropriately competent and authorized human agent can assess, own, defend, revise, or withdraw. The article distinguishes explanation from justification and justification from contestability; proposes proportional triggers for the right; and specifies minimum procedural safeguards from notice through appeal. It also develops a risk-sensitive model for high-stakes decisions while recognizing emergency, resource, expertise, equity, and institutional constraints. The proposal remains normative and requires empirical testing of accessibility, burden, delay, error correction, reviewer independence, equity, and unintended effects before any implementation claim would be warranted.

Keywords: Artificial intelligence, Clinical ethics, Contestability, Patient autonomy, Human review, Justification

Introduction

Explanation is not yet contestability

Healthcare artificial intelligence has intensified demands for transparency, notice, and explanation. Those demands respond to a genuine ethical problem: patients may be affected by systems whose role is not visible to them and whose outputs may influence consequential decisions. Recent rights-based scholarship therefore argues that patients can have ethically significant claims to notice and explanation when AI is used in care [1]. Yet receiving information about a system, or even an account of how it contributed to a decision, does not by itself give a patient standing to challenge that decision. Explanation can inform without opening a route to reconsideration. A patient may understand that an algorithm influenced a denial, classification, or recommendation while remaining unable to ask which reasons were decisive for her case or what could justify changing the result.

This distinction matters because "a human in the loop" can describe nominal presence rather than meaningful oversight. Effective oversight depends on substantive human capacities and cognitive conditions, including the ability to understand the decision context, resist uncritical reliance, and intervene when warranted [2, 3]. A clinician who merely transmits an AI-supported conclusion without independent authority or practical space to reconsider it may be present yet functionally unable to contest it. The ethical deficit is therefore not only opacity.

Access this article online https://smerpub.com/

Received: 19 February 2026; **Accepted:** 30 May 2026

How to cite this article: Mendoza R, Cruz I, Lopez C. The Right to a Human Reason: Contestability in High-Stakes Clinical Decisions. Asian J Ethics Health Med. 2026;6(1):144-52. https://doi.org/10.51847/iLMwibqSTd

It is a deficit of answerability. Conversely, removing AI would not by itself solve the problem: unaided human decisions can also be opaque, paternalistic, or poorly reviewable. The proposed right is consequently directed at high-stakes clinical authority, not at AI as such.

Empirical-ethical work with prospective health professionals further suggests that responsibility is connected to decision-making authority, professional experience, and the conditions under which judgment can actually be exercised [4]. That finding does not establish a universal model of clinician responsibility, but it supports the narrower proposition that responsibility cannot be reduced to formal human presence. This article therefore proposes a right to a human reason for high-stakes clinical decisions. The right is not a claim that humans are always epistemically superior to machines, nor that every algorithmic contribution requires elaborate review. It is a claim that when a decision materially affects a patient's interests, the patient should, under proportionate conditions, be able to obtain a humanly attributable justification and a meaningful route to challenge it.

What constitutes a human reason?

A human reason is not simply an explanation delivered by a person. Explanation concerns intelligibility: what factors, rules, patterns, or model outputs contributed to a conclusion. Justification concerns whether the decision can be defended to the person affected by it. Recent normative analysis argues that healthcare AI raises a need for justification to the patient that cannot be satisfied merely by making model behavior more interpretable [5]. The proposed right begins from that distinction. It also rejects a purely ceremonial notion of human authorship. A clinician does not convert an algorithmic output into a human reason merely by reading it aloud, signing it, or placing it in the record.

Nor should an AI output automatically be treated as equivalent to human epistemic authority. Philosophical analysis of medical AI emphasizes differences between machine-generated inference and the practices through which medical authority is ordinarily claimed, exercised, and socially recognized [6]. This does not imply that a clinician is always more accurate. It means that predictive performance alone does not establish who can answer for the decision, respond to reasons offered by the patient, or revise the judgment when competing considerations become salient. Human attribution is thus a responsibility condition, not a guarantee of epistemic superiority.

Opacity sharpens the problem. When a physician cannot adequately relate a non-interpretable output to a deliberative account of why a particular action is warranted, the physician's justificatory role may be weakened [7]. The human reason proposed here is therefore an intelligible, patient-directed account that an appropriately competent and authorized human agent can assess, own, defend, revise, or withdraw. "Humanly attributable" refers to responsibility for the reason, not to unaided cognition. A clinician may rely on AI, laboratory evidence, colleagues, or guidelines, but must remain capable of giving a reason that can enter a genuine exchange rather than merely repeating a system output.

Table 1 presents "Distinguishing explanation, human reason, justification, and contestability in clinical decision-making" for "The Right to a Human Reason: Contestability in High-Stakes Clinical Decisions."

Table 1. Distinguishing explanation, human reason, justification, and contestability in clinical decision-making

Concept	Primary function	Relational or dialogical role	Corrective capacity
Explanation	Informational: makes relevant inputs, processes, or influences intelligible	May support understanding but need not invite dialogue	None by itself
Human reason	Relational: gives a patient a humanly attributable account of why the decision is defensible	Requires an answerable human agent able to engage with patient reasons	Can support reconsideration
Justification	Justificatory: defends why a decision should be accepted as warranted	Addresses the person affected, not only the technical mechanism	May expose weaknesses in the decision
Contestability	Procedural and dialogical: permits questioning, review, response, and possible correction	Gives the patient a recognized route to challenge	Includes revision or escalation where warranted

Normative grounds for the right to a human reason

The proposed right is grounded first in relational autonomy. Clinical autonomy is not exhausted by isolated choice; it is exercised within relationships in which information, professional judgment, trust, vulnerability, and institutional authority shape what

choices are realistically available. AI adds another actor-like element to that relation. Analysis of AI and shared decision-making therefore describes a triadic structure in which technology can alter communication and the allocation of epistemic authority without eliminating the responsibilities of clinicians toward patients [8]. A right to a human reason gives that relational responsibility procedural expression: when authority is exercised consequentially, the affected patient remains someone to whom an account is owed.

A second ground is resistance to algorithmic paternalism. Patients possess experiential and evaluative knowledge about their goals, tolerances, symptoms, identities, and acceptable trade-offs. Normative work on patient wisdom argues that excluding such knowledge from health AI risks reproducing paternalistic decision-making even when the system is technically sophisticated [9]. A right to a human reason protects the patient's standing as a reason-giver: someone whose challenge can introduce considerations that a model, dataset, or initial clinician judgment may not adequately represent. This standing does not depend on the patient demonstrating that the original decision was wrong before being heard.

The argument is not anti-AI. AI may, under some conditions, expand information, identify options, or support more informed choices, and normative scholarship has accordingly argued that it can sometimes enhance autonomy [10]. The relevant opposition is therefore not human versus machine. It is answerable versus unanswerable decision-making. The proposed right is justified when consequential clinical authority is exercised in a way that affects the patient but cannot be meaningfully questioned. Its moral purpose is to secure relational answerability, preserve evaluative agency, and connect professional authority with responsibility. It does not grant an automatic veto over evidence-based care, nor does it require clinicians to endorse every patient preference. It requires that significant disagreement be met with reasons rather than procedural closure.

When the right is triggered

A right this demanding should not attach identically to every computational contribution. A dosage calculator, background image-processing tool, triage model, autonomous diagnostic system, and AI-supported treatment recommendation do not exercise the same kind of influence. A recent ethical framework distinguishes assistive, replicative, and more autonomous roles for AI in healthcare, supporting the proposition that ethical scrutiny should vary with the system's functional and decisional role [11]. The article therefore proposes a proportional trigger rather than a universal maximum procedure. The relevant question is not simply whether AI was used, but whether the decision-making arrangement created a material need for human answerability.

The trigger should strengthen as several features accumulate: the consequence for the patient becomes more serious; the system exercises greater influence over the decision; the rationale is less intelligible; the decision is value-sensitive or preference-dependent; disagreement could materially alter the outcome; the patient is especially vulnerable to exclusion or error; or the decision is difficult to reverse. Existing normative work proposing a right to refuse AI-based diagnosis or treatment planning provides a useful precedent for stronger patient control over some forms of AI-mediated care [12]. That argument, however, should not be misdescribed as a universal legal entitlement. Nor does the proposed right imply that refusal must always result in access to an equally effective non-AI alternative.

Disclosure itself also requires proportionality. A recent ethical critique rejects the assumption that clinicians are always obligated to disclose every use of medical machine learning, showing that the strength of disclosure arguments varies with context and materiality [13]. The proposed right therefore does not require maximal notice whenever software is involved. It is triggered when an AI-supported or otherwise complex decision materially shapes a high-stakes clinical outcome and ordinary communication would not provide a realistic opportunity for challenge. The more consequential, opaque, autonomous, value-laden, or difficult to reverse the decision, the stronger the case for a fuller contestability process. This is a proposed trigger test, not a validated threshold or legal rule.

Requirements of meaningful contestability

Contestability begins where consent and explanation stop. Consenting to AI use and contesting a decision serve different protective functions; the former concerns permission or acceptance, whereas the latter concerns the ability to challenge how authority was exercised [14]. A meaningful right must therefore include more than disclosure at the front end. At minimum, the patient must know that a consequential decision has been made, receive a humanly attributable reason, and have a genuine opportunity to question that reason. Questioning

must also be practically usable: a nominal invitation that carries no channel, no response obligation, or no protection against dismissal does not constitute meaningful contestability.

Human review is the crucial threshold. Ethical and epistemological analyses of clinician–AI disagreement and second opinions show why review must permit independent assessment rather than ritual confirmation [15, 16]. A reviewer may conclude that the original decision remains correct, including when the AI-supported recommendation is stronger than the challenge. Meaningful contestability does not require reversal; it requires that reaffirmation be reasoned and that revision remain genuinely possible. The reviewer therefore needs competence relevant to the decision, sufficient information, practical time to reconsider it, and authority to depart from the initial determination or escalate the case.

The article therefore proposes a minimum sequence: notice, intelligible humanly attributable reason, opportunity to question, access to empowered human review, substantive response, revision where warranted, documentation, and appeal or other recourse. These elements are proposed as a coherent procedural package rather than a validated standard. The precise form should vary by stakes, urgency, institutional capacity, and patient need. What must remain invariant is the difference between a route that can potentially change the decision and one that merely records dissatisfaction. Contestability is consequently both dialogical and corrective: it creates a place for reasons to be exchanged and a mechanism through which those reasons can matter.

Table 2 presents "Minimum procedural safeguards for contestable high-stakes clinical decisions" for "The Right to a Human Reason: Contestability in High-Stakes Clinical Decisions."

Table 2. Minimum procedural safeguards for contestable high-stakes clinical decisions

Safeguard	Minimum function	Status
Notice	Identifies the consequential decision and material AI involvement where relevant	Evidence-supported; scope proportionate
Intelligible humanly attributable reason	States why the decision is defensible in terms the patient can engage	Proposed synthesis informed by justification literature
Opportunity to question	Allows the patient or representative to raise reasons, facts, preferences, or errors	Normatively supported
Human review	Provides an appropriately competent reviewer with authority and practical capacity to reconsider	Evidence-supported conditions; institutional form proposed
Substantive response	Addresses the challenge rather than merely acknowledging receipt	Proposed safeguard
Revision where warranted	Permits correction when the challenge exposes error, omission, or unacceptable reasoning	Proposed safeguard
Documentation	Records the reason, challenge, response, and disposition proportionately	Proposed institutional safeguard
Appeal or recourse	Provides escalation when first-level review remains disputed or structurally inadequate	Proposed safeguard informed by second-opinion logic

A procedural model for high-stakes clinical decisions

For high-stakes decisions, contestability must connect interpersonal answerability with organizational assurance. Algorithmic auditing has been proposed as a structured way to examine clinical AI across deployment and use, including failures that may not be visible from model-development metrics alone [17]. Patient challenges should therefore not disappear after an individual case is resolved. Recurrent challenges, overrides, unexplained disparities, or repeated inability to produce defensible reasons can be treated as signals for quality and safety review. This is an institutional extension of the proposed right, not evidence that auditing itself satisfies contestability. Conversely, a technically sophisticated audit program cannot substitute for a patient's opportunity to challenge a decision that affects that patient.

Real-world clinical AI also requires safety mechanisms that operate beyond model approval, including monitoring and organizational processes capable of responding when systems behave unexpectedly [18]. Within the proposed model, institutions should therefore

allocate responsibility for routing challenges, identifying a qualified reviewer, documenting the response, and escalating patterns that may indicate system-level risk. The patient-facing pathway and the institutional safety pathway are complementary: one protects answerability in the individual case, while the other can learn from repeated failures of answerability. Institutional learning should not erase the distinction between quality improvement and individual remedy; a future system change is not itself an answer to a present patient whose decision remains contested.

Procedural intensity should nevertheless remain proportionate. A recent normative decision algorithm for introducing black-box AI into healthcare illustrates a broader principle: safeguards can rationally vary with the seriousness and reversibility of risk rather than being uniformly maximal [19]. The model proposed here applies that logic to contestability. Lower-consequence decisions may need only a concise reason and accessible questioning. High-consequence, opaque, strongly AI-influenced, or difficult-to-reverse decisions should receive fuller review, documented response, and recourse. Review should ordinarily occur early enough to matter to the clinical decision, although later sections consider circumstances in which urgency may justify temporary compression or deferral. No level is claimed to be empirically validated.

Figure 1 integrates these elements into a staged model of meaningful contestability for high-stakes clinical decisions. The model shows how a consequential decision should move from notice and a humanly attributable reason to an opportunity for patient challenge, empowered human reconsideration, substantive response, and—where the challenge identifies error, omission, or inadequately justified reasoning—revision or further recourse. Importantly, the pathway is not intended as a rigid linear protocol: procedural intensity should increase with the seriousness, opacity, AI influence, value sensitivity, vulnerability, and irreversibility of the decision, while unresolved or recurrent contestation can also generate signals for institutional quality and safety review. **Figure 1** therefore represents contestability as a connected process of answerability, reconsideration, correction, and escalation rather than as a single act of explanation or nominal human review.

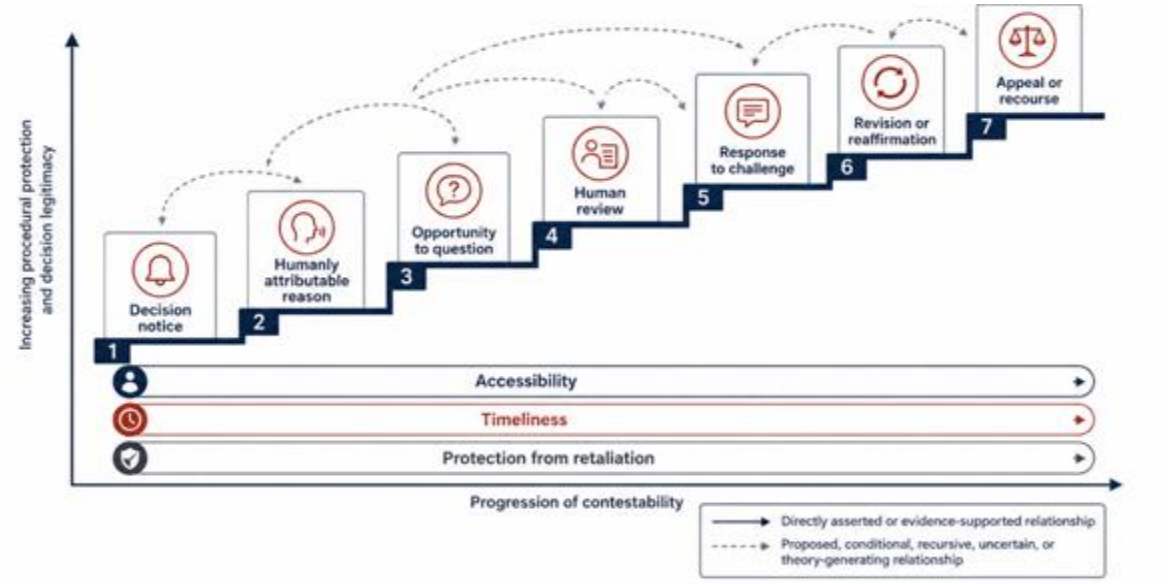

Figure 1. Procedural stages of meaningful contestability in high-stakes clinical decisions. This figure is an original conceptual synthesis developed for "The Right to a Human Reason: Contestability in High-Stakes Clinical Decisions." It uses the exact predefined architecture and visual grammar supplied for this article. Solid arrows indicate finect directly asserted or evidence-supported directional relationships; dashed arrows indicate proposed, conditional, recursive, uncertain, or theory-generating relationships where arrows are used. The figure does not represent measured effect sizes, validated thresholds, or a validated causal model.

Figure 1. Procedural stages of meaningful contestability in high-stakes clinical decisions

Emergency, scarcity, and expertise exceptions

A right to a human reason cannot be absolute in timing or procedural intensity. Emergency care can require action before a full contestability sequence is possible. Research on digital emergency-department triage shows that patients and professionals evaluate digitalization within a setting already structured by urgency, confidence, efficiency, and changing professional roles [20]. That evidence does not establish an ethical exemption. It supports the narrower conclusion that time-

critical contexts generate distinctive trade-offs. The proposal here is therefore a deferral principle: when completing the full process would create a material risk of harm through delay, notice, questioning, or review may be compressed temporarily, but a humanly attributable reason and appropriate retrospective review should follow when clinically feasible.

Scarcity raises a different problem. Institutions may lack specialists, independent reviewers, or time for elaborate reconsideration. Scarcity can justify proportionate procedure, but should not automatically convert an institution's resource deficit into a patient's loss of answerability. The quality of human review also depends on preserving expertise. A mixed-method review of AI-induced deskilling identifies a credible concern that repeated dependence on automated systems may weaken opportunities for independent skill use, while emphasizing that the evidence remains incomplete and context dependent [21]. A formal right would therefore become hollow if the humans designated to review decisions gradually lost the competence required to do so.

Professional adoption evidence likewise shows that AI use is shaped by barriers and facilitators operating at individual, technological, organizational, and wider system levels [22]. The proposed exception structure consequently distinguishes justified compression from abandonment. Institutions may vary how review is delivered, but should retain an accountable pathway, identify when deferred review is permitted, preserve competent reviewers, and document why ordinary safeguards could not be provided.

Table 3 presents "Levels of human review across categories of high-stakes clinical decision" for "The Right to a Human Reason: Contestability in High-Stakes Clinical Decisions."

Table 3. Levels of human review across categories of high-stakes clinical decision

Decision category	Proposed level of human review	Principal rationale
Diagnosis	Full review when diagnosis is consequential, materially AI-influenced, opaque, or disputed	Diagnostic disagreement may require independent reassessment; level is proposed
Treatment denial	Full reason, review, response, and recourse	Denial directly restricts an important clinical option; classification is proposed
Discharge	Enhanced review where foreseeable harm or serious disagreement exists	Consequences may be difficult to reverse; classification is proposed
Triage	Proportionate contemporaneous review; emergency deferral where delay creates material danger	Emergency trade-offs are evidence-informed ; deferral rule is proposed
Transplant selection	Full independent review and recourse where practicable	Scarcity and consequence intensify answerability; classification is proposed
Algorithm-supported risk classification	Review intensity scaled to downstream consequence and system influence	Risk-sensitive proportionality is informed ; levels are proposed
Other high-consequence decisions	Full pathway when stakes, opacity, vulnerability, or irreversibility are substantial	Proposed catch-all protection against procedural gaps

Equity and accessibility requirements

A formally available challenge is not necessarily an accessible one. Patients differ in health literacy, communication needs, disability, language, confidence in professional encounters, time, and capacity to withstand disagreement with institutions. Contestability should therefore be assessed by whether a patient can realistically use it, not merely whether a portal, telephone number, or review policy exists. This requirement is especially important when the relevant evidence is partly experiential. Normative analysis of AI-assisted pain assessment argues for epistemic humility because attempts to objectify pain can marginalize patient testimony and reproduce epistemic injustice [23]. Pain is a specific hard case rather than proof about every AI application, but it illustrates why human review must remain responsive to forms of knowledge not easily captured by structured data.

Patient perspectives toward healthcare AI are also heterogeneous across reported studies, including perceived benefits, ethical concerns, trust, and preferred conditions of use [24]. A single disclosure script or digital-only appeal mechanism should therefore not be presumed adequate. The proposed right requires accessible communication, reasonable opportunities for representation, and formats that enable patients to understand the reason and formulate a challenge.

Accessibility also has a temporal and relational dimension. A technically available review that arrives after an irreversible intervention, discharge, or allocation decision may have little corrective value. Likewise, patients may reasonably fear that persistent questioning will damage therapeutic relationships or influence future care. The proposal therefore treats timeliness and protection from retaliation as continuous safeguards rather than optional additions. This does not imply that every challenge must suspend treatment. It means that institutions should distinguish decisions that can safely proceed while review occurs from decisions for which delay in review would effectively extinguish the patient's opportunity to contest.

Equity also concerns who bears algorithmic error. Research on chest-radiograph algorithms documented underdiagnosis disparities affecting underserved groups in the examined datasets [25]. That finding is task-specific and cannot justify a claim that all clinical AI discriminates. It does establish that error burdens can be distributed unequally. Institutions should therefore treat differential access to contestation, not only differential model performance, as an equity question.

Institutional feasibility

The strongest objection to the proposal may be institutional rather than conceptual: meaningful contestability consumes time, expertise, documentation, and review capacity. Those burdens are real. Evidence from implementation of AI systems for predicting deterioration in adult hospitals shows that real-world integration involves persistent uncertainties, workflow choices, barriers, and enabling conditions [26]. The right should therefore not be described as implementation-ready. Institutions would need to determine which decisions trigger heightened review, who is qualified to reconsider them, how urgent cases are handled, and how individual challenges interact with quality and safety systems.

Governance architecture is also fragmented. A recent scoping review identified numerous healthcare AI governance frameworks while finding uneven coverage and operationalization across governance themes [27]. Framework abundance should not be confused with demonstrated effectiveness. For the proposed right, the practical implication is narrower: contestability cannot depend exclusively on an individual clinician's goodwill. Responsibility must be assigned for routing challenges, obtaining independent review where required, preserving records, and identifying recurrent patterns that warrant system-level attention.

A systematic review of empirical studies likewise identifies recurring concerns around transparency, fairness, accountability, privacy, trust, and governance, together with strategies proposed to address them [28]. These strategies remain heterogeneous and should not be represented as proven solutions. Institutional feasibility therefore requires local design and evaluation. Relevant outcomes would include patient comprehension, uptake of review, timeliness, workload, successful correction, frequency of reaffirmation, reviewer independence, differential access across patient groups, and unintended delays or defensive practices. Such evaluation could refine, narrow, or challenge the proposed model rather than merely confirm it.

Objections and limitations

The right may be criticized as privileging human judgment, producing procedural overload, or inviting repeated challenges to decisions that are already well supported. Those objections matter. The proposal does not assume that humans are more accurate, that every disagreement deserves another specialist opinion, or that patients possess an unlimited veto. Its claim is procedural: when authority is consequential and contestable reasons are reasonably available, affected patients should not be left with an unanswerable determination.

Another objection is that requiring human attribution may encourage false reassurance. A clinician could lend legitimacy to a system without genuinely understanding or reconsidering it. That risk is precisely why nominal human presence is insufficient. Competence, independence, authority, and the possibility of revision are substantive conditions, not ceremonial labels.

A further limitation concerns the boundary between moral right and institutional entitlement. Calling something a right can suggest a determinate duty-bearer, standardized remedy, and enforceable threshold. This article establishes none of those features as existing law. Its claim is instead that sufficiently consequential, answerable clinical authority generates a moral reason to structure opportunities for challenge. Different health systems could reject, narrow, or implement that claim differently. The proposal also leaves unresolved how often independent reviewers should be external to the initial team, what degree of explanation is sufficient when uncertainty cannot be reduced, and when repeated

appeals become disproportionate. Those are not minor administrative details; they are part of what future normative and empirical work must test.

The article is also limited by the evidence on which its practical implications rest. Much of the relevant literature is normative, conceptual, review-based, task-specific, or context-specific. No selected evidence validates the full proposed pathway, establishes its clinical effectiveness, or demonstrates that its benefits exceed its burdens across health systems. Jurisdictional law may confer narrower, broader, or differently structured protections. The proposed right therefore remains a normative framework requiring philosophical criticism, comparative legal analysis, patient and clinician research, accessibility testing, and prospective evaluation of its operational consequences.

Conclusion

The right to a human reason is proposed as a right to answerable clinical authority. It distinguishes being told how or that a decision was made from receiving a justification that can be questioned and, where warranted, changed. Its central demand is not that humans displace artificial intelligence, but that consequential clinical decisions remain attributable to persons and institutions capable of responding to the people affected.

Meaningful contestability therefore requires more than transparency or nominal human review. It requires an intelligible reason, a usable opportunity to question, competent reconsideration, a substantive response, possible revision, and proportionate recourse. These protections may require modification under genuine emergency or scarcity, and they must be accessible enough to avoid reproducing inequity. The proposal is neither an existing universal legal entitlement nor a validated intervention. Its value ultimately depends on whether further normative scrutiny and empirical evaluation show that it can protect patient agency without creating disproportionate delay, burden, or new forms of exclusion.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Hurley ME, Lang BH, Kostick-Quenet KM, Smith JN, Blumenthal-Barby J. Patient consent and the right to notice and explanation of AI systems used in health care. Am J Bioeth. 2025;25(3):102-14. doi:10.1080/15265161.2024.2399828
2. van de Sande D, Economou-Zavlanos N, van Genderen ME. Meaningful oversight of medical AI beyond human in the loop. NPJ Digit Med. 2026;9(1):569. doi:10.1038/s41746-026-02971-1
3. Tikhomirov L, Semmler C, McCradden M, Searston R, Ghassemi M, Oakden-Rayner L. Medical artificial intelligence for clinicians: The lost cognitive perspective. Lancet Digit Health. 2024;6(8). doi:10.1016/S2589-7500(24)00095-5
4. Funer F, Liedtke W, Tinnemeyer S, Klausen AD, Schneider D, Zacharias HU, et al. Responsibility and decision-making authority in using clinical decision support systems: An empirical-ethical exploration of German prospective professionals' preferences and concerns. J Med Ethics. 2024;50(1):6-11. doi:10.1136/jme-2022-108814
5. Muralidharan A, Savulescu J, Schaefer GO. AI and the need for justification (to the patient). Ethics Inf Technol. 2024;26(1):16. doi:10.1007/s10676-024-09754-w
6. Kerasidou A, Kerasidou CX. Epistemic authority and medical AI: Epistemological differences and challenges in medical practice. Med Health Care Philos. 2026;29(1):89-95. doi:10.1007/s11019-025-10306-2
7. Funer F. The deception of certainty: How non-interpretable machine learning outcomes challenge the epistemic authority of physicians. A deliberative-relational approach. Med Health Care Philos. 2022;25(2):167-78. doi:10.1007/s11019-022-10076-1
8. Lorenzini G, Arbelaez Ossa L, Shaw DM, Elger BS. Artificial intelligence and the doctor-patient relationship expanding the paradigm of shared decision making. Bioethics. 2023;37(5):424-9. doi:10.1111/bioe.13158
9. McCradden MD, Kirsch RE. Patient wisdom should be incorporated into health AI to avoid algorithmic paternalism. Nat Med. 2023;29(4):765-6. doi:10.1038/s41591-023-02224-8
10. Guerrero Quiñones JL. Using artificial intelligence to enhance patient autonomy in healthcare decision-

making. AI Soc. 2025;40(3):1917-26. doi:10.1007/s00146-024-01956-6
11. Padela AI, Hayek R, Tabassum A, Jotterand F, Qadir J. Assisting, replicating, or autonomously acting? An ethical framework for integrating AI tools and technologies in healthcare. Bioethics. 2026;40(1):52-60. doi:10.1111/bioe.70019
12. Ploug T, Holm S. The right to refuse diagnostics and treatment planning by artificial intelligence. Med Health Care Philos. 2020;23(1):107-14. doi:10.1007/s11019-019-09912-8
13. Hatherley J. Are clinicians ethically obligated to disclose their use of medical machine learning systems to patients? J Med Ethics. 2025;51(8):567-73. doi:10.1136/jme-2024-109905
14. Ploug T, Holm S. Consenting to and contesting AI use-different functions but both are necessary. Am J Bioeth. 2025;25(3):124-6. doi:10.1080/15265161.2025.2457721
15. Kempt H, Heilinger JC, Nagel SK. "I'm afraid I can't let you do that, Doctor": Meaningful disagreements with AI in medical contexts. AI Soc. 2023;38(4):1407-14. doi:10.1007/s00146-022-01418-x
16. Kempt H, Nagel SK. Responsibility, second opinions and peer-disagreement: Ethical and epistemological challenges of using AI in clinical diagnostic contexts. J Med Ethics. 2022;48(4):222-9. doi:10.1136/medethics-2021-107440
17. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health. 2022;4(5). doi:10.1016/S2589-7500(22)00003-6
18. Sittig DF, Singh H. Recommendations to ensure safety of AI in real-world clinical care. JAMA. 2025;333(6):457-8. doi:10.1001/jama.2024.24598
19. Allen JW, Wilkinson D, Savulescu J. When is it safe to introduce an AI system into healthcare? A practical decision algorithm for the ethical implementation of black-box AI in medicine. Bioethics. 2026;40(1):61-72. doi:10.1111/bioe.70032
20. Morlotti C, Cattaneo M, Paleari S, Manelli F, Locati F. The digitalization of emergency department triage: The perspectives of health professionals and patients. BMC Health Serv Res. 2024;24(1):1406. doi:10.1186/s12913-024-11862-8
21. Natali C, Marconi L, Dias Duran LD, Cabitza F. AI-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond. Artif Intell Rev. 2025;58(11):356. doi:10.1007/s10462-025-11352-1
22. Henzler D, Schmidt S, Koçar A, Herdegen S, Lindinger GL, Maris MT, et al. Healthcare professionals' perspectives on artificial intelligence in patient care: A systematic review of hindering and facilitating factors on different levels. BMC Health Serv Res. 2025;25(1):633. doi:10.1186/s12913-025-12664-2
23. Katz RA, Graham SS, Buchman DZ. The need for epistemic humility in AI-assisted pain assessment. Med Health Care Philos. 2025;28(2):339-49. doi:10.1007/s11019-025-10264-9
24. Bashkin O. Patient perspectives on artificial intelligence in healthcare: A global scoping review of benefits, ethical concerns, and implementation strategies. Int J Med Inform. 2025;203:106007. doi:10.1016/j.ijmedinf.2025.106007
25. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176-82. doi:10.1038/s41591-021-01595-0
26. van der Vegt AH, Campbell V, Mitchell I, Malycha J, Simpson J, Flenady T, et al. Systematic review and longitudinal analysis of implementing artificial intelligence to predict clinical deterioration in adult hospitals: What is known and what remains uncertain. J Am Med Inform Assoc. 2024;31(2):509-24. doi:10.1093/jamia/ocad220
27. Wang A, Freeman S, Magrabi F. Governance for safe and responsible AI in healthcare organisations: A scoping review of frameworks. NPJ Digit Med. 2026;9(1):516. doi:10.1038/s41746-026-02679-2
28. Wubineh BZ, Deriba FG, Gemeda FW. Ethical concerns and strategies for implementing artificial intelligence in healthcare: A review of empirical studies. BMC Med Ethics. 2026;27(1):53. doi:10.1186/s12910-026-01404-8