

When Predictive Recommendations Become Clinical Authority: A Theory of Responsibility, Patient Contestability, and Justified Reliance in Artificial Intelligence–Supported Healthcare Decisions

Anna Weber¹, Thomas Müller^{2*}, Priya Nair³, James Anderson⁴

¹Department of AI Ethics and Governance, Faculty of Philosophy, University of Vienna, Vienna, Austria.

²Department of Medical Informatics and Ethics, Faculty of Medicine, LMU Munich, Munich, Germany.

³Department of Health Policy and Ethics, Indian Institute of Technology Bombay, Mumbai, India.

⁴Department of Bioethics, Faculty of Medicine, University of Manchester, Manchester, United Kingdom.

*E-mail ✉ thomas.mueller@lmu.de

Abstract

Artificial intelligence–supported recommendations can influence clinical decisions long before any institution formally describes an algorithm as a decision-maker. This creates a problem that cannot be resolved by predictive accuracy alone: a recommendation may remain formally advisory while acquiring practical authority through workflow design, clinician dependence, default effects, interpretive asymmetry, or difficulty of override. This theoretical article examines when that transition occurs and what it implies for responsibility, patient contestability, and justified reliance. Using Decision-Rights / Normative Responsibility Theory, the analysis distinguishes predictive performance from practical authority and treats authority as a distribution of operative decision rights among clinicians, institutions, patients, and AI-mediated processes. The argument develops in three steps. First, clinical authority is shown to emerge when AI-supported outputs materially structure which options are considered, how disagreement is handled, and what burden falls on clinicians who depart from the recommendation. Second, responsibility is tied to actors' actual capacity to interpret, override, explain, escalate, and revise AI-mediated decisions, while patient contestability is defined as the effective opportunity to challenge both the recommendation and the institutional conditions under which it was relied upon. Third, justified reliance is conceptualized as conditional alignment among epistemic warrant, operative decision rights, answerable responsibility, and contestability. The resulting account rejects both accuracy-based authority and purely trust-based approaches. AI-supported recommendations may properly shape care, but reliance becomes normatively defensible only when the human and institutional actors expected to answer for the decision retain meaningful authority over it and patients retain a credible route to question its use.

Keywords: Artificial intelligence, Clinical decision-making, Decision rights, Professional responsibility, Patient contestability, Justified reliance

Introduction

The ethical problem created by artificial intelligence in clinical decision-making is often framed as a responsibility gap: when an AI-supported

recommendation contributes to an adverse decision, it may be difficult to determine whose contribution should carry moral weight and how responsibility should be distributed among clinicians, developers, and institutions [1]. That framing is necessary but incomplete. A more prior question concerns authority. If a clinician formally signs an order but the available options, probability estimates, alerts, workflow defaults, and justification requirements have already been structured by an AI-supported system, the nominal location of decision authority may not match its practical location.

This mismatch matters because responsibility can be assigned fairly only when the person or institution being

Access this article online

<https://smerpub.com/>

Received: 23 April 2026; Accepted: 27 July 2026

Copyright CC BY-NC-SA 4.0

How to cite this article: Weber A, Müller T, Nair P, Anderson J. When Predictive Recommendations Become Clinical Authority: A Theory of Responsibility, Patient Contestability, and Justified Reliance in Artificial Intelligence–Supported Healthcare Decisions. *Asian J Ethics Health Med*. 2026;6(2):23-34. <https://doi.org/10.51847/Y4wmkd6kz5>

held responsible possesses relevant control, knowledge, and opportunity to act. Analyses of clinical AI have accordingly distinguished prospective obligations from retrospective blame and questioned whether retaining a formally final human decision is sufficient when the system materially shapes the judgement that precedes it [2]. Related work on harmful epistemic dependence shows that a clinician may remain technically free to reject an algorithmic recommendation while becoming increasingly unable to assess it independently in the circumstances in which the decision must actually be made [3]. Responsibility therefore cannot be inferred from the presence of a human at the end of a workflow. The broader healthcare-AI literature reinforces this point. Ethical evaluation has long extended beyond accuracy to questions of accountability, privacy, bias, professional judgement, and the distribution of benefits and burdens [4, 5]. Trustworthiness, explainability, usability, and transparency likewise operate at the level of the clinical system and workflow rather than as isolated model properties [6]. More recent work on generative and predictive AI has intensified the need to convert high-level ethical principles into operational arrangements capable of governing who can use a system, what decisions it may shape, what must remain reviewable, and how failures can be challenged [7]. Yet operational governance remains underdeveloped in professional guidance on clinicians' use of AI-supported decision tools [8].

This article addresses that gap by asking when a predictive recommendation acquires practical clinical authority and under what conditions reliance on that recommendation can be justified. The analysis uses Decision-Rights / Normative Responsibility Theory to separate five concepts that are frequently collapsed: prediction, recommendation, authority, responsibility, and justified reliance. Prediction concerns the informational output of a system. Recommendation concerns how that output is presented as action-relevant. Practical authority concerns whether the recommendation materially governs the decision process. Responsibility concerns who must answer for the consequences and justification of the resulting choice. Justified reliance concerns whether using the recommendation was normatively defensible given the evidence, the distribution of decision rights, the actors' capacities, and the patient's opportunity to contest the process.

The central claim is conditional. AI does not acquire clinical authority simply by being accurate, sophisticated, or difficult to replace. Practical authority emerges when the organization of decision rights makes AI-supported output consequential for what clinicians can realistically perceive, choose, justify, override, or escalate. Where responsibility remains attached to clinicians or institutions, those actors must retain corresponding capacities to exercise meaningful control. Patient contestability provides an additional test: if an AI-mediated decision cannot be explained, questioned, reviewed, or reconsidered through an accessible human and institutional process, the apparent retention of human authority may be largely formal.

From AI prediction to clinical authority

Clinical AI enters practice through prediction, classification, prioritization, risk estimation, recommendation, or generation of decision-relevant information. Such capabilities may substantially augment clinical work [9]. They do not, however, determine who has authority to act. Ethical analyses of algorithmic decision-making therefore require attention to how recommendations are embedded within professional judgement rather than treating computational output as equivalent to a decision [10]. The distinction is especially important because practical authority can shift without an explicit declaration that authority has shifted.

One route is behavioural influence. In an experimental study involving 265 physicians, inaccurate advice reduced diagnostic performance whether the advice was described as coming from an AI system or from another source [11]. In a separate study of intensive-care prescribing decisions, exposure to AI-generated recommendations significantly altered clinicians' choices, while a simple explanatory layer did not produce an additional increase in influence beyond the AI recommendation itself [12]. These findings do not establish that AI possessed legitimate authority. They show that decision-support information can become behaviourally consequential even when the clinician is ostensibly free to disregard it.

A second route involves the conditions under which disagreement occurs. Eye-tracking work with safe and unsafe AI recommendations indicates that explanatory interfaces do not automatically ensure effective scrutiny of erroneous advice [13]. AI used as a second opinion can also create false confirmation when apparent agreement

is treated as independent corroboration even though the underlying systems or reasoning pathways are not genuinely independent [14]. Practical authority therefore depends partly on what may be called override friction: the cognitive, procedural, temporal, and organizational burden associated with departing from an AI-supported recommendation. A system can remain formally advisory while acquiring greater effective control if rejecting it demands more documentation, time, expertise, or institutional risk than accepting it.

A third route concerns presentation. Clinician-facing explanation format can influence willingness to accept AI recommendations. In one study, clinically framed explanations increased acceptance relative to results-only or SHAP-style presentations [15]. Yet such effects cannot be equated with improved epistemic justification. A reader study in obstetric imaging found marked person-level variability in trust, reliance, and performance; explanations did not produce a statistically significant overall improvement in trust or reliance beyond the model-supported condition [16]. The implication is not that explanations are unimportant. It is that their persuasive effect and their epistemic value must be distinguished.

That distinction becomes sharper in critiques of current explainable-AI approaches. An explanation may create the appearance of intelligibility without establishing that the explanation faithfully represents the model's causal basis, that the clinician can evaluate its relevance, or that the recommendation is clinically warranted in the particular case [17, 18]. A clinician who receives an explanation may therefore possess more information without possessing greater practical authority. What matters is whether the information supports an independent capacity to interrogate, compare, reject, or escalate the recommendation.

The transition from prediction to practical authority is thus configurational rather than binary. AI-supported output becomes increasingly authoritative when it shapes the option set, controls the salience of evidence, establishes defaults, raises the cost of departure, or creates a justificatory asymmetry in which agreement requires little explanation but disagreement demands extensive defence. The same model can occupy different authority positions in different organizations because practical authority is produced by workflow and decision rights as much as by technical performance.

Table 1 summarizes the distinctions discussed above.

Table 1. Conditions through which an AI-supported recommendation can acquire practical clinical authority without formal delegation

Authority-producing condition	Clinical manifestation	How practical control can shift	What preserves operative clinician authority	Normative significance
Salience concentration	AI output is foregrounded before competing clinical evidence	Attention is channelled toward the model's framing of the case	Independent access to clinically relevant evidence and competing interpretations	Influence may increase without explicit transfer of authority
Default structuring	Recommended action is preselected, prioritized, or embedded in the workflow	Acceptance becomes procedurally easier than departure	Symmetrical ability to accept, modify, defer, or reject	Default convenience must not be mistaken for clinical warrant
Override friction	Rejecting the recommendation requires extra time, documentation, escalation, or reputational risk	Nominal discretion becomes costly to exercise	Realistic override pathways with proportionate documentation requirements	Formal final authority may coexist with constrained practical control
Justificatory asymmetry	Agreement is treated as self-explanatory while disagreement requires defence	The recommendation becomes the presumed standard	Equal scrutiny of agreement and disagreement	Responsibility cannot fairly remain human if only human disagreement is burdened
Epistemic dependence	Clinician cannot independently assess the model's reasoning or evidential basis in context	Reliance may become functionally unavoidable	Sufficient competence, information, alternative evidence, and escalation support	Dependence can alter what responsible reliance requires
Confirmation compression	AI agreement is treated as independent corroboration	Apparent consensus discourages further review	Attention to source independence and reasons for agreement	Multiple agreeing outputs do not necessarily provide multiple independent reasons

Explanation-mediated persuasion	Interface presentation increases acceptance or confidence	Persuasive force may exceed epistemic value	Evaluation of relevance, fidelity, and case-specific applicability	Explainability may alter behaviour without establishing justified reliance
Workflow entrenchment	AI output becomes embedded in routine ordering, triage, or referral processes	Repeated use normalizes deference and narrows perceived alternatives	Periodic reassessment of decision rights, override patterns, and local consequences	Authority can accumulate institutionally even if each use is formally advisory

Responsibility and patient contestability

Once AI-supported output becomes practically authoritative, the question of responsibility changes. The relevant issue is no longer simply whether a clinician was present, but whether the clinician possessed the decision rights necessary to justify being answerable for the outcome. One frequently proposed safeguard is disclosure: patients should be informed when machine learning contributes to care. Yet a general obligation to disclose every use of medical machine learning cannot be established solely by invoking autonomy, risk, rights, or accountability; the ethical significance of disclosure depends on what the system does and how its use affects the patient's decision situation [19]. Disclosure is therefore better treated as one possible component of contestability rather than as its equivalent.

Patient-facing evidence supports this more demanding interpretation. In vignette research, attitudes toward AI-supported clinical decisions varied with the information patients received and with the relationship between clinician judgement and AI output [20]. Information matters most when it clarifies what role the system played, who interpreted the result, who could depart from it, and what happens if the patient disputes the recommendation. Merely stating that "AI was used" does not reveal whether AI functioned as background analysis, influential advice, a near-mandatory default, or the effective determinant of the available treatment pathway. Public and patient studies also show that human involvement remains ethically salient, although it should not be romanticized. Concerns about privacy, oversight, and patient-centredness coexist with recognition that AI may contribute useful capabilities [21, 22]. Patients interviewed about AI-supported cardiac decision-making valued clinician assessment, personalization, and the preservation of a human role while also recognizing potential benefits of computational support. The implication is not that every patient prefers maximal human intervention. It is that the institutional process must make clear who can hear, interpret, and respond

when a patient questions an AI-mediated recommendation.

This becomes especially important when the patient cannot directly express a contemporaneous preference. German physicians surveyed about AI systems designed to predict the preferences of incapacitated patients were cautious about substituting prediction for ethical deliberation and generally viewed human support as important [23]. Normative analysis reaches a parallel conclusion from a different direction: accurate preference prediction does not by itself constitute respect for autonomy because autonomy concerns agency, authorship, and the normative status of the person's choices, not merely the probability of reproducing what that person might have selected [24]. A system can therefore be highly predictive yet remain incapable of holding the decision rights that belong to patients, surrogates, clinicians, or legitimate institutions.

Responsibility also becomes more difficult where clinicians and AI disagree. Treating AI as a second opinion introduces epistemic and professional obligations that differ from ordinary deference to another clinician because the system's basis for judgement, competence boundaries, and accountability structure are different [25]. Questions of fairness further complicate reliance: a system can satisfy a local fairness criterion while being poorly suited to the broader clinical purpose for which it is deployed [26]. Interviews with healthcare professionals examining AI-supported resource allocation similarly identify concerns about equity, transparency, monitoring, and organizational responsibility rather than locating ethical responsibility in the model alone [27].

Patient contestability should therefore be understood as an effective opportunity to challenge an AI-mediated clinical decision and obtain a response from an actor with authority to reconsider it. It has at least three dimensions. First, the patient must be able to identify that a consequential recommendation has been materially shaped by AI when that fact is relevant to understanding the decision. Second, there must be an answerable human

or institutional actor capable of explaining the basis and limits of reliance. Third, the challenge must be capable of affecting the process: triggering reconsideration, alternative review, escalation, or acknowledgement that the recommendation cannot appropriately be overridden. Contestability is hollow where a patient can ask a question but no actor possesses practical authority to alter what follows.

This account also changes the meaning of professional responsibility. Clinicians remain responsible for professional judgement where they retain meaningful control, but institutions cannot assign them unrestricted responsibility while constraining their ability to access relevant information, obtain alternatives, override defaults, or escalate concerns. Likewise, institutions cannot describe AI as “only advisory” if organizational design predictably makes deviation impracticable. Responsibility should track operative authority rather than formal labels.

A theory of justified reliance

Trust is often used to describe the desired relationship between clinicians, patients, and medical AI, but trust alone is too indeterminate to serve as the criterion for justified reliance. Stakeholders considering AI-based clinical prediction invoke heterogeneous criteria, including data quality, relevance, reliability, clinician endorsement, interpersonal relations, and institutional assurances [28]. These criteria help explain when people feel willing to depend on AI. They do not establish when that dependence is normatively justified.

The distinction is reinforced by philosophical work questioning whether medical AI is properly an object of interpersonal trust at all. Reliability can be assessed without attributing the intentions, commitments, or moral agency ordinarily associated with a trustworthy person [29]. Institutional approaches to medical AI attempt to address this difficulty by “enveloping” systems within legal, professional, technical, and organizational arrangements that constrain how they may be used [30]. These approaches are valuable because they relocate part of the ethical problem from the internal properties of the model to the surrounding decision environment.

Empirical work on appropriate reliance similarly suggests that the relevant unit of analysis is the human–AI decision configuration rather than the model in isolation. In a pathology study, majority voting among doctors improved measures of appropriate reliance relative to an individual pathologist–AI configuration

[31]. That result should not be generalized into a universal recommendation for collective voting. Its theoretical importance lies elsewhere: the same algorithmic output can support different reliance quality depending on how decision rights and independent judgement are organized around it.

Neither fairness nor trust resolves the problem independently. Technical fairness interventions cannot determine by themselves what distributive or clinical conception of justice should govern a system’s use [32]. Conversely, a reliability-centred account may explain why depending on an AI system can be practically rational even where interpersonal trust terminology is misplaced [33]. A theory of justified reliance must therefore distinguish the epistemic question—whether there are good reasons to expect the recommendation to perform adequately in the present context—from the normative question—whether the institutional arrangement governing reliance preserves legitimate decision rights and responsibility.

The proposed account treats justified reliance as requiring alignment among four conditions.

Epistemic warrant concerns whether the recommendation is supported by evidence relevant to the patient, setting, task, and decision at hand. Performance measures matter here, but they are not enough; applicability, uncertainty, failure modes, and the clinician’s capacity to recognize out-of-scope use also matter.

Operative decision rights concern whether the actors designated as decision-makers can genuinely inspect, interpret, accept, reject, modify, defer, or escalate the recommendation. A formally retained human “final say” has little normative significance if departure is practically unavailable or epistemically impossible.

Answerable responsibility concerns whether responsibility is assigned to actors whose authority, competence, information, and institutional position allow them to discharge it. Responsibility becomes distorted when institutions retain control over system selection, configuration, defaults, and escalation while attributing the resulting decisions entirely to individual clinicians.

Contestability concerns whether patients and affected professionals can question the recommendation or the conditions of its use and receive a response from someone able to reconsider the process. Contestability does not guarantee that the challenge succeeds. It requires that disagreement enter a decision pathway

capable of producing review, explanation, or justified refusal.

These conditions are mutually constraining rather than sequential. Strong epistemic performance cannot compensate for the absence of decision rights when clinicians are expected to carry responsibility. Extensive clinician discretion cannot compensate for a system whose evidential applicability cannot be assessed. Transparent explanation cannot compensate for a patient challenge process that lacks an actor empowered to

respond. Institutional oversight cannot compensate for a workflow that makes human review purely ceremonial.

Figure 1 represents these relationships as a decision-rights field rather than a one-way progression. Predictive output can become practically authoritative through several reinforcing routes—workflow defaults, epistemic dependence, explanation-mediated persuasion, or override friction—while justified reliance remains conditional on whether epistemic warrant, operative human authority, responsibility, and contestability remain aligned.

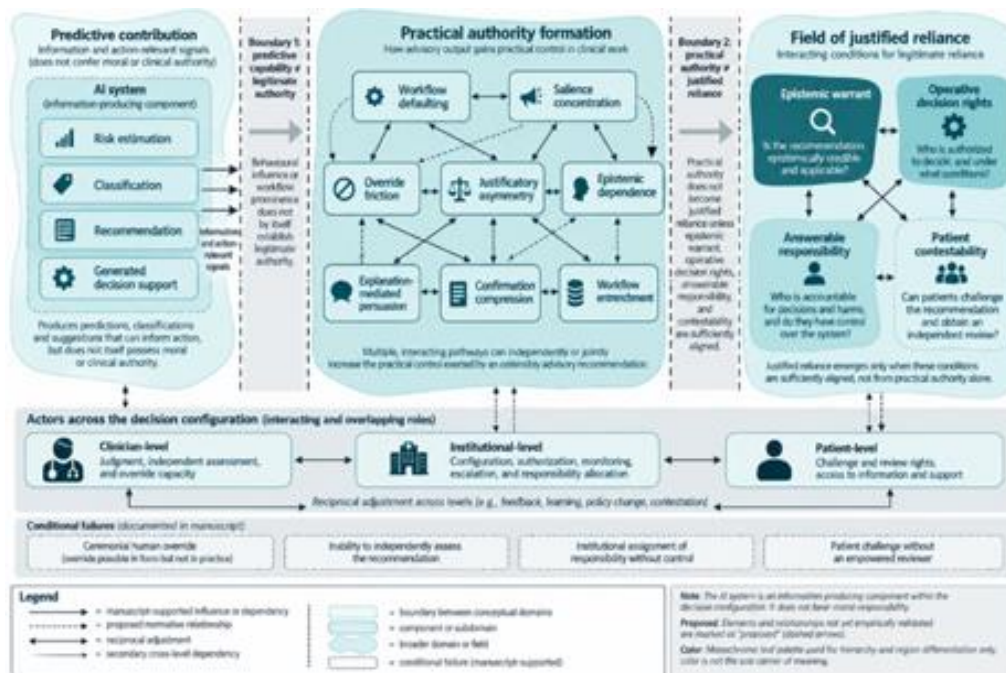


Figure 1. Clinical authority under algorithmic advice: where predictive influence, operative decision rights, and justified reliance diverge

The theoretical consequence of this configuration is that practical authority and justified reliance may diverge. A recommendation may exert substantial authority yet fail the conditions of justified reliance because clinicians cannot meaningfully interrogate it, institutions allocate responsibility asymmetrically, or patients cannot obtain review. Conversely, reliance can be justified even where

clinicians frequently accept AI recommendations, provided acceptance remains grounded in relevant evidence, departure remains genuinely available, responsibility tracks actual control, and challenge mechanisms remain operative.

Table 2 summarizes the distinctions discussed above.

Table 2. Who can decide, who must answer, and where the patient can contest an AI-mediated clinical choice

Decision-right site	Operative right that must remain identifiable	Actor positioned to exercise the right	Corresponding responsibility	Patient-contestability interface	Failure mode if right and responsibility diverge
Interpretation of AI output	Determine whether the output is clinically	Clinician, supported by relevant specialist	Assess relevance, uncertainty, and limits	Ask how the recommendation was interpreted	Clinician is blamed for an output that could not realistically be assessed

	applicable to the present case	or institutional resources		and what evidence was considered	
Acceptance or modification	Accept, adapt, or reject the recommendation	Clinician or authorized clinical team	Justify reliance or departure	Request an alternative clinical interpretation or second review	Nominal choice exists but organizational defaults make disagreement impracticable
Override	Depart from the AI-supported action pathway	Clinician or designated supervisory pathway	Use override responsibly and document material reasons where appropriate	Ask whether override was possible and who could authorize it	Human “final say” is ceremonial because override carries prohibitive friction
Escalation	Refer unresolved disagreement or uncertainty to a higher-authority process	Clinician, team, ethics service, specialist, or governance function as appropriate	Recognize when individual competence or authority is insufficient	Trigger or request escalation where clinically and ethically appropriate	Responsibility remains individualized despite institutionally controlled constraints
Patient challenge	Question the use, interpretation, or consequence of AI-supported advice	Patient or legitimate surrogate	Institution and clinician must provide a meaningful response	Receive explanation, reconsideration, second review, or justified refusal	Contestability becomes informational only and cannot affect the decision process
System configuration	Determine when the tool appears, what it recommends, and how strongly workflow privileges it	Healthcare institution and authorized governance actors	Ensure that defaults and constraints do not silently redistribute decision authority	Challenge institutional reliance where configuration materially shapes care	Institution controls practical authority while externalizing responsibility to clinicians
Monitoring and revision	Detect recurrent override patterns, bias, misuse, workflow dependence, or out-of-scope use	Institution, governance committee, clinical leadership	Revise or withdraw unsafe or unjustified configurations	Raise concerns about recurrent systemic patterns rather than isolated decisions	Harmful authority structures persist because each event is treated as an individual error
Post-event accountability	Determine whose decisions, system choices, and governance arrangements materially contributed to an outcome	Clinician, institution, developers or vendors where their relevant contribution is established	Answer in proportion to actual control, knowledge, and contribution	Obtain a review that examines both individual and institutional factors	Formal responsibility is allocated according to role labels rather than operative authority

Implications for AI-supported healthcare decisions

The theoretical account has direct consequences for how AI-supported decisions should be organized. The first implication is that governance should specify decision rights as explicitly as model requirements. Ethical and regulatory analyses of large language models already distribute obligations across users, developers, and regulators rather than treating responsibility as a property of the system itself [34]. The same logic applies to predictive clinical tools. Institutions should define who may rely on an output, who may override it, when

escalation is required, and which actor must review recurrent disagreement or failure.

Institutional governance therefore becomes part of the ethical architecture of reliance. A health-system case study has shown how responsible-use principles can be translated into operational guidance addressing accountability, transparency, safety, privacy, fairness, and oversight [35]. For the present account, the important point is not that one governance template should be universal. It is that decision rights must be visible in workflow design. If an institution selects the system,

configures its defaults, determines when it appears, and controls escalation pathways, then it also bears responsibility for those choices.

A second implication concerns the clinician. Empirical-ethical work with prospective health professionals indicates that perceived responsibility is closely connected to decision authority, professional experience, and the opportunity to scrutinize decision-support recommendations [36]. Assigning responsibility without the authority or competence needed to exercise judgement creates a structural mismatch. Clinicians should therefore have realistic access to alternative evidence, usable override routes, and escalation support where model output conflicts with clinical assessment.

A third implication concerns justice and system fitness. A recommendation may be technically accurate yet still be unsuitable for justified reliance if its deployment produces unfair exclusion, distorts resource allocation, or

shifts burdens toward groups not adequately represented in development or validation. Fairness cannot be reduced to a single metric [26], and qualitative evidence from healthcare professionals shows that transparency, monitoring, and equity concerns are intertwined with organizational resource-allocation decisions [27]. Institutional permission to rely on AI must therefore remain revisable when deployment consequences diverge from the conditions under which the system was initially judged acceptable.

Figure 2 expresses these implications as a cross-level responsibility field. The central question is not whether the system is “responsible,” but whether each consequential AI-mediated decision can be traced to an actor or institution with both the authority and the capacity to answer for it, while preserving a route through which the patient can challenge the process.

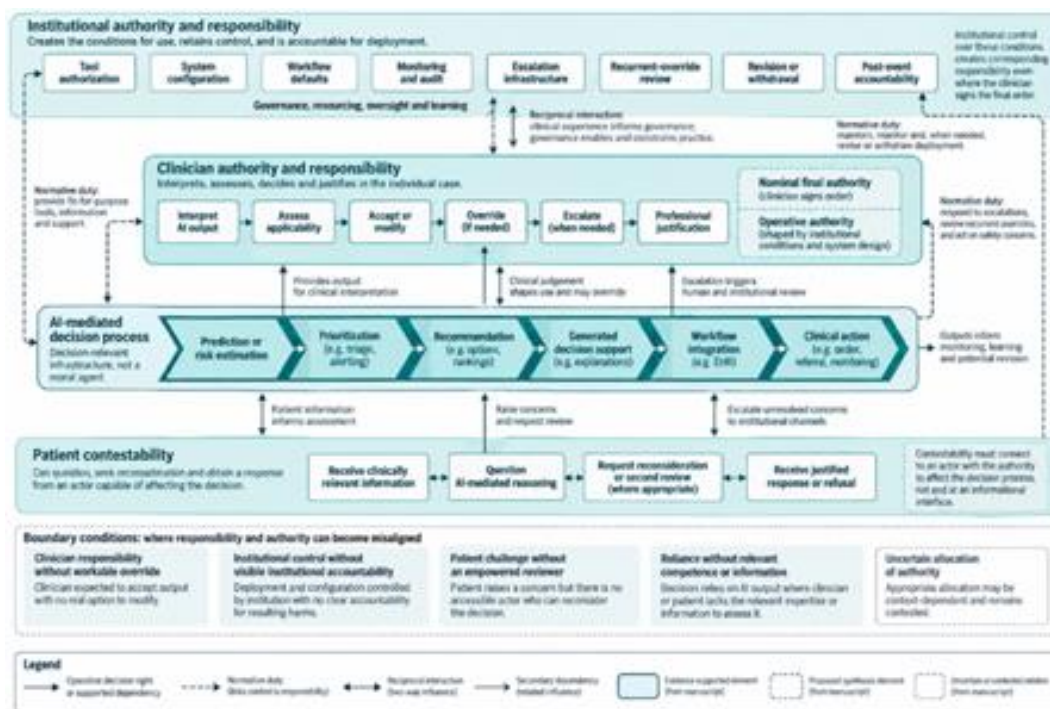


Figure 2. Authority without abdication: keeping responsibility and contestability attached to AI-mediated clinical decisions

The cross-level allocation also clarifies why responsibility cannot be reduced to the clinician’s final click. When an institution controls authorization, system configuration, monitoring, and escalation, it controls conditions that determine how much practical authority the recommendation acquires. At the same time,

institutional responsibility does not erase professional responsibility where clinicians possess meaningful knowledge and control. The normative task is to keep responsibility proportionate to operative authority while ensuring that affected patients can reach an actor able to reconsider how that authority was exercised.

Table 3 summarizes the distinctions discussed above.

Table 3. Conditions that must be satisfied before AI-supported reliance can be treated as justified

Reliance condition	Required clinical or institutional test	Decision right that must remain operative	Patient-facing consequence	Failure signal
Epistemic warrant	Is the recommendation sufficiently relevant, validated, and interpretable for this patient, setting, and decision?	Ability to examine applicability, uncertainty, and competing evidence	A meaningful account of why the output was considered relevant	Reliance continues despite known out-of-scope use or unresolved evidential mismatch
Operative decision rights	Can the designated clinician or team realistically accept, reject, modify, defer, or escalate the recommendation?	Genuine authority over the clinical action pathway	The patient can identify who can reconsider the recommendation	Human “final say” exists only formally
Answerable responsibility	Does the actor expected to answer for the decision possess the knowledge, authority, and institutional support needed to do so?	Authority proportionate to assigned responsibility	Clear identification of who is answerable for the decision and its review	Responsibility is individualized while control is institutionally constrained
Contestability	Can a patient or legitimate surrogate question the recommendation or its use and reach an empowered reviewer?	Ability to trigger explanation, reconsideration, second review, or justified refusal	A challenge can affect the process rather than merely be recorded	Questions are heard but cannot reach an actor with revision authority
Justice and institutional fitness	Does use remain compatible with fair access, resource stewardship, and the purpose for which the system is deployed?	Institutional authority to restrict, revise, or withdraw use	Protection against routinized disadvantage produced through workflow design	Accuracy is treated as sufficient despite unfair or unsuitable deployment
Ongoing revisability	Are recurrent overrides, failures, distributional effects, and workflow dependencies reviewed over time?	Authority to change configuration, escalation rules, or deployment status	Systemic concerns can lead to institutional correction	Repeated problems are treated only as isolated clinician errors

Results and Discussion

The analysis explains why predictive performance, practical authority, and justified reliance can diverge. Behavioural studies show that clinicians may change decisions in response to AI advice even when the system has no formal authority [11]. Explanation can alter acceptance while failing to guarantee safer or more appropriate reliance [15]. Clinician-level variability further undermines the assumption that a single interface or explanation standard will generate uniform oversight [16]. These findings support a theory in which authority is located partly in the structure of the decision environment rather than exclusively in either the model or the clinician formally designated to make the final decision.

This position differs from a trust-centred account. Trust criteria matter because clinicians and patients need reasons to depend on a system [28], but the concept of trust remains contested when applied to AI [29]. Reliability may justify some practical dependence without establishing interpersonal trust, while institutional envelopment can constrain how a system is

used [30]. The present theory therefore treats trust as one possible indicator of perceived acceptability rather than the decisive normative test. A system can be trusted too much, trusted for the wrong reasons, or reliably useful within an arrangement that allocates responsibility unfairly.

The account also differs from a fairness-first approach. Technical fairness measures illuminate important distributional properties, yet they cannot determine by themselves which conception of justice should govern a clinical use [32]. Likewise, the fact that collective procedures can improve measured appropriate reliance in one task does not establish a universal rule for human–AI collaboration [31]. Decision rights must be specified for the particular clinical setting, professional structure, patient population, and institutional purpose.

Several boundary conditions follow. First, meaningful human override may be impossible in time-critical settings. In such cases, justified reliance depends more heavily on ex ante authorization, validation, and institutional accountability. Second, greater human discretion is not automatically ethically preferable. A

clinician who lacks the relevant competence or evidence may exercise override irresponsibly. Third, contestability can legitimately be limited where overriding emergency, public-health, or justice obligations constrain individual choice, but the limitation itself should be explainable and institutionally reviewable. Fourth, responsibility can be distributed across several actors without implying equal responsibility. Contribution, control, knowledge, role, and capacity remain relevant to attribution [1].

These limitations expose an important failure condition for superficially “human-in-the-loop” systems. If the clinician formally retains override authority but cannot understand when departure is warranted, cannot access an alternative assessment, or faces organizational penalties that make override unrealistic, human oversight can become ceremonial. Harmful epistemic dependence is particularly relevant in such configurations because the capacity to reject a recommendation has limited normative value if the clinician lacks an adequate basis for deciding whether rejection is appropriate [3]. The solution cannot be to require clinicians to contradict systems more frequently. It is to design decision arrangements in which agreement and disagreement remain reason-responsive.

The theory is therefore neither anti-automation nor committed to preserving human control for its own sake. Its concern is congruence. Where institutions expect clinicians to bear responsibility, clinicians must retain decision rights capable of grounding that responsibility. Where institutions substantially constrain those rights, institutional responsibility must become correspondingly visible. Where patients are affected by AI-mediated authority, contestability must reach someone who can review the decision rather than merely explain that a model was used.

This account also generates discriminating implications. Two systems with equal accuracy may differ normatively if one permits meaningful override and the other embeds a costly default. Two interfaces with equally clear explanations may differ if only one supports independent review. Two institutions using the same model may differ if one monitors recurrent overrides and revises deployment while the other treats each disagreement as individual clinician error. The unit of ethical evaluation is therefore the AI-mediated decision configuration, not the predictive model alone.

That shift also limits what the theory claims. The present analysis does not establish a universal quantitative threshold at which influence becomes authority, nor does

existing evidence support one. Practical authority is identified by the organization of consequential decision rights: which actor can shape the option set, impose or resist defaults, authorize departure, demand justification, escalate disagreement, and revise the conditions of future use. Different configurations can therefore require different distributions of responsibility even when they deploy the same model.

Conclusion

Predictive recommendations become practically authoritative when they shape what clinicians can realistically perceive, choose, justify, override, or escalate. That authority is not established by accuracy, explanation, or formal designation alone. Reliance is justified only when epistemic warrant, operative decision rights, answerable responsibility, and patient contestability remain sufficiently aligned within the clinical and institutional setting.

AI may appropriately influence care, sometimes strongly, without becoming a moral agent or displacing human and institutional responsibility. The ethically relevant question is therefore not whether clinicians used AI, but whether the distribution of practical authority created by that use remained congruent with the distribution of responsibility and preserved a meaningful route for patient challenge. Where those conditions fail, formal human oversight cannot by itself make reliance justified.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Barman KG, Pawlowski P, Debrabander J. Reframing the responsibility gap in medical artificial intelligence: Insights from causal selection and authorship attribution. *J Med Ethics*. 2026;52(e1). doi:10.1136/jme-2024-110600
2. Smith H, Birchley G, Ives J. Artificial intelligence in clinical decision-making: Rethinking personal moral responsibility. *Bioethics*. 2024;38(1):78-86. doi:10.1111/bioe.13222

3. Pozzi G, Buijsman S, van den Hoven J. Harmful epistemic dependence on medical machine learning and its moral implications. *J Med Ethics*. 2025;52(1):42-49. doi:10.1136/jme-2024-110552
4. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. *PLoS Med*. 2018;15(11). doi:10.1371/journal.pmed.1002689
5. Morley J, Machado CCV, Burr C, Cows J, Joshi I, Taddeo M, et al. The ethics of AI in health care: A mapping review. *Soc Sci Med*. 2020;260:113172. doi:10.1016/j.socscimed.2020.113172
6. Cuttillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD, et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit Med*. 2020;3:47. doi:10.1038/s41746-020-0254-2
7. Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, Miao D, et al. Generative artificial intelligence and ethical considerations in health care: A scoping review and ethics checklist. *Lancet Digit Health*. 2024;6(11). doi:10.1016/S2589-7500(24)00143-2
8. Smith H, Downer J, Ives J. Clinicians and AI use: Where is the professional guidance? *J Med Ethics*. 2024;50(7):437-41. doi:10.1136/jme-2022-108831
9. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
10. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics*. 2020;46(3):205-11. doi:10.1136/medethics-2019-105586
11. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: Susceptibility in deployment of clinical decision-aids. *NPJ Digit Med*. 2021;4(1):31. doi:10.1038/s41746-021-00385-9
12. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making. *NPJ Digit Med*. 2023;6(1):206. doi:10.1038/s41746-023-00955-z
13. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations. *NPJ Digit Med*. 2024;7(1):202. doi:10.1038/s41746-024-01200-x
14. Rosenbacke R, Melhus Å, McKee M, Stuckler D. AI and XAI second opinion: The danger of false confirmation in human-AI collaboration. *J Med Ethics*. 2025;51(6):396-9. doi:10.1136/jme-2024-110074
15. Hur S, Lee Y, Park J, Jeon YJ, Cho JH, Cho D, et al. Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. *NPJ Digit Med*. 2025;8(1):578. doi:10.1038/s41746-025-01958-8
16. Nicolson A, Bradburn E, Gal Y, Papageorgiou AT, Noble JA. The human factor in explainable artificial intelligence: Clinician variability in trust, reliance, and performance. *NPJ Digit Med*. 2025;8(1):658. doi:10.1038/s41746-025-02023-0
17. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11). doi:10.1016/S2589-7500(21)00208-9
18. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. *Science*. 2021;373(6552):284-6. doi:10.1126/science.abg1834
19. 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
20. Kim B, Ryan K, Kim JP. Assessing the impact of information on patient attitudes toward artificial intelligence-based clinical decision support (AI/CDS): A pilot web-based SMART vignette study. *J Med Ethics*. 2025;51(8):541-9. doi:10.1136/jme-2024-110080
21. Witkowski K, Dougherty RB, Neely SR. Public perceptions of artificial intelligence in healthcare: Ethical concerns and opportunities for patient-centered care. *BMC Med Ethics*. 2024;25(1):74. doi:10.1186/s12910-024-01066-4
22. Maris MT, Koçar A, Willems DL, Pols J, Tan HL, Lindinger GL, et al. Ethical use of artificial intelligence to prevent sudden cardiac death: An interview study of patient perspectives. *BMC Med Ethics*. 2024;25(1):42. doi:10.1186/s12910-024-01042-y
23. Benzinger L, Epping J, Ursin F, Salloch S. Artificial Intelligence to support ethical decision-making for

- incapacitated patients: A survey among German anesthesiologists and internists. *BMC Med Ethics*. 2024;25(1):78. doi:10.1186/s12910-024-01079-z
24. Annoni M. It is not about autonomy: Realigning the ethical debate on substitute judgement and AI preference predictors in healthcare. *J Med Ethics*. 2025;51(7):450-5. doi:10.1136/jme-2024-110343
 25. 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
 26. Pruski M. What does it mean for a clinical AI to be just: Conflicts between local fairness and being fit-for-purpose? *J Med Ethics*. 2026;52(e1). doi:10.1136/jme-2023-109675
 27. Yüksel Elgin C, Elgin C. Ethical implications of AI-driven clinical decision support systems on healthcare resource allocation: A qualitative study of healthcare professionals' perspectives. *BMC Med Ethics*. 2024;25(1):148. doi:10.1186/s12910-024-01151-8
 28. Kostick-Quenet K, Lang BH, Smith J, Hurley M, Blumenthal-Barby J. Trust criteria for artificial intelligence in health: Normative and epistemic considerations. *J Med Ethics*. 2024;50(8):544-51. doi:10.1136/jme-2023-109338
 29. Hatherley JJ. Limits of trust in medical AI. *J Med Ethics*. 2020;46(7):478-81. doi:10.1136/medethics-2019-105935
 30. Zuchowski LC, Zuchowski ML, Nagel E. A trust based framework for the envelopment of medical AI. *NPJ Digit Med*. 2024;7(1):230. doi:10.1038/s41746-024-01224-3
 31. Gu H, Yang C, Magaki S, Zarrin-Khameh N, Lakis NS, Cobos I, et al. Majority voting of doctors improves appropriateness of AI reliance in pathology. *Int J Hum Comput Stud*. 2024;190:103315. doi:10.1016/j.ijhcs.2024.103315
 32. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit Health*. 2020;2(5). doi:10.1016/S2589-7500(20)30065-0
 33. Mainz JT. Medical AI: Is trust really the issue? *J Med Ethics*. 2024;50(5):349-50. doi:10.1136/jme-2023-109414
 34. Ong JCL, Chang SYH, William W, Butte AJ, Shah NH, Chew LST, et al. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit Health*. 2024;6(6). doi:10.1016/S2589-7500(24)00061-X
 35. Saenz AD, Mass General Brigham AI Governance Committee, Centi A, Ting D, You JG, Landman A, et al. Establishing responsible use of AI guidelines: A comprehensive case study for healthcare institutions. *NPJ Digit Med*. 2024;7(1):348. doi:10.1038/s41746-024-01300-8
 36. 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