Philosophy of Medicine
Original Research

The Algorithmic Gaze: Subject Occlusion, Clinical Epistemology, and the Reconstruction of the Patient in AI-Mediated Medicine

Daniela Oehring[1]

1 School of Health Professions, University of Plymouth, Plymouth, United Kingdom. Email: daniela.oehring@plymouth.ac.uk

Abstract

The integration of AI into clinical medicine is an event in the genealogy of medical visibilities where causal-mechanistic and contextual reasoning are required. Extending Dai and Guo (2025) through Pearl's causal hierarchy, a four-level taxonomy of consciousness, and a politicized conception of embodied care, this article develops two arguments: (1) only an agent that can be held to account can fail morally, and (2) associative systems cannot take up narrative and contextual material. Five epistemic deficits ground five governance responses, four resting on the knowledge argument; the responsibility argument distinguishes structural conditions from empirical performance and is conditional on non-functionalism.

1. Introduction

The deployment of artificial intelligence (AI) in clinical medicine is typically framed as a technical advance: better diagnostics, faster triage, and more efficient allocation of resources. This framing, pervasive in both the policy literature and the clinical press, obscures a deeper transformation. When an AI system interprets a retinal scan, triages a patient in the emergency department, or recommends a treatment pathway, it does not merely assist the clinician; it restructures the epistemic relationship between doctor, patient, and disease.[1] The clinical encounter (historically a site of embodied interpretation, narrative exchange, and diagnostic judgment) is reconstituted as a data-processing event.

[1] Throughout this article, "clinical AI" refers to a heterogeneous family of systems, including diagnostic imaging tools (for example, the DeepMind retinal disease system discussed in section 2.1 [De Fauw et al. 2018]; the IDx-DR autonomous diabetic retinopathy screener; and chest-imaging triage systems such as Aidoc and Viz.ai), clinical risk-stratification and early-warning systems (for example, the Epic Sepsis Model), and language-model-based clinical applications (for example, ambient documentation systems such as Microsoft/Nuance DAX). The argument generalizes across these systems where they share the associative architecture analyzed in section 2; it is not intended as a uniform claim about every deployed system.

Pitt Open Library Publishing (cc) BY

This work is published by Pitt Open Library Publishing and is licensed under a Creative Commons Attribution 4.0 International License. © The Author(s).

ISSN 2692-3963 | Date Received 01-03-2026 | Date Accepted 24-09-2026 | Date First Published 09-11-2026
Philosophy of Medicine | DOI 10.5195/pom.2026.288 | Volume 7 | Issue 1 | pp. 1–33

By "anthropological event," this article means a transformation in the fundamental practices and self-understanding of clinical practitioners, a change not merely in what tools they use but in how they perceive patients, exercise judgment, and understand their professional identity. The integration of AI constitutes such an event because it reshapes the epistemic and relational foundations upon which clinical medicine rests.

This article examines the philosophical and anthropological dimensions of this transformation. It argues that the integration of AI into medicine produces what may be termed an "algorithmic gaze," a mode of clinical perception that, unlike the physician's gaze described by Michel Foucault (1973), operates on data rather than bodies, correlations rather than causes, and probabilities rather than narratives. Where an encounter turns on causal-mechanistic or contextual reasoning, the consequences are threefold: the patient is reconstructed as a statistical aggregate, the clinician is placed before an epistemic authority they cannot interrogate, and the relational dimension of care (grounded in shared embodiment and vulnerability) comes under pressure. Each of these claims is qualified in the section that establishes it, and section 2.3 marks the encounters to which the argument is addressed.

The concept of the algorithmic gaze builds on but extends beyond Yuxin Dai and Sizhe Guo's (2025) recent Foucauldian critique of AI in clinical diagnosis. Dai and Guo identify three structural paradoxes (accountability, empathy, and care) that establish the ethical-relational case for revisiting algorithmic clinical practice; section 5.2 develops their analysis substantively and specifies where this article agrees, diverges, and adds. The extensions are threefold. Epistemologically, the article incorporates Judea Pearl's causal hierarchy (Pearl 2018) to show that the limitations Dai and Guo identify as ethical and relational are also structural-epistemic, arising from the associative ceiling of current AI architectures. Anthropologically, it develops a four-level consciousness taxonomy (section 3) and a structural-versus-empirical distinction (section 3.4), the latter preempting the parity charge their analysis is sometimes read as inviting. Practically, it connects the governance proposals to the embodied and politicized conception of care developed via Damásio (1994), Friston (2010), Tronto (1993, 2013), Kittay (1999), Mol (2008), and Puig de la Bellacasa (2017), articulated across sections 4 and 7. While Dai and Guo remain within ethical critique, this article demonstrates that the limitations of algorithmic medicine are not merely ethical (questions of power and autonomy) but epistemic (questions of what can be known) and ontological (questions of what kind of agents can be responsible).

Two arguments run through what follows and separating them at the outset is worthwhile because they are distinct and license distinct conclusions. The first concerns responsibility: Only an agent who can be held to account can fail in a way that counts morally, and AI systems are not agents of that kind (sections 3.4 and 4). The second concerns knowledge: Associative architectures cannot take up the narrative and contextual material on which some clinical encounters turn (sections 2 and 5). Both are gathered under the heading of the algorithmic gaze, and they meet at one premise, the semantic gap established in sections 3 and 4, which the knowledge argument draws on in section 5.3. They are otherwise independent. The governance proposals in section 7 rest for the most part on the knowledge argument; the institutional-design argument of section 7.3 draws on both, and the allocation of liability (section 6.1) follows from the responsibility argument alone. Where a conclusion is drawn below, the strand it rests on is named.

The importance of this analysis is not merely academic. The epistemic character of AI systems (what they know, how they know it, and what they cannot know) has direct consequences for clinical governance, medical education, and the regulation of healthcare technology. If AI-mediated medicine operates on a fundamentally different epistemology than the medicine it replaces, governance frameworks designed for the latter may be structurally inadequate for the former.

The methodology is one of critical conceptual analysis, synthesizing scholarship across philosophy of mind, medical anthropology, phenomenology, and care ethics to examine the foundational assumptions of clinical epistemology under conditions of algorithmic mediation. The article does not present empirical data; its contribution is theoretical, offering a unified philosophical framework for understanding the epistemic transformation wrought by clinical AI.

2. Associative Reasoning and the Ladder of Causation

Understanding the epistemic transformation clinical AI produces requires grasping how these systems reason, and how that reasoning differs from that of the clinicians they increasingly assist or supplant. The distinction is not merely technical: It determines what knowledge AI-mediated medicine can produce and, therefore, what clinical encounters it can support.

2.1 Pearl's Three Levels

Pearl's hierarchy of causal inference identifies three levels of cognitive ability (Pearl and Mackenzie 2018). Developed in Pearl's work on causal reasoning and popularized in *The Book of Why,* it provides an unusually precise vocabulary for machine learning's epistemic limitations.

The first level, association, is the domain of statistics and correlation. It answers questions such as: "What symptoms are associated with this disease?" or "Given observation X, how probable is Y?" The clinical AI currently deployed in healthcare (diagnostic imaging models, risk-stratification tools, and clinical-decision support systems) operates predominantly here. The DeepMind system for retinal disease diagnosis, for instance, identifies statistical associations between pixel-level features in optical coherence tomography scans and diagnostic categories, achieving performance comparable to world-leading clinical experts (De Fauw et al. 2018). However, these systems detect patterns without modeling the underlying generative mechanism. They are, as Pearl observes, sophisticated curve-fitting engines (Pearl and Mackenzie 2018, 30): They can identify that a configuration of retinal features is statistically associated with proliferative diabetic retinopathy but cannot articulate the biological mechanism (chronic hyperglycemia damaging the retinal microvasculature) connecting features to disease.

The second level, intervention, involves acting on the world and predicting the result: "What will happen if I administer this drug?" This requires a causal, not merely statistical, model. The distinction is subtle but critical. An associative system can predict that patients given drug A had better outcomes than those given drug B but cannot determine whether this is because drug A is more effective or because healthier patients were more likely to be prescribed it (confounding). Only a causal model supports the interventionist question: "If

I give this patient drug A, what will the outcome be?" Some AI systems act on their environments (reinforcement learning agents, for instance) but typically lack a generalizable model of intervention beyond their narrow training parameters.

The third level, counterfactual reasoning, is the summit of human cognition: "What would have happened if I had acted differently?" This mode (imagining a world that did not occur) is essential for clinical judgment. The experienced clinician reviewing a deteriorating patient asks not only, "What should I do now?" but also, "What would have happened had I intervened earlier?" Currently deployed AI systems lack the causal models needed for counterfactual reasoning in the robust, contextualized sense clinical practice requires (Bishop 2021). They cannot simulate alternative histories or evaluate the avoidability of an outcome. J. Mark Bishop's critique, rooted in John R. Searle's distinction between syntactic manipulation and semantic understanding, suggests that even if computational systems could formally execute counterfactual operations, these would remain empty of the meaning that grounds clinical counterfactual judgment (2021).

This characterization requires clarification. By "robust, contextualized counterfactual reasoning," I mean not formal Pearl-style do-calculation performed during a clinical encounter (sociologically rare in routine practice) but the clinician's capacity to ask and answer counterfactual questions when the encounter requires it. Much routine clinical work proceeds without explicit counterfactual reasoning: Established presentations are managed through tacit, socially constructed "mindlines" (Gabbay and Le May 2004) and recognition-primed pattern matching drawing on accumulated expertise (Eraut 2000). The contexts mobilizing counterfactual reasoning are specific: the diagnostic encounter that does not resolve under the routine pattern; the review of an adverse outcome, where the clinician asks what a different decision might have yielded; the treatment whose benefit must be weighed against what would happen in its absence; the supervision of a junior who must learn not only what to do but what to have considered. It is this capacity for contextualized counterfactual reasoning (mobilized when pattern matching cannot complete the work) that the present argument concerns, not its uniform exercise in every clinical encounter. The structural-versus-empirical distinction of section 3.4 applies here too: The capacity is structurally available to human clinicians and structurally unavailable to current AI; that individual clinicians do not always exercise it does not erase the asymmetry.

Recent work on causal representation learning is beginning to explore computational approaches to counterfactual reasoning (Schölkopf et al. 2021), but these remain largely experimental and uncharacteristic of deployed clinical AI. The argument here concerns the epistemic character of existing systems, not the theoretical limits of future architectures.

2.2 The Significance of the Associative Ceiling

Confining current clinical AI to the associative level has consequences beyond the individual diagnostic act: It shapes the kind of clinical knowledge AI-mediated medicine can produce.

At the associative level, knowledge takes the form of conditional probabilities: Given these inputs, what is the probability of this output? This is extremely powerful for classification: distinguishing malignant from benign lesions, identifying patients at high risk of sepsis, predicting which will deteriorate in the next six hours. Its power lies precisely in detecting patterns that are statistically reliable but cognitively inaccessible to human

observers: correlations across hundreds of variables, nonlinear interactions between features, and subtle distributional differences no human could compute.

Yet associative knowledge is intrinsically limited in three ways that matter for clinical epistemology. First, it cannot distinguish correlation from causation. That patients with feature set X tend to develop outcome Y does not establish that X causes Y; a third factor may cause both, or the association may be an artifact of the training data. This constrains AI for treatment decisions (as opposed to diagnostic classification): Knowing that a drug correlates with better outcomes is not knowing it causes them.

Second, associative knowledge cannot generalize beyond its training distribution. A system trained on one population may fail unpredictably in another, because the associations it learned may not hold in a different context. This "dataset shift" is a well-documented failure mode of clinical AI with no analogue in causal medical knowledge (Finlayson et al. 2021; Moreno-Torres et al. 2012). The clinician who understands a disease's causal mechanism can apply that understanding across populations; the AI that has learned a statistical pattern is trapped within the distribution that produced it.

Third, associative knowledge cannot support clinical reasoning that imagines alternatives. The clinician weighing surgery asks, "What will happen if I operate, and what will happen if I do not?": intervention-level reasoning (Pearl's Level 2). Reviewing an adverse outcome, she asks, "Would the patient have survived if I had acted differently?": counterfactual reasoning (Level 3). A system that computes only associations can answer neither.

Two contrasts make this limitation concrete. In the first, a model trained on structured electronic health-record data assigns a low-risk score to a patient whose presentation, in the training distribution, was associated with nonserious outcomes. But the encounter includes contextual features (a recent migration, a family history not yet documented, an inarticulate but emphatic patient concern) poorly represented in the training data that would, in a clinician's causal reasoning, alter the differential. The system has no mechanism to identify the relevant difference; it returns its low score with high confidence, and the clinician must decide whether to act on grounds the system cannot represent. In the second, a high-volume ophthalmic referral triage in a screening program runs a deep learning model on retinal images for signs of vision-threatening pathology. Here, the task is well characterized, the input distribution consistent, and the relevant pathological features precisely those the model has learned to detect; the system can match or exceed expert performance at scale (De Fauw et al. 2018), and the human contribution lies less in classification than in configuring the system within a clinical pathway. The gap between human and machine reasoning is not uniformly consequential: It matters where causal-mechanistic and contextual reasoning are required to complete what pattern matching cannot, and less where consistent associative classification is what the task itself demands. This article concerns the first kind of case, where the structural limits of associative reasoning are most directly at stake.

This does not claim causal AI is impossible in principle. The critique concerns clinical AI systems currently deployed in healthcare, which overwhelmingly operate at Pearl's associative level, and the philosophical consequences of their associative architecture. Recent work in causal representation learning (Schölkopf et al. 2021) and hybrid mechanistic-statistical models suggests future systems may implement genuine intervention-level reasoning. Even if realized, the deeper arguments of this article

(concerning embodied understanding, phenomenological access, and normative agency; sections 3–4) would remain: A causal AI that could answer the question "What will happen if I intervene?" would still lack the existential stake in the answer that grounds clinical care. The critique is therefore both contingent (directed at current deployed systems) and structural (directed at the gap between computation and care). Whether future systems achieve causal reasoning does not diminish the urgency of understanding what present systems cannot do.

2.3 The Epistemic Significance of Clinical Knowledge

The distinction between associative and causal reasoning is not merely philosophical: It determines what knowledge a clinical AI system can produce and what clinical encounters it can sustain. The question is not whether AI is everywhere inferior or superior to human clinical reasoning but where the boundary between associative-sufficient and causal-necessary tasks lies, and how a clinical system should be designed to honor it. The two cases in section 2.2 illustrate the poles. What follows concerns encounters requiring causal-mechanistic and contextual reasoning, not because they are the majority, but because in them clinical knowledge's epistemic character is most at stake.

An associative system can identify that certain retinal features correlate with cardiovascular risk, but it cannot explain why: It cannot articulate the mechanism linking retinal microvasculature to cardiac pathology. The clinician reasons causally: Hypertension damages small vessels systemically, so retinal changes mark systemic vascular disease. This causal understanding lets the clinician contextualize, qualify, and communicate the finding as the algorithm cannot. The clinician can tell the patient: "The changes in your eyes suggest that your blood pressure may be damaging blood vessels throughout your body, including in your heart. This is why I recommend…". The algorithm can only say, in effect: "Pixels in this configuration are statistically associated with cardiovascular events."

This epistemic difference matters for practice. If the AI's knowledge is associative, its recommendations are, from the clinician's perspective, oracular: reliable in aggregate but opaque in the particular case. The clinician must decide whether to trust a recommendation they cannot evaluate on its merits, a situation that, as section 5 argues, transforms the doctor-patient encounter in ways existing frameworks of clinical epistemology are ill-equipped to address. The encounter shifts from the clinician interpreting the patient's condition to the clinician translating the machine's prediction: Clinical knowledge's epistemological structure moves from interpretation to transmission.

3. A Taxonomy of Mental Attribution

The ontological gap between human and machine agency is often framed as binary (AI either "thinks" or it does not), a framing too imprecise for philosophical and clinical purposes. What matters for clinical epistemology is not whether AI "thinks" in some general sense but what specific capacities it possesses and lacks, and how those capacities (or their absence) affect the structure of clinical knowledge.

3.1 Four Levels

It is more productive to distinguish four levels of mental attribution, which law and clinical practice implicitly conflate for human agents; these come apart when applied to AI.

Phenomenal consciousness refers to subjective experience, the "what it is like" of Thomas Nagel's celebrated formulation (Nagel 1974). A clinician examining a patient has a subjective experience of the encounter: skin color, tissue texture, the tone of the patient's voice. The phenomenological tradition, from Edmund Husserl's analysis of intentional consciousness (Husserl [1913] 2014) through Maurice Merleau-Ponty's account of embodied perception (Merleau-Ponty [1945] 2013), holds this experiential dimension to be not incidental to clinical knowledge but constitutive of it. The experienced clinician's capacity to "read" a patient (recognizing subtle signs of distress, the incongruence between verbal report and physical presentation) depends on a perceptual engagement with the irreducibly phenomenal patient.

Intentional states refer to beliefs, desires, and goals attributable to an agent. The clinician believes the patient has pneumonia; desires a good outcome; intends to prescribe antibiotics. These states structure clinical reasoning: The diagnosis is a belief formed on evidence, and the treatment plan an intention directed toward a desired outcome: in Michael Bratman's (1987) terms, a future-directed intention embedded in a larger plan. Attributing intentional states to AI ("the algorithm believes the patient has cancer") is metaphorical at best, misleading at worst. The system has no beliefs in the sense Searle (1980) defines intentionality, mental states about something, directed at objects under aspects; it has probability distributions.

Normative agency refers to the capacity to recognize and respond to reasons, to understand that one ought to have acted differently. This capacity grounds moral and legal responsibility: The negligent clinician failed to respond to reasons they should have recognized. As T.M. Scanlon (2000) argues, moral responsibility requires the capacity to evaluate one's own actions against reasons for and against them, which Peter F. Strawson (1962) linked to the reactive attitudes (resentment, gratitude, indignation) that constitute our practices of holding one another accountable. Normative agency requires not merely computational competence but an understanding of the significance of one's actions, grasping that a misdiagnosis matters because a human being will suffer.

Legal attribution refers to the formal assignment of rights and duties by a legal system. This level operates by convention rather than ontology: A corporation can be a legal person without phenomenal consciousness, as established in *Salomon v. A. Salomon & Co. Ltd.* [1897] AC 22 (HL) and formalized across common law jurisdictions.

Human clinicians satisfy all four levels simultaneously, which is why clinical practice can treat diagnostic failure as both factual and moral failing. Current AI systems satisfy at most the fourth, and only by legal fiction. They lack phenomenal consciousness (no subjective experience of diagnosis), genuine intentional states (no beliefs about patient welfare), and normative agency (no capacity to recognize a duty of care as a reason for action).

This taxonomy is deliberately heterogeneous. Phenomenal consciousness and intentional states are categories of mental state; normative agency is a capacity for moral responsiveness; legal attribution is an institutional fact. The four do not form a stack, nor derive from a single ontological axis. They sit together in human clinicians because legal and clinical discourse treats personhood as a bundle, conflating ontological, capacity-based,

and institutional features for practical purposes. The taxonomy's purpose is to unbundle these and ask, of any candidate clinical agent, which of the four is satisfied and on what grounds. That permits the precise question pursued in the sections that follow: not whether AI "thinks" but which conditions of clinical responsibility AI does and does not meet.

3.2 The Dennettian Objection and Its Limits

Daniel C. Dennett has argued that "understanding" may be functional competence rather than inner phenomenal state, and that sufficiently complex symbol manipulation might constitute genuine comprehension (Dennett 2017). On the functionalist view, if a system's outputs are indistinguishable from genuine understanding, whether it "really" understands is merely terminological.

More recently, proponents point to emergent capabilities in large language models (LLMs) (chain-of-thought reasoning, zero-shot generalization, multi-step inference) that appear to exceed simple pattern matching (Wei et al. 2022). The challenge these capabilities pose to the taxonomy is real and must be addressed directly. Chain-of-thought prompting produces outputs superficially resembling causal reasoning, but the resemblance is structural, not semantic. LLMs perform learned sequential token prediction conditioned on corpora describing causal reasoning; they reproduce the linguistic form of causal inference without constructing causal models of the world (Marcos 2023). This matters clinically: A system generating "drug A caused the adverse event because it inhibits enzyme X" has produced a syntactically valid causal claim but predicted it from distributional regularities in text, rather than deriving it from a causal model. Jason Wei and colleagues (2022) document task-level performance thresholds, not the acquisition of causal reasoning in Pearl's sense. The boundary between sophisticated associative processing and genuine causal modeling remains analytically important even if empirically contested.

The empirical foundation of the emergent-capabilities argument has itself come under scrutiny. Rylan Schaeffer et al. (2023) show that apparent discontinuities in LLM performance characterized as "emergent abilities" are largely artifacts of nonlinear or discontinuous metric choices. Substituting linear, continuous metrics for the same task-model pairs flattens the discontinuities and performance improves smoothly with scale: precisely what one expects from sophisticated associative scaling rather than a qualitative shift to interventional or counterfactual reasoning. Their meta-analysis of BIG-Bench results found more than 92 percent of claimed emergent abilities appeared under only two metrics, and they induced previously unobserved "emergent" abilities in vision tasks by selecting appropriate metrics. The argument here does not depend on this finding; even if some emergent capabilities are genuine, the structural distinction between metric-improved pattern recognition and acquired causal architecture remains. But it tightens the case that current LLM "reasoning" is associative scaling, not the acquisition of higher-level cognitive structure.

But the deeper issue is epistemological, not purely computational. Even if an AI system were behaviorally indistinguishable from a clinician (same outputs, same diagnostic calls, apparent causal reasoning), clinical responsibility requires more than behavioral equivalence. It requires what may be called an *existential stake* in the outcome: a phenomenological structure within which clinical outcomes register as morally significant to the agent, such that a diagnostic error is not merely a parameter to update but an event

that matters. This is not to claim every empirical clinician experiences this stake in every encounter. Burnout, time pressure, and routinization may flatten the felt weight of clinical decisions, and the literature on moral injury and physician disengagement is evidence that the stake is not uniformly present in clinical work (Dean et al. 2019). The claim is rather that the *capacity* for existential stake (and the biological and social conditions that make it intelligible) is structurally available to human clinicians and structurally unavailable to AI systems. The clinician who has stopped caring can be held responsible for having stopped, in a way that an AI system that has never been the kind of thing that could care cannot. This is the same structural-versus-empirical distinction developed in section 3.4. The condition that survives even Harry Frankfurt's ([1969] 2018) rejection of alternate possibilities as a requirement of moral responsibility (that the agent acts on reasons of her own and can recognize that she ought to have acted differently) is not erased by the empirical fact that some agents fail to honor it. Without this normative responsiveness as a structural possibility, even flawless behavioral performance leaves the essential problem of clinical epistemology unresolved.

This article does not seek to resolve this philosophical debate definitively. Even on the most generous functionalist interpretation, the practical conditions for clinical responsibility require normative agency in a thick sense: The clinician must grasp reasons as reasons, experience the moral weight of a diagnostic decision, and be held accountable within a web of reactive attitudes (Strawson 1962). Current AI systems lack these capacities. Whether a future system might satisfy them is a separate question; the argument here concerns the clinical AI systems that exist and are being deployed.

One might object that normative responsiveness could be socially scaffolded (that institutions could create a functional equivalent of normative agency). This article acknowledges social scaffolding's role in cultivating human normative agency but observes that such scaffolding presupposes a phenomenologically responsive agent: one who can experience professional obligation, feel a duty of care, and suffer moral injury when that duty is breached. The scaffold supports responsiveness; it does not create it.

3.3 The Clinical Significance of the Taxonomy

The taxonomy is not mere classification but has direct implications for clinical epistemology and governance: Its consequences are *normative,* specifying what responsible clinical practice requires and what is foreclosed when that requirement fails, not empirical predictions about clinician or AI behavior (see section 3.4).

The absence of phenomenal consciousness means AI cannot perceive the patient phenomenologically: The clinician who notes a patient "looks unwell" draws on a perceptual capacity rooted in shared embodiment: uneven, susceptible to fatigue, routinization, and cultural and cognitive bias, but structurally available to humans and structurally unavailable to algorithms. The absence of intentional states means AI cannot form genuine clinical beliefs: The clinician's diagnosis is a belief embedded in a network of pathophysiology, prognosis, and treatment; the AI's "probability of pneumonia: 0.87" is a computation, not a belief, with no implications beyond the number. The absence of normative agency means AI cannot be held morally responsible for its outputs: The clinician's awareness that errors cause suffering, for which they bear moral and professional consequences, normatively grounds the epistemic virtues (caution, thoroughness, humility)

good clinical practice requires. That individual clinicians may fail to cultivate or sustain them does not undermine the claim; the failure is itself morally evaluable, precisely because the condition is available to humans and structurally absent in AI.

3.4 Structural Capacity and Empirical Exercise

A reasonable objection to the taxonomy in sections 3.1–3.3 is that it contrasts an idealized clinician (one with phenomenal consciousness, genuine intentional states, and normative agency in the thick sense) with an empirical AI system, an unfair comparison since actual clinicians do not always meet these conditions. The diagnostic-error literature documents pervasive cognitive biases (anchoring, premature closure, availability, confirmation bias) producing diagnostic inaccuracies in 36.5–77 percent of studied case scenarios (Saposnik et al. 2016; Croskerry [2003] catalogues at least 30 such cognitive dispositions to respond). Sociological work on clinical knowledge shows much routine practice proceeds through tacit, socially constructed "mindlines" rather than explicit counterfactual reasoning (Gabbay and Le May 2004; Eraut 2000). The burnt-out physician who treats patients as technical problems, the time-pressured generalist who follows heuristic over inference, the experienced clinician whose intuition has hardened into bias: these are recurrent features of practice, not edge cases. Were the article's comparison one between idealized humans and empirical machines, it would indeed be vulnerable to a charge of false parity.

The comparison is not of that kind. The four-level taxonomy makes no empirical claim about any individual clinician's performance relative to any individual AI system; it claims that certain *structural conditions* determine whether an agent can be held morally responsible for clinical judgment. Phenomenal consciousness, intentional states in Searle's (1980) thick sense, and normative agency in Strawson's (1962) sense are conditions of possibility for clinical responsibility: what an agent must possess in principle to be the kind of being that can fail in a morally evaluable way. These are structural prerequisites, not empirical achievements, and empirical failure presupposes structural capacity. The clinician who misses a diagnosis through anchoring bias can be held morally responsible (indeed, this is what makes professional accountability intelligible) precisely because the capacity to have reasoned otherwise was structurally available, even if unexercised in that instance. An AI system producing the same error cannot be held morally responsible in this sense, because that capacity, in any morally significant way, is constitutively unavailable to it. The asymmetry is thus not a claim that actual clinicians outperform actual AI systems (the empirical record on cognitive bias makes such a claim manifestly false in many domains, particularly consistent high-volume associative classification, where AI now matches or exceeds expert performance) but a claim that clinicians occupy a normative position they can fail to honor, whereas AI systems cannot occupy that position at all.

This distinction has practical consequences for the article's argument. The governance proposals in section 7 do not assume human clinical reasoning is uniformly superior to algorithmic processing; they rest on the more defensible claim that responsible clinical practice requires a particular structural ground: one that makes professional accountability and moral repair possible and cannot be relocated to a computational substrate by any extension of current AI architectures. Why this structural ground is nontrivial (why no sufficiently elaborate functional analogue can replace it, and why it arises specifically from

the biological and social conditions of being a mortal creature among mortal creatures) is the subject of section 4.

4. Biological Homeostasis and the Roots of Care

4.1 The Somatic Marker Hypothesis

Why do clinicians care about their patients? Why does a diagnostic error produce anxiety, guilt, or distress? These are not trivial questions: They concern the motivational structure underlying clinical judgment and the ethical foundations of the therapeutic relationship.

For Damásio (1994), the answer is homeostasis: the biological imperative to maintain a stable internal environment for survival. His somatic marker hypothesis holds that human rationality is not purely logical but guided by emotional signals from the body: Weighing a risky treatment decision, a clinician's body generates a "gut feeling" (a somatic marker based on past experiences of similar situations) that prunes the decision tree, steering them from negative outcomes before conscious calculation occurs (Damásio 1996).

This is neuroscience, not mysticism. Damásio's clinical work with patients with damage to the ventromedial prefrontal cortex showed that those who lose the capacity for somatic marking (who reason abstractly but cannot feel the emotional weight of their decisions) make catastrophically poor real-world choices despite normal intellectual function. The implication for clinical epistemology is significant: the experienced clinician's "clinical intuition" is not an ineffable gift but a somatic process, the body's accumulated record of diagnostic encounters, encoded in visceral and emotional memory, informing judgment in ways formal reasoning alone cannot replicate.

4.2 Free Energy, Existential Stakes, and the Grounding of Care

Karl Friston's free energy principle extends this analysis from individual decision-making to the fundamental structure of biological agency: Biological agents act to minimize "free energy" (the discrepancy between an organism's predictions and its sensory input) to ensure their continued existence, giving them an existential stake in the accuracy of those predictions because their survival depends on it (Friston 2010). A clinician who misdiagnoses a dangerous condition fails not merely technically but in a way that reverberates through their embodied existence, knowing they could have acted otherwise: They may lose sleep, lose confidence, or suffer the moral injury that Wendy Dean et al. (2019) describe as a defining feature of clinical distress.

AI systems, by contrast, are disembodied in the clinically relevant sense, though this requires qualification. The free energy principle is mathematically substrate independent (active inference models have been applied to artificial agents, including robots and computational simulations [Friston et al. 2024; Matsumura et al. 2024]) and a system minimizing variational free energy can in principle exhibit goal-directed behavior and adaptive control without biological tissue. This article does not deny that possibility; its claim is that the existential stake grounding clinical care (the clinician's embodied vulnerability, mortality, and capacity for moral injury) is constituted not by computational homeostasis alone but by the biological and social conditions of being a mortal creature

among mortal creatures. Minimizing a loss function, or even variational free energy, does not thereby give an AI system what Strawson (1962) calls the reactive attitudes (resentment, gratitude, guilt) that constitute our practices of holding one another accountable and motivate the clinician's commitment to the patient's welfare.

This asymmetry needs precise statement. The claim is not that the mathematics of active inference is unavailable to artificial systems; it is, as Friston et al. (2024) themselves acknowledge, substrate independent. The claim concerns the relationship between the formalism and what it is solving for. A biological organism does not minimize variational free energy *as an end in itself* but as a condition of remaining the kind of organism that has anything to lose: mortality, vulnerability to suffering, and the social exposure of being a creature among other creatures are not parameters in the equation but the conditions under which it has any motivational force at all. Friston et al.'s (2024) own staged-development account makes this visible: Current AI systems sit at what they label S0, systemic intelligence (general-purpose optimization), while the capacities clinical reasoning draws on (sentience [S1], counterfactual sophistication [S2], theory-of-mind [S3]) are developmental targets, not achieved properties of existing architectures. Tadayuki Matsumura et al.'s (2024, 280) "empathic active inference," which produces socially calibrated robotic behavior through a reward term encoding others' expectations, is technically impressive but illustrates the same point: The empathy is a function in the expected free energy, not a stance the agent occupies. The Damasian argument, then, is not that the formalism is biologically restricted but that the *significance* it captures (outcomes registering as mattering) presupposes a creature for whom they matter, which an artificial agent solving the same equations is not.

This move presupposes a philosophical commitment not all readers share. Dennett's (2017) functionalism would deny that mortality is constitutive of meaning: On that view, meaning is the role a formalism plays, and biological vulnerability is one realization of that role rather than its necessary condition. The argument here does not claim to settle that debate. It claims, more modestly, that if one accepts the Strawsonian framework of reactive attitudes as constitutive of clinical accountability (Strawson 1962; see section 3.2), the conditions under which those attitudes can be felt (exposure to suffering as a creature among other creatures, the moral injury of breaching a duty of care) are not optional add-ons to the formalism but the conditions under which it has any clinical force at all. The substrate-independence concession is therefore conditional on the contested status of functionalism in this domain, not a refutation of it.

The argument from embodiment is not a metaphysical claim about what computation cannot in principle achieve; it is a claim about what currently grounds clinical care as a normative practice: shared vulnerability, phenomenal access to suffering, and the capacity for moral injury. An AI that misclassifies a tumor registers a loss function (a mathematical penalty that adjusts its parameters) but experiences no pain, regret, or fear of death.

4.3 Two Consequences: Value Alignment and the Ethics of Care

The first consequence concerns value alignment. Lacking biological imperatives, AI must have its "values" explicitly programmed as objective functions. This creates a risk (the structural point applies equally to narrow clinical AI): An AI might pursue a goal (such as "maximize patient throughput") to a pathological extreme (such as premature discharge)

because it lacks the commonsense, homeostatic check that harming humans is intrinsically undesirable (Bostrom 2014). The problem is not that AI has the wrong values but that it has none at all, only optimization targets. The gap between an optimization target and a genuine value is the gap between syntax and semantics, between computation and care.

The second consequence concerns the ethics of care. In healthcare, "care" is not merely a service, but a relational act grounded in shared vulnerability. Nel Noddings (2013), Edmund D. Pellegrino and David C. Thomasma (1993), and Shannon Vallor (2011) hold that the clinician's obligation arises from the asymmetry of vulnerability between patient and professional: The patient is exposed (physically, emotionally, existentially) in ways that create a moral claim on the clinician. The clinician's homeostatic body makes this asymmetry intelligible: The doctor can recognize the patient's suffering as suffering because they share the capacity for it: They understand pain because they have been in pain, and the fear of death because they are mortal.

Two further developments of the care-ethics tradition are central here. First, Joan C. Tronto (1993, 2013) shows that care is not a single act but a chain of distinct phases: caring-about (recognizing that care is needed), caring-for (taking responsibility for the response), care-giving (the direct work of care), and care-receiving (the recipient's response, which itself completes the act). This matters because the algorithmic gaze redistributes the phases unevenly: AI systems can be configured to perform aspects of caring-about (population-level risk identification) and can support care-giving (workflow automation, decision support) but cannot occupy the position from which caring-for is exercised (the assumption of responsibility for a particular other), nor complete care-receiving as a moral relation. Tronto's political reformulation, which insists that care is not a private virtue but a structural feature of just institutions, bears directly on section 7: The institutional design proposals there must protect the conditions under which all four phases of care can be sustained, not only those the algorithm performs well.

Second, Eva Feder Kittay (1999) argues that dependency is species typical rather than exceptional or pathological. This grounds the present article's appeal to shared vulnerability in a politicized conception of human life: Clinicians and patients are not autonomous agents who happen to meet in the clinical encounter but mutually dependent creatures whose interdependence is the condition of social existence. Against any purely behaviorist reading of care, Kittay also insists that good caring "seems not only to involve the overt actions but also an attitude, a sentiment, or an emotion that accompanies behavior" (Kittay 2002, 244). This distinction is decisive for the argument that follows: An AI system can be designed to perform care-actions (Pepper's reassuring gesture [Rincon et al. 2018], the conversational system's empathic phrasing) but cannot, by virtue of its constitution, possess the care-attitude that makes the action genuinely caring. The synthetic empathy of clinical AI is the appearance of care decoupled from its motivational ground.

An AI system can simulate the expressions of care ("synthetic empathy") but cannot care in the Damasian sense, because it has no self to preserve and, in Kittay's terms, no attitude that accompanies the behavior. The care robot that comforts a distressed patient with a preprogrammed gesture of reassurance performs an action that looks like care but is decoupled from the motivational structure (the embodied concern for a fellow creature, the species-typical recognition of dependency) that gives care its moral weight. The synthetic gesture is care-as-action without care-as-relation; it satisfies the behavioral criterion while leaving the relational and political criteria unmet.

This relational and attitudinal conception of care has institutional implications. Annemarie Mol (2008) argues that the "logic of care" (improvisational, ongoing, responsive to the particular patient) differs structurally from the "logic of choice," under which care is reconfigured as a service to be selected. Algorithmic clinical systems, in their current deployment, intensify the logic of choice: They generate options, scores, and recommendations from which the patient or clinician selects. María Puig de la Bellacasa extends this by insisting that care is "everything that is done … to maintain, continue, and repair the world so that all can live in it as well as possible" (2017, 161), a definition that extends care beyond the human-human dyad to the infrastructural and material conditions under which clinical work occurs. Together, these point toward the institutional analysis in section 7.3: The question is not only how individual clinicians can preserve relational care under AI mediation, but how institutional design can protect the logics, attitudes, and infrastructural commitments that make care possible at all. Recent work in critical posthuman nursing (Smith and Willis 2020; Hopkins-Walsh et al. 2022) develops this institutional and infrastructural framing further by tracing how clinical care is always already an assemblage of human, technological, and material elements, a framing this article does not fully adopt (see section 9) but takes seriously as an indication that the human/machine binary on which the argument partly relies is itself contestable.

4.4 The Searle Problem and the Syntax-Semantics Gap

The argument from embodiment raises a deeper point about understanding. Searle's "Chinese Room" argument (Searle 1980, 417–418) holds that a system manipulating symbols by formal rules, however complex, does not thereby understand their meaning. The debate is extensive in the philosophy of mind: The systems reply holds that understanding belongs to the whole system rather than any component, and functionalist critics argue more broadly that understanding is a matter of functional organization, not substrate (Dennett 2017). Whatever one's verdict on the Chinese Room in the abstract, the clinical context offers a distinctive test case that illuminates the limits of the functionalist position.

Consider a clinical AI system that correctly identifies a malignant tumor from a histopathology slide: It processes pixel data, computes features, and outputs a classification. A functionalist might say the system "understands" the image in the relevant sense, since it extracts the diagnostically relevant information and produces the correct output. But the clinician examining the same slide understands more: what the diagnosis means—this patient has cancer; their life may be shortened; they face difficult treatment decisions. This is not a computational addition to the classification but a different kind of engagement with the same data, grounded in the clinician's own mortality, empathy, and awareness of the human stakes involved. It is partial, situated, and susceptible to bias but structurally available to the clinician in a way it is not to any computational process of the kind now deployed.

The syntax-semantics gap is therefore not merely a theoretical curiosity but the gap between computation and care. A system that processes symbols without understanding their meaning can classify but cannot care, in the sense developed in section 3.4: The capacity for care is structurally unavailable to it; it can compute but cannot comprehend. Where the consequences of classification are human suffering and death, this gap is not merely philosophically interesting but clinically consequential.

Table 1 summarizes the contrasts developed in sections 2–4. It compares the human clinician, current clinical AI, and a theoretical active inference agent across six dimensions: cognitive basis, primary drive, embodiment, semantics, characteristic errors, and ethical basis. The first row restates each agent's position on Pearl's ladder of causation (section 2); the rows on primary drive, embodiment, and ethical basis condense the argument from biological homeostasis, free energy minimization, and the ethics of care (sections 4.1–4.3); and the semantics row restates the syntax-semantics gap discussed above. The third column is included because active inference offers the most developed formal account of how an artificial agent might approximate biological self-regulation; as section 4.2 argued, however, an agent that met its formal specification would solve the same equations without the existential stakes that give them motivational force for a living clinician.

Table 1: Comparative ontology of intelligent agents

Dimension	Human clinician	Current AI	Active inference agent (theoretical)
Cognitive basis	Causal, counterfactual, abductive (Pearl Levels 2–3)	Associative, correlative, inductive (Pearl Level 1)	Causal inference, Bayesian updating
Primary drive	Biological homeostasis and social survival	Loss function optimization	Free energy minimization
Embodiment	Fully embodied (somatic markers)	Disembodied (server based)	Simulated embodiment (Markov blankets, Friston et al. 2021)
Semantics	Situated meaning-making grounded in mortality and social context	Formal symbol manipulation (under classical computationalism, syntactic)	Generative-model inference (semantically constrained by the model)
Error characteristic	Cognitive bias, fatigue, knowledge gap	Hallucination, overfitting, adversarial fragility	False inference, model mismatch
Ethical basis	Empathy, shared vulnerability, social contract	Value alignment (reinforcement learning from human feedback)	Epistemic value (curiosity)

5. The Algorithmic Gaze: From Anatomy to Algorithm

5.1 Foucault and the Medical Gaze

In *The Birth of the Clinic* (1973), Foucault described the shift from "bedside medicine," in which the physician listened to the patient's narrative, to "hospital medicine," in which the physician examined the patient's anatomy. This was the birth of the medical gaze (*le regard médical*), which separated the patient's body from their personhood to reveal the "truth" of the disease (Foucault 1973, 107–123). The reorganization of clinical space (the hospital

ward, the anatomy theatre, the classificatory schema) created a new regime of visibility in which certain phenomena (lesions, anatomical correlations) became epistemically admissible while others (the patient's narrative of illness, their subjective experience) were rendered peripheral (146–151).

Foucault's analysis was genealogical, not merely descriptive: The medical gaze was not a natural or inevitable way of seeing but a historically specific configuration of knowledge and power that emerged in the reorganization of European medicine in the late eighteenth and early nineteenth centuries, making anatomical pathology and lesion-symptom correlations visible while rendering the patient's experience, social context, and narrative invisible.

The Foucauldian gaze was constituted around visible objects (the lesion, the palpable organ, the percussed chest), but its epistemic structure was not perceptual in any straightforward sense: What could be said about what was seen, and which seen things could be admitted as evidence at all, was determined by the discursive formation of pathological anatomy that organized clinical perception. The body was the object of examination, but the discursive apparatus that rendered its interior legible (lesions classified, pathologies named) was itself the condition of the gaze: a structural production rather than a natural achievement of trained sight. This article takes that already-structural character as its starting point.

5.2 The Algorithmic Gaze in a Genealogy of Medical Visibilities

The term "algorithmic gaze" was brought into the critique of clinical diagnosis by Dai and Guo (2025), who deploy it in a Foucauldian critique of AI in clinical diagnosis identifying three structural paradoxes. The accountability paradox: Algorithmic opacity transforms the meaning of consent and contestation, leaving the dissenting patient no surface against which to articulate dissent and the clinician no reasoning structure to contest. The empathy paradox: The "calculated empathy" of conversational AI, a statistically optimized performance of caring discourse, domesticates the language of suffering without engaging it existentially. The care paradox: Splitting emotional labor (increasingly delegated to AI) from technical intervention (retained by physicians) fragments the holistic therapeutic relationship in which trust has historically been constituted. Their constructive directions (enforceable accountability, reallocating automation-freed time to dialogical engagement, participatory data governance, and ethics embedded in algorithmic design from inception) are commensurate with those conditions.

This article shares the diagnostic framing but extends it in three respects Dai and Guo's ethical critique does not pursue. First, it grounds the analysis in Pearl's causal hierarchy (sections 2–3), so limitations they cast as ethical and relational appear also as structural-epistemic, arising from the associative ceiling of current AI architectures. Second, it develops the consciousness taxonomy of section 3 and the structural-versus-empirical distinction of section 3.4, which disambiguate what the legal and clinical framing of "AI as agent" systematically conflates and preempt the parity charge their analysis is sometimes read as inviting. Third, it grounds the relational critique in the embodied-care argument of section 4, particularly the Tronto-Kittay framing of dependency and care-as-attitude that connects clinical care to institutional design.

This article thus uses the term in a specific sense and genealogy. The algorithmic gaze is not a rupture from Foucault's medical gaze but its latest moment: a regime of visibility produced by datafication and model inference, in which the "truth" of the patient's condition is constituted by computational pattern-recognition rather than perceptual encounter. Just as late eighteenth-century anatomical medicine intensified and reconfigured an earlier symptomatology by reorganizing the hospital, the classificatory schema, and the anatomy theatre, the algorithmic gaze intensifies and reconfigures the anatomical gaze by reorganizing the database, the model, and the deployed inference pipeline. The discursive formations conferring authority on clinical knowledge (biomedical markers, statistical inference, hierarchies of evidence, populations conceived as comparable distributions) are inherited rather than overturned; what is new is the technique of inscription and inference applied to them. Rosi Braidotti's (2013, 2019) analyses of cognitive capitalism and the Fourth Industrial Revolution situate this transformation politically and economically: The algorithmic gaze is the most recent crystallization of a long-running conversion of clinical knowledge into computable form, accelerated by the architectures of contemporary machine learning and the data infrastructures that sustain them. The novelty claim is accordingly not that a wholly new kind of clinical knowledge has appeared but that the genealogy has reached a moment in which specific epistemic features make a categorial difference.

Five features make the difference:

- *Opacity to expert inspection:* Where the anatomical gaze was opaque to the patient but legible to the trained clinician, the algorithmic gaze is opaque to both. The deep-learning model identifies patterns (nonlinear correlations across thousands of variables) not cognitively recoverable by any expert, because the seeing is not constituted in any perceptually accessible space.
- *Dependence on training distributions:* The gaze sees what its corpus permits, and a model trained on one population's data may misclassify or fail unpredictably on another in ways no clinical-mechanistic account would have predicted (Finlayson et al. 2021).
- *Occlusion of the narrative register:* The algorithmic gaze admits as evidence only what has been pre-quantified or reducible to feature vectors; narrative, experiential, and contextual material is either filtered out at data collection or compressed into representations retaining correlational structure without experiential meaning (the three-mechanism account is developed at section 5.3).
- *Substitution of statistical inference for embodied perception:* The clinician's perception of distress in a fellow creature is replaced, in the algorithmic register, by inference over signals correlated with prior labels of distress.
- *Semantic noncomprehension:* Even at perfect classification accuracy, the algorithmic gaze does not understand what it sees in the sense of sections 3 and 4; it does not grasp the diagnosis as significant for a person whose life is implicated by it.

No single feature is wholly novel; the categorial difference lies in their conjunction at scale and under expert opacity. Several of these features are developed further in the sections that follow; others have been treated in sections 3 and 4 and are returned to in section 7. Together, they mark the algorithmic gaze as discursively continuous with the medical gaze and epistemically distinct from it.

The previous regime located clinical authority in a discursive formation the trained clinician could occupy: Pathology classifications could be learned, examined, and contested by experts. The algorithmic gaze relocates that authority to a computational process no human (not the clinician, not the developer, not the patient) can fully inspect. That authority is not self-authorized by the algorithm; it is ascribed by institutions, regulators, payers, and clinicians themselves, who treat the machine's output as evidence under the same hierarchies of evidence that authorize other clinical knowledge. Displacing the clinician from the epistemic center of the encounter is thus an institutional and discursive act as well as a technical one, examined in turn by the next sections, particularly the rituals of human-in-the-loop sign-off discussed in section 6.

5.3 Subject Occlusion and the Loss of the Narrative

The anthropological consequence of the algorithmic gaze is subject occlusion. The patient's subjective experience (their pain, history, fear, values) is filtered out as "noise" if it cannot be quantified. The algorithm enforces a normative homogenization, treating the patient as a statistical aggregate rather than a singular individual (Dai and Guo 2025; Peerson 1995). What resists quantification is not merely ignored but epistemically excluded from clinically relevant knowledge.

Posed sharply, this may seem to overstate the discontinuity. Displacing narrative and experiential knowledge from epistemic admissibility is not novel to AI-mediated medicine; it has been contested in clinical practice for at least half a century, since the codification of evidence-based medicine (EBM) and the spread of structured templates, point-of-care prompts, and computerized decision support. Trisha Greenhalgh et al. (2014) document how "well intentioned efforts to automate use of evidence through computerised decision support systems, structured templates, and point of care prompts can crowd out the local, individualised, and patient initiated elements of the clinical consultation," and how inexperienced clinicians may engage with such systems "mechanically and defensively," stifling more nuanced clinical expertise. The algorithm completes what EBM began but does so under epistemic conditions that EBM's most thoughtful proponents had already warned against. AI's distinctive feature is therefore not introducing quantificationism into a previously holistic practice (medicine had begun that conversion decades earlier) but intensifying the quantification tendency while progressively removing the corrective check. Under the algorithmic gaze, the clinician who would have overridden the quantified score with narrative judgment is epistemically subordinated to a process that cannot, in principle, register what the override would have responded to. The distinction worth drawing is between quantification-as-translation and quantification-as-substitution. Narrative-based instruments such as the McGill pain questionnaire are quantifications-as-translation: They retain the patient's narrative content as input. The algorithmic risk score is quantification-as-substitution: generated without reference to the patient's narrative articulation, with the encounter then structured around the score rather than the score around the encounter.

This exclusion is not neutral. The features that resist quantification (the patient's lived experience of illness, their cultural understanding of disease, their relational context, their values and preferences) are precisely those that clinical practice has traditionally considered most important for holistic care. The phenomenological tradition in medicine, from Karl Jaspers ([1913] 1997) through S. Kay Toombs (1992), has argued that the patient's

experience of illness (*Erlebnis*) is not merely a subjective overlay on objective disease but a constitutive dimension of the clinical reality the physician must engage with.

Three distinct mechanisms of subject occlusion should be distinguished:

- *Data-level exclusion:* The structured datasets on which clinical AI systems are trained systematically omit narrative, experiential, and contextual information that resists quantification.
- *Representational reduction:* Even when narrative data is included (as in multimodal systems processing clinical notes), it is reduced to feature vectors that capture correlational patterns but not experiential meaning.
- *Semantic noncomprehension:* Even a system with complete patient data, processed through the most sophisticated algorithms, cannot understand what the patient's experience means in the phenomenological sense discussed in section 3.

These levels are cumulative: Current systems typically exhibit all three, but even an ideal future system that eliminated data-level exclusion would remain constrained by representational reduction and semantic noncomprehension.

Consider a 58-year-old woman presenting with persistent fatigue and diffuse abdominal discomfort. An AI triage system, trained on structured data (blood panels, imaging results, demographic variables), assigns her a low-risk score and recommends routine monitoring. It has no access to the qualitative narrative a clinician would elicit: that her mother died of ovarian cancer at 62; that she has night sweats she considers too trivial to report; and that she describes a feeling she can articulate only as "something is wrong". These phenomenological cues (experiential, embodied, and resistant to quantification) are precisely the clinical information the algorithmic gaze occludes.

The scenario does not demonstrate that AI is inferior to human judgment in all cases. Rather, it illustrates the epistemic loss that occurs when the patient is reconstructed as a data vector and the clinician's interpretive role is subordinated to algorithmic output. The patient's narrative (which contains clinically relevant information structured data cannot capture) is not merely supplementary to the algorithmic analysis; it is a different kind of knowledge, operating in a different epistemic register that the algorithmic gaze cannot access.

A potential objection emerges: As AI systems become multimodal, incorporating narrative transcripts, audio recordings, and unstructured clinical notes alongside structured data, will this mitigate data-level occlusion? Future systems processing the patient's spoken account of their symptoms, the temporal pattern of their presentations, and verbatim notes of previous clinicians might seem to restore access to the phenomenological material current triage systems omit. But this would conflate data access with epistemological access. The gap is not merely about input breadth but about the kind of knowledge that input supports. Even a system processing complete narrative data (the patient's full testimony, recorded in their own voice, contextualized by years of clinical notes) remains constrained at Pearl's Level 1, the associative level. It can correlate narrative features (words, patterns, linguistic markers of distress or urgency) with diagnostic outcomes; it cannot thereby understand the phenomenological content of the narrative. The system performs pattern matching on narrative text, not comprehension of experiential meaning. The patient who says "something is wrong" is articulating an embodied, existential unease, a felt sense that something fundamental has shifted in their relationship to their own body, a pre-reflective awareness that consciousness itself has changed. An AI

system correlating it with delayed-diagnosis cases has not understood what "something is wrong" means in the way a clinician can: as an expression of vulnerability, of a person recognizing that they are no longer in secure possession of their own body. Subject occlusion is therefore not merely a data-access problem but an epistemic one, rooted in the structural gap between processing narrative data (however rich and multimodal) and understanding experiential meaning.

A final case shows what is at stake infrastructurally. In engineering, a "digital twin" is a dynamic computational model of a physical system, continuously updated with real-time sensor data so as to mirror the system's state and behavior for prediction, optimization, and decision support (Grieves and Vickers 2017); Eric J. Topol (2019) and others have advocated extending it to medicine, giving each patient a twin that integrates genomic data, clinical history, imaging, and real-time physiological monitoring that would enable personalized prediction and treatment. An aircraft engine can be modeled faithfully because it is finite, deterministic at the relevant scale, and fully instrumentable. A patient is none of these things: biological processes only partially understood, psychosocial factors that resist quantification, and subjective experience that is in principle inaccessible to third-person measurement. The twin is therefore a selective representation rather than a faithful model, and the occlusion described above becomes infrastructural: What the twin excludes as unquantifiable is rendered clinically invisible, since clinical AI can act only on what the twin includes.

5.4 The Transformation of Clinical Knowledge

These epistemic constraints affect the structure and transmission of clinical knowledge. The claims below about educational atrophy are not claims about present states but risk hypotheses extrapolated from the epistemic architecture analyzed above, offered as grounds for precautionary governance.

Medical education teaches causal reasoning, connecting observed signs to underlying mechanisms so that the clinician can reason beyond standard presentations. AI-mediated diagnosis inverts this pathway: It proceeds from statistical association to classification without the causal understanding that clinical reasoning demands. The AI does not know why these features associate with this diagnosis, only that they do. Its classification may be more accurate than the clinician's but is epistemically impoverished, lacking the causal depth that enables explanation and generalization.

This inversion affects the transmission of clinical knowledge across generations. If clinicians increasingly defer to algorithmic classification, the causal understanding that historically constituted clinical expertise risks gradual atrophy (a risk hypothesis warranting investigation). The phenomenon parallels aviation's "automation complacency" (Parasuraman and Manzey 2010): As pilots rely more on automated systems, flying ability deteriorates. The clinical analogue: doctors who can operate AI tools but cannot diagnose without them.

Studies of computer-aided detection in radiology document changes in visual search patterns under algorithmic assistance (Tourassi et al. 2013), and the aviation parallel is suggestive. Neither, alone, establishes that clinical deskilling under AI will follow these trajectories, and the radiology evidence especially has been read in multiple ways. The argument therefore does not rest on that empirical prediction but on the epistemic claim

that the conditions cultivating causal-mechanistic reasoning, embodied perception, and dissent may not be reproducible under progressive algorithmic mediation. The governance proposals in section 7 draw their precautionary force from that epistemic claim regardless of whether deskilling, as empirically documented, ultimately materializes.

6. The Ritual of the Human in the Loop and the Uncanny Valley of Care

6.1 Human in the Loop as Ritual

Legal and regulatory frameworks, including the European Union's AI Act and the United Kingdom's General Data Protection Regulation, require a human in the loop (HITL) to preserve human agency, assuming the clinician safeguards decisions by evaluating algorithmic output, applying clinical judgment, and taking responsibility for the outcome. Anthropological analysis suggests that in practice HITL frequently reduces to a ritual of verification rather than a functional safeguard (Shneiderman 2022; Zheng et al. 2024).

Calling HITL "ritual" is analytic, not pejorative: Sign-off has the structural properties of ritual action as understood in social anthropology (Bell 1992) (a symbolic performance that legitimates by converting algorithmic output into an attributable human decision), while its causal efficacy, substantive correction of algorithmic error, is attenuated by the conditions of its performance. The clinician is justified in signing off as an epistemic safeguard, yet opacity plus workload yield low substantive scrutiny in practice, while sign-off still performs liability-transfer and institutional legitimation. This liability sink is HITL's primary function in many high-stakes environments: not to correct the machine, often assumed more accurate, but to absorb legal liability. Pressing "approve" converts the machine's output into a human decision, shifting liability from the software vendor to the individual professional (Lawton et al. 2024).

This is compounded by automation bias, the well-documented cognitive tendency to over-rely on automated systems and stop critically evaluating their outputs (Parasuraman and Manzey 2010). Because the brain conserves cognitive energy (the "cognitive miser" model), when paired with a highly reliable automated system, humans tend to treat machine outputs as authoritative (Nash et al. 2017, 894; Strickland et al. 2023, 860). The "loop" becomes a rubber-stamping exercise.

The resulting paradox of control is structurally irresolvable within the current framework: Agreeing with the AI renders the human functionally redundant, its output passing through unmodified, while disagreeing creates a double bind—overrule the superior machine and risk being wrong, or follow it and risk being wrong through blind reliance. This produces decision paralysis and moral distress (Kellar 2025). The HITL, conceived as a safeguard, becomes a mechanism for distributing blame.

6.2 The Phenomenology of the Double Bind

The double bind above is also phenomenological, not merely decision theoretic. A clinician confronted with an AI recommendation they cannot evaluate experiences a distinctive epistemic anxiety: the awareness of being asked to take responsibility for a judgment that is not, in any meaningful sense, their own. This differs from ordinary clinical uncertainty,

about whether a diagnosis is correct or a treatment will work; here, the uncertainty concerns the epistemic status of the information one is asked to act on.

Ordinary clinical uncertainty is reducible, at least in principle, through further investigation: more tests, more history, more consultation. Uncertainty generated by the algorithmic gaze is irreducible, because it concerns the epistemic process itself. The clinician cannot resolve it by examining the AI's reasoning since that reasoning is inaccessible. They can only decide whether to trust a process they cannot inspect, a decision that is existential rather than epistemic.

This connects to this article's broader argument: The algorithmic gaze changes not only what the clinician knows about the patient but also their relationship to their own knowledge. In pre-algorithmic medicine, that knowledge was earned through training, experience, and interpretive engagement with the patient, knowledge they could defend, explain, and take responsibility for. In AI-mediated medicine, the clinician increasingly acts on knowledge that is not their own, produced by a process they cannot access or evaluate. Their epistemic autonomy is diminished, and with it the moral grounding that connects knowledge to responsibility.

6.3 Synthetic Empathy and the Question of Care

In social robotics, the anthropological challenge is often framed as authenticity: Does it care, or merely simulate care? An influential counter-thesis, sometimes called "posthuman well-being," holds that this is the wrong question: What matters morally is not the caregiver's internal states but the patient's experienced outcomes (Šabanović 2010; DeFalco 2020). If a robot reduces loneliness and lowers cortisol, the absence of an inner stance is, on this view, morally irrelevant.

The thesis deserves serious engagement, though the critical care-robotics literature has long argued the opposite: Synthetic empathy without the homeostatic commitment that gives human care its moral weight betrays the vulnerable (Sharkey and Sharkey 2010). In resource-limited settings, pragmatic considerations may nonetheless legitimately favor algorithmic efficiency. But moving from "this technology produces the experience of being cared for" to "this technology is equivalent to care" is a substantive philosophical commitment, not a neutral empirical observation: It treats care as a service exhausted by its measurable effects, not the moral relation whose value depends partly on the kind of being occupying the caring position (section 4.3). Clinical AI's current deployment implicitly adopts the outcome-based view by default, not explicit deliberation; section 7's institutional response takes that anthropological choice as one to be made deliberately, not absorbed by ambient pressure.

7. Implications for Clinical Governance and Medical Education

AI should not be excluded from clinical practice: Its capabilities are real. But integrating AI into medicine requires explicit attention to the epistemic and anthropological transformations it produces, which current governance frameworks have not adequately addressed. The proposals below rest, for the most part, on the knowledge argument: They answer what the algorithmic gaze cannot bring into view (sections 2 and 5). Sections 7.1, 7.2, and 7.4 would be warranted even if the responsibility argument of sections 3.4 and 4

were rejected; section 7.3 draws on both, since institutional design for care (section 4.3) presupposes an agent who can assume responsibility for a particular patient. The allocation of liability (section 6.1, restated in section 8) follows from the responsibility argument and not from the knowledge argument.

7.1 Preserving Epistemic Pluralism

The most immediate practical implication is the need to preserve epistemic pluralism in clinical settings. If the algorithmic gaze produces one kind of knowledge (statistical, associative, bounded by Pearl's Level 1; sections 2 and 5) and the clinical gaze produces another (causal, narrative, embodied), both must be maintained as legitimate and complementary forms of clinical knowledge. The risk of AI integration is not that it will replace clinical judgment overnight but that it will gradually marginalize the forms of knowledge it cannot produce: narrative understanding, phenomenological perception, and relational attunement. Institutional structures (workflow design, quality metrics, performance evaluation) must be deliberately designed to value both computational and clinical knowledge.

7.2 Rethinking Medical Education

Medical education must equip clinicians to understand AI's epistemic character (what it knows, how it knows, and what it cannot know), which requires a philosophical dimension to training. Clinicians must distinguish associative from causal knowledge, recognize when algorithmic classification needs supplementing by clinical reasoning, and articulate the tools' epistemic limitations.

"AI literacy" usually means narrow technical competence: using the tool, interpreting outputs, recognizing error (Zhao et al. 2024). Instead, an epistemic literacy is needed, to explain why some clinical questions suit algorithmic analysis and others do not.

Three pedagogical interventions follow. First, a required preclinical module, "Epistemology of Clinical AI," should cover Pearl's hierarchy of causal inference, the associative ceiling, the distinction between pattern correlation and causal explanation, and the practical consequences for clinical decision-making, framing AI as a knowledge modality with characteristic strengths (pattern detection at scale, consistency) and structural limitations (substrate independence from embodied understanding, associative constraint). Students would ask questions such as: "What kind of knowledge does this tool produce? For what clinical questions is this knowledge sufficient? For what questions does it fail?"

Second, clinical reasoning courses should use case-based exercises comparing AI classifications with causal diagnoses reached through embodied clinical reasoning. A pulmonology module might present a dyspnea case: Students would first review the patient's data (imaging, labs, vital signs) and the AI risk score, then the clinician's assessment after history and physical exam, emphasizing embodied perceptual cues (posture, the character of breathing, tone of voice). This exercise highlights not only differing outputs but also differing epistemic process: what the algorithm can see versus what the clinician perceives through trained, embodied engagement with the patient.

Third, monthly residency case reviews examining divergences between AI recommendations and clinical judgment would build the epistemic virtues (caution, humility, willingness to dissent) good practice requires. These reviews would also serve the pedagogical function identified in section 2.1: Counterfactual reasoning is partly transmitted through supervision, as the experienced clinician's reasoning about what should have been considered becomes visible to the trainee, a transmission route AI-mediated clinical environments may foreclose without explicit institutional countermeasures. Assessment should include written questions asking residents to identify whether a recommendation rests on associative evidence (pattern correlation) or causal evidence (mechanistic understanding), and to articulate the epistemic stakes of that distinction for patient care.

7.3 The Question of Institutional Design

At the institutional level, the analysis suggests that hospitals and healthcare systems deploying AI must attend to its anthropological consequences, not merely its technical performance. If the algorithmic gaze occludes the patient's subjective experience (the structural claim of sections 3 and 5), institutional design must create countervailing structures (time for narrative, space for embodied encounter, cultures of relational care) that keep the occlusion from becoming total. The efficiency gains of AI-mediated triage must be balanced against the epistemic losses of algorithmic subject occlusion.

This is not merely about preserving a humanistic atmosphere within technical workflows; it is a question of which conception of care institutional design instantiates. The argument of section 4.3 applies: Tronto's (1993, 2013) political reformulation insists that care is a structural feature of just institutions, distributed across four phases (caring-about, caring-for, care-giving, care-receiving), each faring differently under algorithmic mediation. Caring-about, at population level, is precisely what AI does best: identifying who needs attention before they appear in clinic. Caring-for (assuming responsibility for the particular patient) cannot be discharged by an algorithm; it requires an agent who can in principle be held accountable. Care-giving, the direct work of care, is supportable by algorithmic systems where the work is well defined (medication reminders, monitoring) and unsupportable where it depends on improvisational responsiveness. Care-receiving, the recipient's response completing the act, requires an addressee who can be addressed as a fellow creature, which the algorithm cannot be. Kittay (1999) insists dependency is species typical, not pathological. Institutional design must honor the patient's dependency as normal, not manage it as deviation. Mol's (2008) distinction between the logic of care and the logic of choice deepens the point. Current clinical AI intensifies the logic of choice, leaving unprotected the logic of care that improvisational, ongoing clinical work depends on. Puig de la Bellacasa (2017) adds that care includes clinical work's infrastructural and material conditions—the architectures, schedules, incentives, and metrics that determine whether relational care can be sustained at all. "Institutional design for AI" is, then, institutional design for the conditions under which care, in the rich sense developed in section 4, remains possible.

The posthuman care thesis complicates this. Following Braidotti (2013), care has never been purely human but always involved technological mediation. This article contends that

current AI (associative, acausal, phenomenologically empty) differs in kind from prior technologies. A stethoscope extends perception; an algorithm substitutes for judgment.

7.4 A Concrete Proposal: Narrative Preservation Protocols

This proposal builds on an established tradition of narrative medicine. Rita Charon (2001, 2006) developed narrative medicine as a model of empathic, reflective practice grounded in "narrative competence": the capacity to absorb, interpret, and respond to stories (Charon 2001, 1897). Arthur W. Frank (1995) explains why narrative is constitutive of illness experience, not supplementary to it. Illness is a "call for stories" that disrupts the self's temporal coherence, and restoring coherence requires the patient be received as a teller, not only assessed as a case. Charon's five features of medical narrative—temporality, singularity, causality and contingency, intersubjectivity, and ethicality (Charon 2006)—supply the grid against which the algorithmic gaze can be measured. Each is transformed or eliminated when the patient becomes a data vector.

This article's contribution to that tradition is not rediscovering narrative competence but a structural-epistemological account of why narrative practices become institutionally necessary under AI mediation. Charon's case is humanistic and interpretive: Physicians should attend to stories because doing so makes them better physicians. The case here is structural: Where the algorithmic gaze systematically occludes the narrative register (section 5.3), narrative preservation protocols are not optional cultivations of physician virtue but institutional countermeasures against an epistemic foreclosure that would otherwise become permanent. Charon's tradition supplies the methodology; this argument supplies the epistemic warrant for embedding it in workflow and governance.

The foregoing analysis suggests a specific institutional intervention: mandatory narrative preservation protocols in clinical settings deploying AI decision support, operationalizing the epistemic pluralism advocated in section 7.1 by embedding structured narrative practices into AI-mediated workflows.

Three elements are proposed:

- *Narrative intake requirements:* Before any AI-generated risk score or diagnostic classification reaches the clinician, the patient must have the opportunity to give an unstructured narrative account of their condition (what phenomenologists term the *Erlebnis,* or lived experience of illness [Toombs 1992]), documented in the record alongside, not subordinated to, the algorithmic output, so that the patient's subjective experience retains epistemic standing rather than being displaced by quantified data.
- *Structured dissent documentation:* When a clinician disagrees with an AI recommendation, the workflow should require (and institutionally reward) documenting the clinician's reasoning for overriding the algorithm. Current HITL frameworks treat override as an anomaly (Agudo et al. 2024); narrative preservation protocols would treat it as an exercise of epistemic autonomy that strengthens rather than undermines clinical governance, creating an institutional record of cases where embodied clinical judgment diverges from algorithmic classification and generating data that could inform the iterative refinement of both human and machine performance.

- *Interdisciplinary narrative rounds:* regular case conferences discussing cases where the algorithmic gaze and the clinical gaze diverged, serving both an educational function (developing the epistemic literacy described in section 7.2) and a governance function, giving institutional visibility to forms of clinical knowledge algorithmic systems cannot capture. The model draws on the tradition of morbidity and mortality conferences (Lecoanet et al. 2016) but reorients them toward epistemic rather than purely clinical audit.

Three implementation considerations follow. Narrative intake should be structured but flexible (five to ten minutes of open-ended prompts), documented alongside structured data so the patient's lived experience is registered before algorithmic output appears. Dissent documentation should be embedded seamlessly in electronic health record workflows rather than added as an auxiliary form, with a structured template capturing the algorithmic recommendation, the clinician's reasoning, and the epistemological ground of the divergence. Narrative rounds should run as monthly multidisciplinary conferences reviewing cases of substantive algorithmic-clinical divergence, asking what the clinical gaze perceived that the algorithmic gaze could not. A pilot-first approach is strongly recommended: implementation in a single department for six to twelve months allows assessment of burden, acceptance, and cultural change before hospital-wide rollout. Success would be measured by shifts in institutional epistemic culture as well as clinical outcomes.

8. Conclusion

The argument of this article is that the integration of artificial intelligence into clinical medicine is not, in the philosophical-anthropological sense, a smooth technical upgrade. It constitutes an event in the genealogy of medical visibilities, a transformation in how the patient is perceived, how the clinician reasons, and how care is conceptualized. The algorithmic gaze, in the sense Dai and Guo (2025) gave it, and this article has extended, reconstructs the patient as a high-dimensional data vector, occludes the narrative and experiential register, and displaces the clinician's embodied judgment toward a computational process that no human (clinician, developer, or patient) can fully inspect. This transformation is not uniform across clinical activities. As section 2.2 illustrated, AI may match or exceed expert performance where the task is consistent, the input distribution stable, and the relevant features well characterized. The argument concerns the encounters in which causal-mechanistic and contextual reasoning are required, where the structural limits of associative reasoning become epistemically and morally consequential.

Five epistemic deficits identified in the foregoing analysis ground five corresponding governance responses, four of them developed in section 7. Those four follow from the knowledge argument. The associative ceiling of current AI architectures (Pearl's Level 1 limit; section 2.1) recommends a deliberate division of clinical labor, with AI assigned to high-volume associative tasks and human reasoning preserved for the interventional and counterfactual questions that pattern matching cannot answer (section 7.1). The semantic gap between computation and comprehension (sections 3 and 4) recommends narrative-preservation protocols that maintain the patient's lived experience as epistemically admissible in the clinical record before algorithmic output is presented (section 7.4). The opacity of algorithmic inference to expert inspection recommends the structured

documentation of clinician dissent, which returns epistemic visibility to a process that would otherwise remain inscrutable (section 7.4). The systematic occlusion of the narrative register through the three mechanisms of section 5.3 recommends interdisciplinary narrative rounds that institutionalize visibility for the forms of knowledge the algorithmic gaze cannot produce (section 7.4). The fifth response follows from the responsibility argument rather than the knowledge one, and none of the preceding four depends on it: The structural absence of normative agency from AI (sections 3.4 and 4) recommends liability structures that hold institutions and not individual clinicians accountable for AI-mediated outcomes, refusing the configuration in which clinicians become liability sinks for algorithmic decisions they did not author (section 6.1; Lawton et al. 2024). The governance proposals are, in this sense, not a checklist of institutional adjustments but the practical inheritance of the epistemic analysis the article has advanced.

These responses together preserve the conditions under which clinical knowledge (in the full sense developed in sections 3 and 4) can continue to be possible. The challenge is not the rejection of AI, whose capabilities are real and whose contributions to consistent associative classification are substantial, but the integration of AI under conditions that protect rather than erode the epistemic and relational dimensions of clinical practice that pattern matching cannot replicate. The algorithmic gaze reveals what data can reveal; the clinical gaze reveals what only embodied, mortal, caring beings can perceive. Medicine, as the previous sections have argued, requires both, and requires institutions designed to make both possible, rather than allowing one to absorb the other by default.

The empirical research program that would test the theoretical claims advanced here is substantial. Ethnographic studies of AI-mediated clinical encounters would document how the theoretical transformation manifests in practice, how clinicians and patients actually negotiate algorithmic authority, how narrative is preserved or occluded in real workflows, and what epistemic norms emerge in human-AI collaborative practice. Comparative work across medical specialties would clarify whether the governance proposals advanced here apply uniformly or differentially across acute, chronic, surgical, and medical domains. Longitudinal studies of clinical expertise, diagnostic reasoning, and the transmission of medical knowledge under deepening AI deployment would test the deskilling hedge committed to in section 5.4.

The limitations of the argument (its dependence on the European phenomenological tradition, its Global North case material, its provisional treatment of the human/machine binary, its nonengagement with disability studies and decolonial perspectives) are set out in section 9 and circumscribe what the present article can and cannot conclude.

9. Limitations

This argument is conceptual and theoretical. Several empirical claims extrapolate from the analyzed epistemic architecture, rather than being findings established at scale. The deskilling argument is deliberately a risk hypothesis (section 5.4); the narrative-occlusion claim is a structural prediction whose empirical realization in particular clinical settings would require ethnographic investigation (section 8); the moral-injury claims connect documented clinical phenomena to the epistemic conditions analyzed here but do not themselves establish causal links. Where empirical evidence supports the article's claims, it

is cited; where it does not, the claim is framed as a precautionary epistemic position rather than as an established empirical fact.

The article's scope is bounded geographically and philosophically. Its clinical-system examples and governance frameworks are drawn from US, UK, and EU contexts; whether the argument applies to low- and middle-income settings, where clinical AI may serve different functions (compensating for clinician shortages rather than supplementing established expertise), is not addressed here and would require its own treatment. The phenomenological tradition drawn upon is European, principally Husserl and Merleau-Ponty; non-Western accounts of personhood, mind, and care (African *Ubuntu* on personhood as relational, Buddhist analyses of mind, Confucian ethics of relational responsibility) offer resources for thinking about non-individualist personhood that would enrich the embodied-care argument and are not engaged here. These are real boundaries of scope rather than incidental omissions.

Three bodies of work that would complicate the argument's framing are acknowledged but not substantively engaged. Disability studies (particularly Havi Carel's [2016] *Phenomenology of Illness*) has critiqued accounts of embodiment that implicitly privilege "normal" bodies, raising whether the embodied-care argument's shared-vulnerability foundation tacitly assumes a body more uniform than illness, impairment, and disability allow. Decolonial scholarship on global health (Affun-Adegbulu and Adegbulu 2020) argues that biomedical individualism is itself a Western framework whose epistemology obscures relational and structural determinants of health, and whose datasets the algorithmic gaze inherits and may amplify. Feminist science and technology studies (Suchman 2007; Haraway 1991), with the critical-posthuman nursing scholarship of the Compost Collaborative (Smith and Willis 2020; Hopkins-Walsh et al. 2022), contest the human/machine binary on which the article's analytic structure partly relies, treating clinical work as an assemblage of human, technological, and material elements, rather than a separation between distinct agents.

To be clear, the article does not argue that AI should be excluded from clinical practice; its capabilities are real and substantial. It does not claim its empirical predictions about deskilling and narrative occlusion are demonstrated; the precautionary framing is preserved throughout. It does not deny that individual clinicians can fail to exhibit the capacities the four-level taxonomy treats as constitutive of responsible practice; the structural-versus-empirical distinction of section 3.4 holds that case open and morally evaluable. And it does not claim the analytic framework adopted here (Pearl plus the consciousness taxonomy plus embodied care) is the only valid way of analyzing these questions; alternative framings will produce alternative emphases.

However, the algorithmic gaze can tell us what the data permit; it cannot tell us what the patient means, and a medicine that forgets the difference will have optimized its way out of the very knowledge that made it medicine.

Acknowledgments

The author is grateful to the colleagues whose conversations informed this article, and in particular to Vicky Gregory (the University of Law) for the discussions that helped shape its argument.

Disclosure Statement
The author declares no competing interests.

References
Affun-Adegbulu, Clara, and Opemiposi Adegbulu. 2020. "Decolonising Global (Public) Health: From Western Universalism to Global Pluriversalities." *BMJ Global Health* 5, e002947. https://doi.org/10.1136/bmjgh-2020-002947.

Agudo, Ujué, Karlos G. Liberal, Miren Arrese, and Helena Matute. 2024. "The Impact of AI Errors in a Human-in-the-Loop Process." *Cognitive Research: Principles and Implications* 9, no. 1, 1. https://doi.org/10.1186/s41235-023-00529-3.

Bell, Catherine. 1992. *Ritual Theory, Ritual Practice*. Oxford University Press.

Bishop, J. Mark. 2021. "Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It." *Frontiers in Psychology* 11, 513474. https://doi.org/10.3389/fpsyg.2020.513474.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Braidotti, Rosi. 2013. *The Posthuman*. Polity Press.

———. 2019. *Posthuman Knowledge*. Polity Press.

Bratman, Michael. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press.

Carel, Havi. 2016. *Phenomenology of Illness*. Oxford University Press.

Charon, Rita. 2001. "Narrative Medicine: A Model for Empathy, Reflection, Profession, and Trust." *JAMA* 286, no. 15: 1897–1902. https://doi.org/10.1001/jama.286.15.1897.

———. 2006. *Narrative Medicine: Honoring the Stories of Illness*. Oxford University Press.

Croskerry, Pat. 2003. "The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them." *Academic Medicine* 78, no. 8: 775–780. https://doi.org/10.1097/00001888-200308000-00003.

Dai, Yuxin, and Sizhe Guo. 2025. "Algorithmic Gaze and Subject Occlusion: A Medical Ethical Critique of Artificial Intelligence Diagnosis and Treatment from a Foucaultian Perspective." *Medicine, Health Care and Philosophy* 29: 523–534. https://doi.org/10.1007/s11019-025-10318-y.

Damásio, António. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. G.P. Putnam's Sons.

———. 1996. "The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 351, no. 1346: 1413–1420. https://doi.org/10.1098/rstb.1996.0125.

Dean, Wendy, Simon Talbot, and Austin Dean. 2019. "Reframing Clinician Distress: Moral Injury Not Burnout." *Federal Practitioner* 36, no. 9: 400–402. https://www.mdedge.com/fedprac/article/207458/mental-health/reframing-clinician-distress-moral-injury-not-burnout.

DeFalco, Amelia. 2020. "Towards a Theory of Posthuman Care: Real Humans and Caring Robots." *Body & Society* 26, no. 3: 31–60. https://doi.org/10.1177/1357034X20917450.

De Fauw, Jeffrey, Joseph R. Ledsam, Bernardino Romera-Paredes, et al. 2018. "Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease." *Nature Medicine* 24, no. 9: 1342–1350. https://doi.org/10.1038/s41591-018-0107-6.

Dennett, Daniel C. 2017. *From Bacteria to Bach and Back: The Evolution of Minds*. Allen Lane.

Eraut, Michael. 2000. "Non-Formal Learning and Tacit Knowledge in Professional Work." *British Journal of Educational Psychology* 70, no. 1: 113–136. https://doi.org/10.1348/000709900158001.

Finlayson, Samuel G., Adarsh Subbaswamy, Karandeep Singh, et al. 2021. "The Clinician and Dataset Shift in Artificial Intelligence." *New England Journal of Medicine* 385, no. 3: 283–286. https://doi.org/10.1056/NEJMc2104626.

Foucault, Michel. 1973. *The Birth of the Clinic: An Archaeology of Medical Perception*. Routledge.

Frank, Arthur W. 1995. *The Wounded Storyteller: Body, Illness, and Ethics*. University of Chicago Press.

Frankfurt, Harry. (1969) 2018. "Alternate Possibilities and Moral Responsibility." In *Moral Responsibility and Alternative Possibilities,* edited by Michael McKenna and David Widerker. Routledge.

Friston, Karl. 2010. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience* 11, no. 2: 127–138. https://doi.org/10.1038/nrn2787.

Friston, Karl, Lancelot Da Costa, Danijar Hafner, Casper Hesp, and Thomas Parr. 2021. "Sophisticated Inference." *Neural Computation* 33, no. 3: 713–763. https://doi.org/10.1162/neco_a_01351.

Friston, Karl J., Maxwell J.D. Ramstead, Alex B. Kiefer, et al. 2024. "Designing Ecosystems of Intelligence from First Principles." *Collective Intelligence* 3, no. 1. https://doi.org/10.1177/26339137231222481.

Gabbay, John, and Andrée Le May. 2004. "Evidence Based Guidelines or Collectively Constructed 'Mindlines?' Ethnographic Study of Knowledge Management in Primary Care." *British Medical Journal (Clinical research ed.)* 329, no. 7473, 1013. https://doi.org/10.1136/bmj.329.7473.1013.

Greenhalgh, Trisha, Jeremy Howick, and Neal Maskrey. 2014. "Evidence Based Medicine: A Movement in Crisis?" *British Medical Journal* 348, g3725. https://doi.org/10.1136/bmj.g3725.

Grieves, Michael, and John Vickers. 2017. "Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems." In *Transdisciplinary Perspectives on Complex Systems,* edited by Franz-Josef Kahlen, Shannon Flumerfelt, and Anabela Alves. Springer International Publishing.

Haraway, Donna J. 1991. "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century." In *Simians, Cyborgs and Women: The Reinvention of Nature*. Routledge.

Hopkins-Walsh, Jane, Jessica Dillard-Wright, Brandon Blaine Brown, Jamie Smith, and Eve Willis. 2022. "Critical Posthuman Nursing Care: Bodies Reborn and the Ethical Imperative for Composting." *Witness: The Canadian Journal of Critical Nursing Discourse* 4, no. 1: 16–35. https://doi.org/10.25071/2291-5796.126.

Husserl, Edmund. (1913) 2014. *Ideas for a Pure Phenomenology and Phenomenological Philosophy: First Book*. Hackett Publishing.

Jaspers, Karl. (1913) 1997. *General Psychopathology*. Edited by J. Hoenig and Marian W. Hamilton. Johns Hopkins University Press.

Kellar, Robert. 2025. "AI in Healthcare: Redefining Liability for Doctors and Hospitals." *British Journal of Hospital Medicine* 86, no. 9: 1–6. https://doi.org/10.12968/hmed.2025.0212.

Kittay, Eva Feder. 1999. *Love's Labor: Essays on Women, Equality, and Dependency*. Routledge.

———. 2002. "Love's Labor Revisited." *Hypatia* 17, no. 3: 237–250. https://doi.org/10.1353/hyp.2002.0061.

Lawton, Tom, Phillip Morgan, Zoe Porter, et al. 2024. "Clinicians Risk Becoming 'Liability Sinks' for Artificial Intelligence." *Future Healthcare Journal* 11, no. 1, 100007. https://doi.org/10.1016/j.fhj.2024.100007.

Lecoanet, André, Gwenaëlle Vidal-Trecan, Frédéric Prate, et al. 2016. "Assessment of the Contribution of Morbidity and Mortality Conferences to Quality and Safety Improvement: A Survey of Participants' Perceptions." *BMC Health Services Research* 16, no. 1, 176. https://doi.org/10.1186/s12913-016-1431-5.

Marcos, Henrique Jerônimo Bezerra. 2023. "Consciousness and Liability of Non-Human Intelligence." *Revista de Estudos Constitucionais, Hermenêutica e Teoria do Direito* 14, no. 3: 382–401. https://doi.org/10.4013/rechtd.2022.143.06.

Matsumura, Tadayuki, Kanako Esaki, Shao Yang, Chihiro Yoshimura, and Hiroyuki Mizuno. 2024. "Active Inference with Empathy Mechanism for Socially Behaved Artificial Agents in Diverse Situations." *Artificial Life* 30, no. 2: 277–297. https://doi.org/10.1162/artl_a_00416.

Merleau-Ponty, Maurice. (1945) 2013. *Phenomenology of Perception*. Routledge.

Mol, Annemarie. 2008. *The Logic of Care: Health and the Problem of Patient Choice*. Routledge.

Moreno-Torres, Jose García, José A. Sáez, and Francisco Herrera. 2012. "Study on the Impact of Partition-Induced Dataset Shift on *K*-Fold Cross-Validation." *IEEE Transactions on Neural Networks and Learning Systems* 23, no. 8: 1304–1312. https://doi.org/10.1109/TNNLS.2012.2199516.

Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *The Philosophical Review* 83, no. 4: 435–450. https://doi.org/10.2307/2183914.

Nash, Robert A., Kimberley A. Wade, Maryanne Garry, and James S. Adelman. 2017. "A Robust Preference for Cheap-and-Easy Strategies over Reliable Strategies When Verifying Personal Memories." *Memory* 25, no. 7: 890–899. https://doi.org/10.1080/09658211.2016.1214280.

Noddings, Nel. 2013. *Caring: A Relational Approach to Ethics and Moral Education*. 2nd edition. University of California Press.

Parasuraman, Raja, and Dietrich H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors* 52, no. 3: 381–410. https://doi.org/10.1177/0018720810376055.

Pearl, Judea. 2018. "Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution." *arXiv*. https://doi.org/10.48550/arXiv.1801.04016.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Peerson, Anita. 1995. "Foucault and Modern Medicine." *Nursing Inquiry* 2, no. 2: 106–114. https://doi.org/10.1111/j.1440-1800.1995.tb00073.x.

Pellegrino, Edmund D., and David C. Thomasma. 1993. *The Virtues in Medical Practice*. Oxford University Press.

Puig de la Bellacasa, María. 2017. *Matters of Care: Speculative Ethics in More Than Human Worlds*. University of Minnesota Press.

Rincon, Jaime A., Angelo Costa, Paulo Novais, Vicente Julian, and Carlos Carrascosa. 2018. "A New Emotional Robot Assistant That Facilitates Human Interaction and Persuasion." *Knowledge and Information Systems* 60, no. 1: 363–383. https://doi.org/10.1007/s10115-018-1231-9.

Šabanović, Selma. 2010. "Robots in Society, Society in Robots." *International Journal of Social Robotics* 2, no. 4: 439–450. https://doi.org/10.1007/s12369-010-0066-7.

Saposnik, Gustavo, Donald Redelmeier, Christian C. Ruff, and Philippe N. Tobler. 2016. "Cognitive Biases Associated with Medical Decisions: A Systematic Review." *BMC Medical Informatics and Decision Making* 16, no. 1, 138. https://doi.org/10.1186/s12911-016-0377-1.

Scanlon, T.M. 2000. *What We Owe to Each Other*. Belknap Press.

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2023. "Are Emergent Abilities of Large Language Models a Mirage?" In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. https://doi.org/10.52202/075280-2425.

Schölkopf, Bernhard, Francesco Locatello, Stefan Bauer, et al. 2021. "Toward Causal Representation Learning." *Proceedings of the IEEE* 109, no. 5: 612–634. https://doi.org/10.1109/JPROC.2021.3058954.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3: 417–424. https://doi.org/10.1017/S0140525X00005756.

Sharkey, Amanda, and Noel Sharkey. 2010. "Granny and the Robots: Ethical Issues in Robot Care for the Elderly." *Ethics and Information Technology* 14, no. 1: 27–40. https://doi.org/10.1007/s10676-010-9234-6.

Shneiderman, Ben. 2022. *Human-Centered AI*. Oxford University Press.

Smith, Jamie B., and Eva Willis. 2020. "Interpreting Posthumanism with Nurse Work." *Journal of Posthuman Studies* 4, no. 1: 59–75. https://doi.org/10.5325/jpoststud.4.1.0059.

Strawson, Peter F. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48: 1–25.

Strickland, Luke, Russell J. Boag, Andrew Heathcote, Vanessa Bowden, and Shayne Loft. 2023. "Automated Decision Aids: When Are They Advisors and When Do They Take Control of Human Decision Making?" *Journal of Experimental Psychology: Applied* 29, no. 4: 849–868. https://doi.org/10.1037/xap0000463.

Suchman, Lucy A. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd edition. Cambridge University Press.

Toombs, S. Kay. 1992. *The Meaning of Illness: A Phenomenological Account of the Different Perspectives of Physician and Patient*. Kluwer Academic Publishers.

Topol, Eric J. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books.

Tourassi, Georgia, Sophie Voisin, Vincent Paquit, and Elizabeth Krupinski. 2013. "Investigating the Link Between Radiologists' Gaze, Diagnostic Decision, and Image Content." *Journal of the American Medical Informatics Association* 20, no. 6: 1067–1075. https://doi.org/10.1136/amiajnl-2012-001503.

Tronto, Joan C. 1993. *Moral Boundaries: A Political Argument for an Ethic of Care*. Routledge.

———. 2013. *Caring Democracy: Markets, Equality, and Justice*. New York University Press.

Vallor, Shannon. 2011. "Carebots and Caregivers: Sustaining the Ethical Ideal of Care in the Twenty-First Century." *Philosophy & Technology* 24, no. 3: 251–268. https://doi.org/10.1007/s13347-011-0015-x.

Wei, Jason, Yi Tay, Rishi Bommasani, et al. 2022. "Emergent Abilities of Large Language Models." *arXiv*. https://doi.org/10.48550/arXiv.2206.07682.

Zhao, Xin, Andrew Cox, and Liang Cai. 2024. "ChatGPT and the Digitisation of Writing." *Humanities and Social Sciences Communications* 11, no. 1, 482. https://doi.org/10.1057/s41599-024-02904-x.

Zheng, Elise Li, Weina Jin, Ghassan Hamarneh, and Sandra Soo-Jin Lee. 2024. "From Human-in-the-Loop to Human-in-Power." *American Journal of Bioethics* 24, no. 9: 84–86. https://doi.org/10.1080/15265161.2024.2377139.