Original Scientific Paper/Original naučni rad UDC/UDK: 004.8:616-07
Paper Submitted/Rad primljen:31.12.2025.
Paper Accepted/Rad prihvaćen:20.01.2026.
DOI: 10.5937/SJEM2600076M

Veštačka inteligencija u medicinskoj dijagnostici: Kritički pregled rizika, odgovornosti i epistemoloških ograničenja velikih jezičkih modela

Marjan Marjanović[1], Luka Latinović [2]

[1] Clinic for General, Visceral and Thoracic Surgery - InnKlinikum, Altötting, Germany
[2]Belgrade School of Engineering Management, Beopolis University, Belgrade, Serbia, luka.latinovic@fim.rs

Sažetak: Brza integracija veštačke inteligencije u dijagnostičku medicinu predstavlja istovremeno tehnološki napredak i epistemološki poremećaj. Ovaj narativni kritički pregled ispituje kako veliki jezički modeli i srodni sistemi veštačke inteligencije utiču na dijagnostičko rasuđivanje, kliničku odgovornost i medicinsku etiku. Iako je veštačka inteligencija pokazala izuzetan potencijal u prepoznavanju obrazaca i sintezi podataka, njena logika zasnovana na korelacijama lišena je uzročno-posledičnog razumevanja, interpretabilnosti i moralne odgovornosti. Rad identifikuje ključne rizike kao što su automatizaciona pristrasnost, pristrasnost skupa podataka, neprozirnost modela i asimetrija odgovornosti između programera i lekara. Integracija veštačke inteligencije izaziva tradicionalne granice profesionalne odgovornosti i informisanog pristanka, dovodeći u pitanje epistemičku validnost i poverenje pacijenata. Postojeći okviri upravljanja, prilagođeni statičkoj regulativi medicinskih uređaja, pokazuju se neadekvatnim za sisteme koji kontinuirano uče. Istraživanje pokazuje da očuvanje integriteta medicinskog rasuđivanja u doba veštačke inteligencije zahteva obnovljenu posvećenost epistemičkoj skromnosti, raspodeljenoj odgovornosti i očuvanju ljudskog suda u okviru algoritamski posredovane zdravstvene nege.

Ključne reči: epistemički integritet, dijagnostičko rasuđivanje, medicinska odgovornost, pristrasnost automatizacije, regulatorna etika.

Artificial Intelligence in Medical Diagnostics: A Critical Narrative Review of Risks, Responsibility, and the Epistemological Limits of Large Language Models

Abstract: The rapid integration of artificial intelligence (AI) into diagnostic medicine represents both a technological advance and an epistemological disruption. This narrative critical review examines how large language models and related AI systems influence diagnostic reasoning, clinical accountability, and medical ethics. While AI has demonstrated remarkable potential in pattern recognition and data synthesis, its correlation-based logic lacks causal understanding, interpretability, and moral agency. The review identifies key risk domains including automation bias, dataset bias, model opacity, and liability asymmetries between developers and physicians. It argues that AI's integration challenges traditional boundaries of professional responsibility and informed consent, raising concerns over epistemic validity and patient trust. Current governance frameworks, adapted from static medical device regulation, remain ill-suited for continuously learning systems. The research shows that safeguarding the integrity of medical reasoning in the age of AI requires a renewed commitment to epistemic humility, distributed accountability, and the preservation of human judgment within computationally mediated care.

Keywords: epistemic integrity, diagnostic reasoning, medical accountability, automation bias, regulatory ethics.

1. Introduction

The deployment of artificial intelligence (AI) in clinical medicine has advanced rapidly in recent years, propelled by improved computational capacity, expanding digital-health datasets, and the emergence of large language models (LLMs) capable of complex pattern recognition, natural language processing and decision-support tasks (Fahim et al., 2025; Graili & Farhoudi, 2025; Krishnan et al., 2023). For instance, reviews report that AI systems

are widely applied in diagnostic workflows—spanning medical imaging, biosignal interpretation, electronic health-record analytics and predictive modelling of disease onset or progression (Sadr et al., 2025; Sinha, 2024). This growth promises considerable benefits: faster detection of pathology, more consistent triage, support for clinicians in assimilating broad swathes of data, and ultimately, improved patient-outcomes (McGenity et al., 2024; van Diest et al., 2024).

Nevertheless, alongside this promise there are substantive risks and unresolved questions—especially when AI is used to assist or partially automate clinical decision-making such as diagnosis (Chustecki, 2024; Khosravi et al., 2024a). Critically, while clinicians remain legally and ethically responsible, AI tools such as LLMs increasingly influence those decisions (Jones et al., 2023; MacIntyre et al., 2023; Pham, 2025). In that context, it is imperative to examine the epistemological foundations (how diagnoses are arrived at), the clinical and ethical ramifications (when AI errs), and the accountability structures (who bears responsibility). The diagnostic domain is particularly sensitive because diagnostic error has direct patient-harm implications. Medical literature indicates that many AI diagnostic systems are trained and validated under idealised conditions, using curated datasets and retrospective designs; their performance in real-world heterogeneous clinical settings often remains uncertain (Angus et al., 2025; Sourlos et al., 2024). Moreover, many LLM-based systems operate as "black boxes" with limited transparency of how inputs map to outputs, raising concerns about trust, interpretability and clinician over-reliance (automation bias) (Savage et al., 2024; Ullah et al., 2024). These features generate a latent danger: doctors relying on AI-driven suggestions may defer too readily to the model, reducing their own critical scrutiny.

Equally important is the question of epistemic validity: the AI tool may recognise statistical correlations but lacks genuine causal understanding of pathophysiology, which may lead to plausible-looking but erroneous diagnoses (Sanchez et al., 2022; Xu & Shuttleworth, 2024). As the literature underscores, AI's diagnostic suggestions must be viewed as probabilistic, supportive, rather than definitive; yet in practice, the clinician's stamp often converts the suggestion into a formal diagnosis with attendant legal/ethical weight (Alowais et al., 2023; Bozyel et al., 2024).

In this narrative review, we focus specifically on the risks associated with the use of LLMs and other AI tools in medical diagnostics, in contexts where a physician remains the final accountable decision-maker. We analyse existing literature on diagnostic performance, error modes, bias, interpretability, workflow integration and responsibility. Our aim is to provide a critical account of how these tools interplay with clinical decision-making, where their limitations lie, and how responsibility and epistemic integrity might be preserved in an era of AI-augmented diagnosis.

1.1. Approach and Scope

This paper adopts a **narrative critical review** design rather than a systematic or scoping review, owing to the conceptual nature of the topic and the interpretive orientation required to assess epistemological and ethical risks in AI-assisted diagnostics. The goal is not to quantify evidence across studies but to critically examine how artificial intelligence—particularly large language models (LLMs)—reconfigures clinical reasoning, medical accountability, and diagnostic epistemology. The narrative review approach is appropriate where the research problem transcends purely empirical boundaries and involves theoretical, ethical, and philosophical considerations that cannot be meaningfully reduced to numerical synthesis (Greenhalgh et al., 2018; Sukhera, 2022). In this case, diagnostic accuracy metrics, error rates, and algorithmic benchmarks, while informative, are insufficient to illuminate how AI influences the clinician's judgment, responsibility, and interpretive autonomy.

The review draws upon peer-reviewed journal articles, conceptual papers, policy reports, and interdisciplinary analyses published in the last half-decade across medicine, cognitive science, bioethics, and science-and-technology studies. These sources were identified through purposive reading rather than database-driven screening, ensuring inclusion of perspectives that critically interrogate not only performance outcomes but also the epistemic underpinnings of diagnostic reasoning. The critical-interpretive orientation of this review entails three guiding premises:

1. **Contextual evaluation** – each source is interpreted in relation to the broader clinical, legal, and epistemological context in which AI systems operate.
2. **Conceptual synthesis** – emphasis is placed on tracing how recurring themes—such as automation bias, opacity, and accountability—interrelate across disciplines.
3. **Normative reflection** – beyond describing empirical findings, the analysis interrogates the moral and epistemic consequences of delegating diagnostic reasoning to non-human agents.

By integrating these perspectives, the narrative review allows for a coherent critical account of AI's role in diagnostic medicine—an account that would be obscured by the procedural constraints of systematic evidence aggregation. This interpretive flexibility is essential for capturing the complexity of emerging clinical–technological assemblages where algorithms increasingly participate in, but do not yet own, the act of medical judgment.

2. Artificial Intelligence in Clinical Diagnostics

The integration of artificial intelligence into clinical diagnostics represents a decisive inflection point in the evolution of medical practice. Traditionally, diagnostic reasoning has relied upon the physician's capacity to synthesise heterogeneous information—symptom presentation, medical history, laboratory findings, and imaging data—into a coherent clinical hypothesis. The arrival of AI systems has introduced computational models capable of performing many of these interpretive functions autonomously or semi-autonomously, thus altering both the epistemological structure of diagnosis and the practical organisation of healthcare delivery (D'Adderio & Bates, 2025; Saenz et al., 2023; Wu et al., 2024).

AI's involvement in diagnostics initially emerged through **rule-based expert systems**, designed to emulate decision trees reflecting human medical reasoning (Alowais et al., 2023; Kulikowski, 2015; Solutions, 2024). These early systems proved rigid, as their performance depended heavily on the explicitness and completeness of the encoded knowledge base. The contemporary landscape is instead dominated by **data-driven approaches**, most notably deep learning architectures and large language models (Wang & Zhang, 2024; Zhou et al., 2025). These systems infer diagnostic associations from vast quantities of data—imaging repositories, genomic databases, or electronic health records—without presupposing an explicit theoretical model of disease.

This shift from symbolic to sub-symbolic reasoning has generated unprecedented diagnostic capabilities. Machine vision models now achieve or exceed human-level accuracy in detecting malignancies on radiographs, histopathological slides, and dermatological images (Jeong et al., 2022; Krakowski et al., 2024; McGenity et al., 2024; Waqas et al., 2023). Predictive models assist in early identification of cardiovascular, metabolic, and neurodegenerative conditions by integrating multi-modal data streams (DeGroat et al., 2024; Xue et al., 2024). Most recently, large language models have entered the diagnostic domain through their ability to process unstructured clinical narratives, suggest differential diagnoses, and assist in generating clinical documentation (Nazi & Peng, 2024; Singhal et al., 2023; Vrdoljak et al., 2025; Zhou et al., 2025). Their flexible linguistic competence allows them to operate across specialties, functioning as conversational diagnostic aides that simulate expert dialogue.

However, this transformation also entails a profound **epistemic displacement**. Whereas human diagnostic reasoning is anchored in pathophysiological understanding and clinical context, AI models operate through correlation-based pattern recognition (Boge & Mosig, 2025; Tikhomirov et al., 2024). They may produce correct answers for the wrong reasons, or plausible but incorrect answers that appear clinically credible. The opacity of neural-network inference processes precludes meaningful introspection into why a particular diagnosis was proposed, complicating both verification and accountability.

In practice, AI is increasingly positioned as a co-participant in diagnostic reasoning (Goh et al., 2024). Clinical workflows now involve hybrid arrangements in which AI systems pre-screen imaging studies, flag anomalies, or generate ranked diagnostic suggestions for the physician's review. Such configurations ostensibly preserve human oversight, yet they also shift the cognitive landscape of clinical decision-making (Korfiatis et al., 2025). The physician's role transitions from primary diagnostician to meta-analyst of algorithmic output. This role change carries consequences for professional identity, epistemic authority, and legal responsibility.

Another emergent dynamic concerns **automation bias**, wherein clinicians may develop an uncritical trust in algorithmic recommendations, particularly when confronted with time pressure or cognitive fatigue (Abdelwanis et al., 2024; Goh et al., 2025; Khosravi et al., 2024b). Conversely, distrust of AI systems—especially following a perceived error—may lead to over-correction and rejection of potentially valuable input (Rosenbacke et al., 2024; Tun et al., 2025). The dialectic between overreliance and overcaution illustrates that AI's integration into diagnostic medicine is not merely a technical implementation challenge but a cognitive and cultural transformation of medical epistemology itself.

Therefore, the advance of AI in clinical diagnostics cannot be understood solely as a technological progression. It represents a reconfiguration of how diagnostic knowledge is produced, validated, and acted upon. The subsequent sections of this paper address these transformations by examining the epistemological limits of AI reasoning, the

associated risk domains, and the evolving moral and legal frameworks through which medical responsibility is being renegotiated.

2.1. Epistemological Foundations of AI-Driven Diagnosis

The epistemological foundations of medical diagnosis have always rested on the clinician's capacity to connect empirical observation with theoretical understanding. Diagnostic reasoning traditionally involves a dialectical movement between inductive inference from patient-specific findings and deductive application of biomedical knowledge (Shin, 2019). In contrast, artificial intelligence—particularly large language models and deep learning systems—operates through statistical correlation rather than causal comprehension. This difference marks a fundamental epistemic rupture: AI systems recognise patterns in data but lack awareness of what those patterns represent within the biological or clinical ontology of disease (Nicholson et al., 2023).

2.2. Correlation versus Causation in Algorithmic Reasoning

While human diagnostic thought aspires to causal explanation, AI models are optimised for predictive accuracy. They learn associations that maximise statistical fit, not explanatory coherence. This epistemic asymmetry means that an algorithm may correctly predict that a given constellation of symptoms aligns with a particular pathology, yet remain oblivious to the underlying mechanisms producing those symptoms. As such, its "knowledge" is non-conceptual; it consists of distributed weights within a network, not propositions about biological reality. Consequently, even when AI models achieve high diagnostic accuracy, they do not possess the epistemic properties that confer medical understanding—intentionality, contextual judgment, or counterfactual reasoning.

This limitation has practical implications. When the underlying data are biased, incomplete, or unrepresentative, correlations learned by the model may amplify pre-existing distortions (Cross et al., 2024; Mittermaier et al., 2023). A model trained primarily on imaging data from one demographic group may misinterpret normal variations in another as pathological. Because the system does not "know" why its prediction holds, it cannot recognise when its reasoning falls outside valid clinical context (Koçak et al., 2025; Tejani et al., 2024; Vrudhula et al., 2024). In human medicine, such epistemic self-monitoring—awareness of uncertainty, reflection on plausibility, and revision of hypotheses—is integral to diagnostic reliability. AI systems, by contrast, lack meta-cognitive capacity; they cannot interrogate their own inferences.

2.3. The Problem of Opacity and the "Black Box" Dilemma

A defining challenge of deep learning and LLM-based diagnostics is their opacity. The internal operations of these models, involving billions of parameters, are not interpretable in a way that corresponds to human reasoning (Chen et al., 2022; Zhao et al., 2024). The clinician cannot meaningfully trace how input variables—symptoms, images, test results—lead to a specific diagnostic output. This "black box" character undermines transparency and complicates trust. In medicine, where the justification of a diagnosis must be communicable and defensible, opacity introduces epistemic risk. A diagnostic conclusion whose provenance cannot be articulated cannot be ethically or legally justified (Freyer et al., 2024).

Attempts to enhance interpretability through "explainable AI" frameworks offer partial solutions, but these often provide post hoc rationalisations rather than genuine insight into the model's inferential process (Acun & Nasraoui, 2025; Sadeghi et al., 2024). Visual heatmaps, saliency maps, and token importance rankings may show where the model "focused," but they do not explain *why* it reached a particular conclusion (Rao & Aalami, 2023; Saarela & Podgorelec, 2024). Thus, even well-intentioned efforts to make AI more interpretable often remain epistemologically superficial.

2.4. Epistemic Authority and the Physician's Cognitive Displacement

The introduction of AI diagnostic tools has also begun to redistribute epistemic authority within clinical settings. When algorithmic recommendations are integrated into decision support systems, they implicitly acquire a status of evidentiary credibility (Jones et al., 2023). Physicians, conscious of AI's statistical prowess, may defer to its suggestions, especially in ambiguous cases (Kostick-Quenet & Gerke, 2022). This subtle shift can lead to a **cognitive displacement**, wherein the clinician's reasoning becomes secondary to the machine's inference (Patil et al., 2025; Tikhomirov et al., 2024). Over time, such displacement may erode the epistemic autonomy that has historically defined medical professionalism.

This transformation also challenges the social contract between physician and patient. The act of diagnosis traditionally entails a moral responsibility grounded in human judgment—the clinician stands accountable for both error and care. When part of that reasoning is delegated to an opaque computational entity, the ethical foundation of that accountability becomes unstable. The physician may still bear formal responsibility, but the epistemic control over the decision has become distributed and partially inaccessible. This asymmetry—responsibility without full epistemic authority—creates an untenable moral configuration in which practitioners are held accountable for outcomes they cannot fully understand or verify. The epistemological tension between human and machine reasoning, therefore, lies at the core of AI-assisted diagnostics. While AI systems can expand the range and precision of pattern recognition, they simultaneously displace the causal and interpretive foundations upon which medical diagnosis depends. The next section examines how these epistemic vulnerabilities manifest as concrete risks—statistical, procedural, and ethical—within real-world clinical environments.

3. Risk Domains in AI-Assisted Diagnostics

Artificial intelligence, while promising in diagnostic medicine, introduces a spectrum of risks that extend beyond technical error. These risks manifest at the intersection of epistemology, human cognition, and institutional responsibility. Each domain reflects a distinctive way in which the deployment of AI challenges the reliability, safety, and ethical legitimacy of clinical decision-making.

3.1. Diagnostic Error and Model Hallucination

AI diagnostic systems are not immune to error; in fact, their errors are often opaque and systematic (Xu & Shuttleworth, 2024). Deep learning models and large language models are prone to what has been described as *hallucination*—the confident generation of incorrect information or conclusions (Asgari et al., 2025; Chelli et al., 2024). In a medical context, such hallucination can produce diagnostic suggestions that are linguistically convincing yet clinically false. Unlike human reasoning, which typically encodes some awareness of uncertainty, AI systems tend to express outputs with unwarranted confidence, creating an illusion of precision (Afroogh et al., 2024; Rezaeian et al., 2025). This dynamic heightens the risk of misdiagnosis when clinicians interpret machine-generated statements as authoritative.

Another source of diagnostic error stems from dataset bias. Training data frequently overrepresent certain demographic groups, institutions, or disease categories, resulting in models that generalise poorly across populations (Koçak et al., 2025). Errors can thus disproportionately affect under-represented groups, compounding existing health inequities. Because many models are developed in research settings with highly controlled inputs, their apparent performance often degrades under real-world clinical variability—differences in imaging devices, patient populations, and record-keeping conventions (Wellnhofer, 2022; Yang et al., 2024).

3.2. Data Bias and Representational Inequities

The epistemic core of AI diagnostics rests on data representation, and representation is never neutral (Pagano et al., 2023; Stinson, 2022). Every training dataset reflects historical practices, institutional priorities, and structural inequities (Agarwal et al., 2023; Hanna et al., 2025; Jui & Rivas, 2024; Zajko, 2022). When these inequities are encoded into the model, they become perpetuated at scale. For instance, diagnostic tools trained predominantly on data from high-income regions may fail to recognise pathologies common in low-resource contexts (Celi et al., 2022; Joseph, 2025). Similarly, gender or racial bias in symptom reporting can distort probabilistic inference (Colacci et al., 2025; Mittermaier et al., 2023). The danger lies not only in biased predictions but in the false appearance of objectivity that accompanies algorithmic output.

Such representational distortions are epistemologically pernicious because they are difficult to detect once embedded in the model (Hasanzadeh et al., 2025). Bias audits and fairness metrics offer partial safeguards, yet they remain reactive—addressing observed disparities rather than confronting the social and epistemic structures that produce biased data in the first place. The ethical consequence is that AI may reinforce diagnostic hierarchies under the guise of technological neutrality.

3.3. Automation Bias and Overreliance by Clinicians

Automation bias describes the human tendency to overtrust algorithmic systems, accepting their outputs even when inconsistent with clinical judgment. In diagnostic contexts, this bias is amplified by the authority ascribed

to quantitative or computational processes. When AI systems are presented as advanced, evidence-based, or statistically superior, clinicians may defer to them in situations of uncertainty or cognitive fatigue (Saadeh et al., 2025). The problem is not simply psychological; it reflects a deeper shift in epistemic hierarchy where algorithmic reasoning displaces human interpretive authority.

Overreliance on AI recommendations can erode critical engagement with clinical evidence (Klingbeil et al., 2024). Physicians might shortcut their reasoning processes, assuming that the model's suggestion has already integrated all relevant data (Cross et al., 2024; Klingbeil et al., 2024). Conversely, when AI systems are known to make errors, excessive scepticism may lead to systematic disregard of useful insights—a phenomenon known as *automation aversion (Kostick-Quenet & Gerke, 2022)*. Balancing these extremes requires deliberate epistemic discipline and institutional mechanisms that preserve reflective judgment.

3.4. The Problem of Accountability and Legal Liability

Perhaps the most contentious risk domain concerns responsibility for harm when AI contributes to diagnostic error (Cestonaro et al., 2023; Contaldo et al., 2024). Current legal frameworks assign ultimate accountability to the licensed clinician, regardless of the degree of machine involvement (Jones et al., 2023; Maliha et al., 2021). Yet this arrangement becomes ethically fragile when the AI system's internal logic is inscrutable even to its developers. Physicians may thus be held liable for decisions shaped by models whose reasoning they cannot reconstruct.

This tension exposes a structural mismatch between existing doctrines of medical liability and the distributed nature of AI-mediated decision-making. Hospitals, software vendors, and data providers all contribute to the diagnostic process, yet the chain of accountability remains unarticulated. Without clear delineation of responsibility, both clinicians and patients inhabit a zone of epistemic uncertainty. Moreover, legal ambiguity can foster defensive approach to practicing medicine or reluctance to adopt beneficial innovations, thereby constraining the very progress AI was meant to deliver (Maliha et al., 2021).

The risk landscape of AI-assisted diagnostics, thus encompasses more than technical reliability. It includes cognitive, ethical, and institutional vulnerabilities that arise when probabilistic systems participate in normative acts of medical judgment. The following section addresses these vulnerabilities within the broader ethical and professional frameworks that govern the practice of medicine.

4. Ethical and Professional Implications

The ethical implications of integrating artificial intelligence into diagnostic medicine extend beyond the management of technological risk; they concern the reconfiguration of moral agency and the professional identity of the physician (Ackerhans et al., 2025; Dankwa-Mullan, 2024; Elgin & Elgin, 2024). Medicine is a domain historically defined by fiduciary duty—the moral obligation of the physician to exercise independent judgment for the patient's welfare. As AI systems enter diagnostic reasoning, this duty becomes shared, yet asymmetrically so, between human and machine. The ensuing moral configuration challenges established notions of responsibility, trust, and professional integrity.

4.1. Responsibility and Informed Consent in AI-Mediated Decisions

At the heart of the ethical dilemma lies the distribution of responsibility. When an AI system contributes to a diagnostic conclusion, the physician's decision is no longer purely autonomous. It becomes partially shaped by algorithmic inference, which may be correct, biased, or flawed in ways imperceptible to the clinician. Nevertheless, legal and professional accountability remain anchored to the human practitioner (Jones et al., 2023; Maliha et al., 2021). This discrepancy raises the question of whether moral responsibility presupposes epistemic control—can a clinician be fully responsible for an outcome generated by a process they cannot audit or comprehend? Informed consent further complicates this relationship. Traditional consent assumes that patients are informed about who, or what, is making diagnostic and therapeutic decisions. When AI systems influence these processes, the patient's understanding of agency becomes blurred. Ethically, it could be argued that patients should have a right to know when a diagnosis has been algorithmically assisted, and to what extent the recommendation reflects machine reasoning (H. J. Park, 2024; Ploug et al., 2025; Shaw et al., 2025). Failing to disclose this undermines patient autonomy and erodes trust in the therapeutic relationship. Yet few institutional protocols currently mandate explicit disclosure of AI participation in clinical decision-making (Bignami et al., 2025; Khosravi et al., 2024b).

4.2. Moral Agency, Trust, and the Patient–Doctor–AI Triad

The introduction of AI into diagnostics effectively creates a triadic relationship: patient, physician, and algorithm. Trust, once dyadic, now depends on how the physician mediates between human and non-human sources of authority. The physician must not only interpret the patient's condition but also interpret the algorithm's reasoning—or lack thereof. This dual interpretive burden redefines what it means to be a trustworthy clinician. If physicians rely too heavily on algorithmic outputs, they risk diminishing their role as moral agents who exercise judgment in the face of uncertainty. Conversely, rejecting AI assistance altogether can constitute a different form of negligence—one that disregards tools capable of improving diagnostic accuracy. The ethical ideal lies not in technological abstinence or blind adoption, but in cultivating *epistemic humility*: the recognition that both human and machine reasoning are fallible, and that good clinical practice involves managing, not eliminating, uncertainty.

Trust in AI-mediated diagnostics also depends on transparency and validation. Patients should be given assurance that algorithmic tools have been evaluated transparently and that their limitations are acknowledged. The moral obligation of the clinician extends to communicating these limitations clearly, without overstating the reliability or intelligence of the system. Overpersonalisation of AI—treating it as a quasi-expert or colleague—can foster misplaced trust and distort moral accountability (MacIntyre et al., 2023).

4.3. Ethical Boundaries of Algorithmic Assistance

The ethical boundary between assistance and delegation is particularly delicate. Assistance implies that AI supports the physician's reasoning; delegation implies partial transfer of that reasoning to the machine (Funer & Wiesing, 2024; Pham, 2025). In practice, the distinction is often blurred, especially when AI systems generate direct diagnostic statements rather than probabilistic cues. Once an algorithm offers a seemingly definitive conclusion, the cognitive pressure on the physician to either accept or reject it becomes ethically consequential (Funer & Wiesing, 2024).

Furthermore, algorithmic systems do not share human moral intuitions. They cannot prioritise patient dignity, contextual sensitivity, or compassion—values central to clinical ethics. Their outputs, no matter how sophisticated, remain instrumental rather than empathetic (Shen et al., 2024; Sirgiovanni, 2025). Ethical practice in AI-assisted medicine thus requires preserving human presence as the locus of empathy and judgment (M. K. Park et al., 2025). The physician's role is then not diminished but transformed: from an exclusive decision-maker to a moral interpreter of technological cognition.

Finally, institutions deploying diagnostic AI should acknowledge collective ethical responsibility. Hospitals and health systems bear moral obligations to ensure that deployed models are transparent, validated, and aligned with the ethical principles of beneficence, non-maleficence, autonomy, and justice (Gorelik et al., 2025; Pham, 2025; Weidener & Fischer, 2024). Failing to meet these obligations might converts technological innovation into ethical negligence.

5. Governance and Regulation

The rapid adoption of artificial intelligence in diagnostic medicine has far outpaced the establishment of corresponding governance structures. While technological capabilities expand with remarkable velocity, ethical, legal, and institutional frameworks remain fragmented and reactive (Bouderhem, 2024; Mennella et al., 2024; Nasir et al., 2025). This asymmetry between innovation and regulation poses profound challenges for patient safety, professional accountability, and societal trust in AI-mediated healthcare.

5.1. Current Regulatory Landscape for AI in Healthcare

At present, regulatory mechanisms governing AI in medicine are largely adapted from pre-existing frameworks for medical devices and software (Fraser et al., 2023; Santra et al., 2024). In many jurisdictions, AI diagnostic systems are categorised as *software as a medical device (SaMD)*, subject to evaluation based on safety, efficacy, and data protection standards (Health, 2025a, 2025b; Navarro, 2024). Such classifications, however, inadequately capture the adaptive, non-deterministic nature of machine learning models, particularly those capable of continuous learning or unsupervised pattern recognition (Onitiu et al., 2024; Pantanowitz et al., 2024).

Traditional regulatory paradigms assume that medical devices are static, their performance fixed at the time of approval. AI systems, by contrast, evolve as they process new data, potentially altering their diagnostic behaviour

post-deployment (Babic et al., 2025; Onitiu et al., 2024). This dynamic quality renders conventional approval models obsolete, as periodic re-certification or post-market surveillance may fail to detect subtle but clinically significant drift in algorithmic performance. Furthermore, many existing regulations emphasise product conformity but not epistemic transparency—compliance is assessed by procedural documentation rather than demonstrable interpretability. The situation is further complicated by the proliferation of *general-purpose* large language models adapted for medical use. Unlike domain-specific tools trained on curated clinical datasets, general LLMs often lack formal medical validation yet are widely employed by clinicians for drafting, summarising, or differential diagnosis (Busch et al., 2025; Nazi & Peng, 2024; Singhal et al., 2023). Regulatory agencies currently lack the authority or technical capacity to oversee such hybrid, continuously evolving systems (Health, 2025a; Muralidharan et al., 2024).

5.2. The Limits of Existing Medical Liability Frameworks

Medical liability systems presuppose a clear chain of causation between practitioner action and patient outcome. In AI-assisted diagnostics, this causal chain becomes diffused. A misdiagnosis may result from flawed training data, an opaque inference, or an inappropriate reliance on machine output. Yet existing legal norms hold the physician solely accountable, even when the underlying model is proprietary and inaccessible for forensic examination (Cestonaro et al., 2023; *Fault Lines in Health Care AI – Part Two*, 2025).

This legal asymmetry may disincentivise transparency. Developers, shielded by intellectual property protections, may restrict access to algorithmic details, leaving clinicians to bear the ethical and legal risk of using systems they cannot audit (Cestonaro et al., 2023; Lawton et al., 2024). Hospitals, in turn, adopt AI tools under institutional pressure to modernise, often without fully understanding their operational boundaries. The result is a regulatory vacuum in which responsibility is fragmented among actors such as physicians, developers, data providers, and institutions, without any coherent framework for shared liability. A reformed governance approach must therefore move beyond individual culpability to embrace *distributed accountability*. This would entail legally recognising that diagnostic outcomes in AI-mediated environments emerge from socio-technical systems, not isolated agents. Ethical responsibility should be proportionally allocated among stakeholders according to their degree of epistemic control and capacity to prevent harm.

6. Discussion

The emergence of artificial intelligence in diagnostic medicine signifies not merely a technological innovation but a paradigmatic reordering of epistemic and moral structures within healthcare. The preceding sections have outlined how LLMs and other AI systems introduce new forms of diagnostic reasoning, new loci of error, and new challenges for ethical and legal accountability. This discussion section synthesises these insights to evaluate how medicine must adapt to preserve its normative foundations—truthfulness, responsibility, and professional integrity—while embracing the benefits of computational intelligence.

The central paradox of AI in medicine is that the same properties that make these systems powerful—autonomy, adaptivity, and scalability—also make them epistemically fragile. AI can detect subtle statistical regularities across massive datasets, but it cannot situate those regularities within causal or moral frameworks. The medical profession, however, is founded on causal reasoning and ethical deliberation. Integrating AI thus requires a dual commitment: to innovation as a means of improving care, and to responsibility as a safeguard of meaning and trust.

Physicians should remain the primary custodians of diagnostic interpretation, not because humans outperform machines in statistical inference, but because human reasoning is embedded in moral context. The legitimacy of clinical judgment stems from its accountability to both evidence and ethical norms—something AI cannot replicate. To sustain this legitimacy, clinicians should cultivate a reflective stance toward algorithmic assistance: using AI as a tool for hypothesis generation rather than as an epistemic authority. Institutional frameworks must reinforce this stance through training programs that emphasise critical literacy about AI outputs, cognitive biases, and uncertainty management.

Moreover, AI-assisted diagnostics compel a redefinition of what it means to be an expert. Traditionally, expertise implied mastery of knowledge and the capacity to reason through complex uncertainty. In AI-mediated contexts, expertise increasingly involves the ability to evaluate and contextualise algorithmic reasoning. The competent physician of the near future must understand not only disease mechanisms but also model architectures, data provenance, and the interpretive limits of statistical inference. This epistemic expansion suggests that clinical

expertise will become more interdisciplinary, drawing on informatics, ethics, and systems theory. Yet there is a risk that the medical profession could devolve into a supervisory role—technically overseeing systems that perform the substantive cognitive work of diagnosis. To prevent such displacement, medical education should integrate *algorithmic epistemology*—a structured understanding of how AI systems know, err, and evolve. Only by mastering this knowledge can clinicians preserve their agency as interpreters rather than custodians of machine output.

Finally, epistemic humility is emerging as a vital virtue in AI-mediated medicine. It entails recognising the partiality of both human and algorithmic knowledge. It can be argued that physicians should resist the seduction of technological omniscience, while AI developers should acknowledge the social and moral dimensions of diagnostic reasoning. Transparency, both technical and communicative, becomes the ethical counterpart of humility. For AI systems, transparency requires intelligible documentation of training data, model scope, and known failure modes. For physicians, transparency involves disclosing to patients when AI systems have contributed to diagnostic reasoning and articulating the degree of confidence and uncertainty involved. For institutions, transparency means cultivating environments where algorithmic errors can be reported without punitive consequence, thereby transforming individual blame into collective learning.

The broader societal dimension of transparency concerns trust. Public trust in medicine has historically rested on the assumption of human accountability—someone who can explain, justify, and empathise. As AI increasingly mediates diagnosis, sustaining that trust demands mechanisms through which algorithmic processes become morally legible. This does not mean anthropomorphising AI, but rather embedding it within systems of human explanation and oversight.

7. Conclusion

Artificial intelligence has become an indispensable instrument in the evolution of diagnostic medicine, yet its integration exposes deep tensions between technological capability and epistemic legitimacy. The medical act of diagnosis has never been purely technical; it has always been an interpretive, moral, and social process grounded in human judgment and responsibility. By introducing systems that can simulate diagnostic reasoning without understanding, AI disrupts the foundations upon which that process rests. This review has argued that the diagnostic use of large language models and related AI systems represents both a remarkable advance and a profound epistemological challenge. These models expand the reach of diagnostic inference by uncovering correlations invisible to human cognition, but they also erode the transparency and accountability that lend medicine its moral authority. The physician remains legally and ethically accountable, yet increasingly relies on outputs generated by algorithms whose internal logic cannot be meaningfully inspected or contested. This asymmetry produces a new form of clinical vulnerability: responsibility without full comprehension. The key insight emerging from this analysis is that the successful coexistence of human and artificial intelligence in medicine depends not on technological perfection but on institutional wisdom. AI should remain a tool of reasoning, not a substitute for it. Its diagnostic suggestions should be treated as probabilistic hypotheses, always subject to the clinician's contextual judgment. Preserving this hierarchy requires an ethical infrastructure that prioritises transparency, continuous oversight, and distributive accountability among all actors in the diagnostic chain. At the epistemological level, the rise of AI calls for a renewed humility within medical science—a recognition that neither human expertise nor algorithmic intelligence possesses complete knowledge of disease. The task ahead is to construct systems in which human and machine reasoning complement, rather than compete with, each other. This balance can be achieved only if medicine resists the temptation to conflate predictive accuracy with understanding, and if it safeguards the interpretive and moral dimensions of diagnosis as distinctly human responsibilities.

Author Contributions

Conceptualisation, M.M. and L.L.; methodology, M.M.; validation, L.L., M.M; investigation, L.L., M.M; data curation, L.L. and M.M; writing—original draft preparation, M.M; writing—review and editing, L.L; supervision, L.L. All authors have read and agreed to the published version of the manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare no conflict of interest.

Literature

1. Abdelwanis, M., Alarafati, H. K., Tammam, M. M. S., & Simsekler, M. C. E. (2024). Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis. *Journal of Safety Science and Resilience*, *5*(4), 460–469. https://doi.org/10.1016/j.jnlssr.2024.06.001
2. Ackerhans, S., Wehkamp, K., Petzina, R., Dumitrescu, D., & Schultz, C. (2025). Perceived Trust and Professional Identity Threat in AI-Based Clinical Decision Support Systems: Scenario-Based Experimental Study on AI Process Design Features. *JMIR Formative Research*, *9*(1), e64266. https://doi.org/10.2196/64266
3. Acun, C., & Nasraoui, O. (2025). Pre Hoc and Co Hoc Explainability: Frameworks for Integrating Interpretability into Machine Learning Training for Enhanced Transparency and Performance. *Applied Sciences*, *15*(13), 7544. https://doi.org/10.3390/app15137544
4. Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications*, *11*(1), 1568. https://doi.org/10.1057/s41599-024-04044-8
5. Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., & Gao, G. (2023). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy and Technology*, *12*(1), 100702. https://doi.org/10.1016/j.hlpt.2022.100702
6. Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education*, *23*(1), 689. https://doi.org/10.1186/s12909-023-04698-z
7. Angus, D. C., Khera, R., Lieu, T., Liu, V., Ahmad, F. S., Anderson, B., Bhavani, S. V., Bindman, A., Brennan, T., Celi, L. A., Chen, F., Cohen, I. G., Denniston, A., Desai, S., Embí, P., Faisal, A., Ferryman, K., Gerhart, J., Gross, M., … JAMA Summit on AI. (2025). AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence. *JAMA*. https://doi.org/10.1001/jama.2025.18490
8. Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J. A., & Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *Npj Digital Medicine*, *8*(1), 274. https://doi.org/10.1038/s41746-025-01670-7
9. Babic, B., Glenn Cohen, I., Stern, A. D., Li, Y., & Ouellet, M. (2025). A general framework for governing marketed AI/ML medical devices. *Npj Digital Medicine*, *8*(1), 328. https://doi.org/10.1038/s41746-025-01717-9
10. Bignami, E., Darhour, L. J., Franco, G., Guarnieri, M., & Bellini, V. (2025). AI policy in healthcare: A checklist-based methodology for structured implementation. *Journal of Anesthesia, Analgesia and Critical Care*, *5*, 56. https://doi.org/10.1186/s44158-025-00278-3
11. Boge, F., & Mosig, A. (2025). Causality and scientific explanation of artificial intelligence systems in biomedicine. *Pflügers Archiv - European Journal of Physiology*, *477*(4), 543–554. https://doi.org/10.1007/s00424-024-03033-9
12. Bouderhem, R. (2024). Shaping the future of AI in healthcare through ethics and governance. *Humanities and Social Sciences Communications*, *11*(1), 416. https://doi.org/10.1057/s41599-024-02894-w
13. Bozyel, S., Şimşek, E., Koçyiğit Burunkaya, D., Güler, A., Korkmaz, Y., Şeker, M., Ertürk, M., & Keser, N. (2024). Artificial Intelligence-Based Clinical Decision Support Systems in Cardiovascular Diseases. *The Anatolian Journal of Cardiology*, *28*(2), 74. https://doi.org/10.14744/AnatolJCardiol.2023.3685
14. Busch, F., Hoffmann, L., Rueger, C., van Dijk, E. H., Kader, R., Ortiz-Prado, E., Makowski, M. R., Saba, L., Hadamitzky, M., Kather, J. N., Truhn, D., Cuocolo, R., Adams, L. C., & Bressem, K. K. (2025). Current applications and challenges in large language models for patient care: A systematic review. *Communications Medicine*, *5*(1), 26. https://doi.org/10.1038/s43856-024-00717-2
15. Celi, L. A., Cellini, J., Charpignon, M.-L., Dee, E. C., Dernoncourt, F., Eber, R., Mitchell, W. G., Moukheiber, L., Schirmer, J., Situ, J., Paguio, J., Park, J., Wawira, J. G., Yao, S., & Data, for M. C.

(2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health*, *1*(3), e0000022. https://doi.org/10.1371/journal.pdig.0000022

16. Cestonaro, C., Delicati, A., Marcante, B., Caenazzo, L., & Tozzo, P. (2023). Defining medical liability when artificial intelligence is applied on diagnostic algorithms: A systematic review. *Frontiers in Medicine*, *10*, 1305756. https://doi.org/10.3389/fmed.2023.1305756
17. Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research*, *26*(1), e53164. https://doi.org/10.2196/53164
18. Chen, H., Gomez, C., Huang, C.-M., & Unberath, M. (2022). Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. *Npj Digital Medicine*, *5*(1), 156. https://doi.org/10.1038/s41746-022-00699-2
19. Chustecki, M. (2024). Benefits and Risks of AI in Health Care: Narrative Review. *Interactive Journal of Medical Research*, *13*, e53616. https://doi.org/10.2196/53616
20. Colacci, M., Huang, Y. Q., Postill, G., Zhelnov, P., Fennelly, O., Verma, A., Straus, S., & Tricco, A. C. (2025). Sociodemographic bias in clinical machine learning models: A scoping review of algorithmic bias instances and mechanisms. *Journal of Clinical Epidemiology*, *178*, 111606. https://doi.org/10.1016/j.jclinepi.2024.111606
21. Contaldo, M. T., Pasceri, G., Vignati, G., Bracchi, L., Triggiani, S., & Carrafiello, G. (2024). AI in Radiology: Navigating Medical Responsibility. *Diagnostics*, *14*(14), 1506. https://doi.org/10.3390/diagnostics14141506
22. Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, *3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651
23. D'Adderio, L., & Bates, D. W. (2025). Transforming diagnosis through artificial intelligence. *NPJ Digital Medicine*, *8*(1), 54. https://doi.org/10.1038/s41746-025-01460-1
24. Dankwa-Mullan, I. (2024). Health Equity and Ethical Considerations in Using Artificial Intelligence in Public Health and Medicine. *Preventing Chronic Disease*, *21*. https://doi.org/10.5888/pcd21.240245
25. DeGroat, W., Abdelhalim, H., Peker, E., Sheth, N., Narayanan, R., Zeeshan, S., Liang, B. T., & Ahmed, Z. (2024). Multimodal AI/ML for discovering novel biomarkers and predicting disease using multi-omics profiles of patients with cardiovascular diseases. *Scientific Reports*, *14*(1), 26503. https://doi.org/10.1038/s41598-024-78553-6
26. Elgin, C. Y., & Elgin, C. (2024). Ethical implications of AI-driven clinical decision support systems on healthcare resource allocation: A qualitative study of healthcare professionals' perspectives. *BMC Medical Ethics*, *25*, 148. https://doi.org/10.1186/s12910-024-01151-8
27. Fahim, Y. A., Hasani, I. W., Kabba, S., & Ragab, W. M. (2025). Artificial intelligence in healthcare and medicine: Clinical applications, therapeutic advances, and future perspectives. *European Journal of Medical Research*, *30*(1), 848. https://doi.org/10.1186/s40001-025-03196-w
28. *Fault lines in health care AI – Part two: Who's responsible when AI gets it wrong?* (2025, June 26). https://carey.jhu.edu/articles/fault-lines-health-care-ai-part-two-whos-responsible-when-ai-gets-it-wrong
29. Fraser, A. G., Biasin, E., Bijnens, B., Bruining, N., Caiani, E. G., Cobbaert, K., Davies, R. H., Gilbert, S. H., Hovestadt, L., Kamenjasevic, E., Kwade, Z., McGauran, G., O'Connor, G., Vasey, B., & Rademakers, F. E. (2023). Artificial intelligence in medical device software and high-risk medical devices – a review of definitions, expert recommendations and regulatory initiatives. *Expert Review of Medical Devices*, *20*(6), 467–491. https://doi.org/10.1080/17434440.2023.2184685
30. Freyer, N., Groß, D., & Lipprandt, M. (2024). The ethical requirement of explainability for AI-DSS in healthcare: A systematic review of reasons. *BMC Medical Ethics*, *25*(1), 104. https://doi.org/10.1186/s12910-024-01103-2
31. Funer, F., & Wiesing, U. (2024). Physician's autonomy in the face of AI support: Walking the ethical tightrope. *Frontiers in Medicine*, *11*. https://doi.org/10.3389/fmed.2024.1324963
32. Goh, E., Bunning, B., Khoong, E. C., Gallo, R. J., Milstein, A., Centola, D., & Chen, J. H. (2025). Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. *Communications Medicine*, *5*(1), 59. https://doi.org/10.1038/s43856-025-00781-2
33. Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H.

(2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, *7*(10), e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

34. Gorelik, A. J., Li, M., Hahne, J., Wang, J., Ren, Y., Yang, L., Zhang, X., Liu, X., Wang, X., Bogdan, R., & Carpenter, B. D. (2025). Ethics of AI in healthcare: A scoping review demonstrating applicability of a foundational framework. *Frontiers in Digital Health*, *7*. https://doi.org/10.3389/fdgth.2025.1662642
35. Graili, P., & Farhoudi, B. (2025). The intersection of digital health and artificial intelligence: Clearing the cloud of uncertainty. *DIGITAL HEALTH*, *11*, 20552076251315621. https://doi.org/10.1177/20552076251315621
36. Greenhalgh, T., Thorne, S., & Malterud, K. (2018). Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation*, *48*(6), e12931. https://doi.org/10.1111/eci.12931
37. Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., & Rashidi, H. H. (2025). Ethical and Bias Considerations in Artificial Intelligence/Machine Learning. *Modern Pathology*, *38*(3). https://doi.org/10.1016/j.modpat.2024.100686
38. Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., & White, J. A. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *Npj Digital Medicine*, *8*(1), 154. https://doi.org/10.1038/s41746-025-01503-7
39. Health, C. for D. and R. (2025a). Artificial Intelligence in Software as a Medical Device. *FDA*. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device
40. Health, C. for D. and R. (2025b, July 23). *Software as a Medical Device (SaMD)*. FDA; FDA. https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd
41. Jeong, H. K., Park, C., Henao, R., & Kheterpal, M. (2022). Deep Learning in Dermatology: A Systematic Review of Current Approaches, Outcomes, and Limitations. *JID Innovations*, *3*(1), 100150. https://doi.org/10.1016/j.xjidi.2022.100150
42. Jones, C., Thornton, J., & Wyatt, J. C. (2023). Artificial intelligence and clinical decision support: Clinicians' perspectives on trust, trustworthiness, and liability. *Medical Law Review*, *31*(4), 501–520. https://doi.org/10.1093/medlaw/fwad013
43. Joseph, J. (2025). Algorithmic bias in public health AI: A silent threat to equity in low-resource settings. *Frontiers in Public Health*, *13*. https://doi.org/10.3389/fpubh.2025.1643180
44. Jui, T. D., & Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, *15*(8), 3095–3125. https://doi.org/10.1007/s13042-023-02083-2
45. Khosravi, M., Zare, Z., Mojtabaeian, S. M., & Izadi, R. (2024a). Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews. *Health Services Research and Managerial Epidemiology*, *11*, 23333928241234863. https://doi.org/10.1177/23333928241234863
46. Khosravi, M., Zare, Z., Mojtabaeian, S. M., & Izadi, R. (2024b). Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews. *Health Services Research and Managerial Epidemiology*, *11*, 23333928241234863. https://doi.org/10.1177/23333928241234863
47. Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352. https://doi.org/10.1016/j.chb.2024.108352
48. Koçak, B., Ponsiglione, A., Stanzione, A., Bluethgen, C., Santinha, J., Ugga, L., Huisman, M., Klontzas, M. E., Cannella, R., & Cuocolo, R. (2025). Bias in artificial intelligence for medical imaging: Fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and Interventional Radiology*, *31*(2), 75–88. https://doi.org/10.4274/dir.2024.242854
49. Korfiatis, P., Kline, T. L., Meyer, H. M., Khalid, S., Leiner, T., Loufek, B. T., Blezek, D., Vidal, D. E., Hartman, R. P., Joppa, L. J., Missert, A. D., Potretzke, T. A., Taubel, J. P., Tjelta, J. A., Callstrom, M. R., & Williamson, E. E. (2025). Implementing Artificial Intelligence Algorithms in the Radiology Workflow: Challenges and Considerations. *Mayo Clinic Proceedings: Digital Health*, *3*(1). https://doi.org/10.1016/j.mcpdig.2024.100188

50. Kostick-Quenet, K. M., & Gerke, S. (2022). AI in the hands of imperfect users. *Npj Digital Medicine*, *5*(1), 197. https://doi.org/10.1038/s41746-022-00737-z
51. Krakowski, I., Kim, J., Cai, Z. R., Daneshjou, R., Lapins, J., Eriksson, H., Lykou, A., & Linos, E. (2024). Human-AI interaction in skin cancer diagnosis: A systematic review and meta-analysis. *Npj Digital Medicine*, *7*(1), 78. https://doi.org/10.1038/s41746-024-01031-w
52. Krishnan, G., Singh, S., Pathania, M., Gosavi, S., Abhishek, S., Parchani, A., & Dhar, M. (2023). Artificial intelligence in clinical medicine: Catalyzing a sustainable global healthcare paradigm. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1227091
53. Kulikowski, C. A. (2015). An Opening Chapter of the First Generation of Artificial Intelligence in Medicine: The First Rutgers AIM Workshop, June 1975. *Yearbook of Medical Informatics*, *10*(1), 227–233. https://doi.org/10.15265/IY-2015-016
54. Lawton, T., Morgan, P., Porter, Z., Hickey, S., Cunningham, A., Hughes, N., Iacovides, I., Jia, Y., Sharma, V., & Habli, I. (2024). Clinicians risk becoming 'liability sinks' for artificial intelligence. *Future Healthcare Journal*, *11*(1), 100007. https://doi.org/10.1016/j.fhj.2024.100007
55. MacIntyre, M. R., Cockerill, R. G., Mirza, O. F., & Appel, J. M. (2023). Ethical considerations for the use of artificial intelligence in medical decision-making capacity assessments. *Psychiatry Research*, *328*, 115466. https://doi.org/10.1016/j.psychres.2023.115466
56. Maliha, G., Gerke, S., Cohen, I. G., & Parikh, R. B. (2021). Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation. *The Milbank Quarterly*, *99*(3), 629–647. https://doi.org/10.1111/1468-0009.12504
57. McGenity, C., Clarke, E. L., Jennings, C., Matthews, G., Cartlidge, C., Freduah-Agyemang, H., Stocken, D. D., & Treanor, D. (2024). Artificial intelligence in digital pathology: A systematic review and meta-analysis of diagnostic test accuracy. *Npj Digital Medicine*, *7*(1), 114. https://doi.org/10.1038/s41746-024-01106-8
58. Mennella, C., Maniscalco, U., De Pietro, G., & Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*, *10*(4), e26297. https://doi.org/10.1016/j.heliyon.2024.e26297
59. Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). Bias in AI-based models for medical applications: Challenges and mitigation strategies. *Npj Digital Medicine*, *6*(1), 113. https://doi.org/10.1038/s41746-023-00858-z
60. Muralidharan, V., Adewale, B. A., Huang, C. J., Nta, M. T., Ademiju, P. O., Pathmarajah, P., Hang, M. K., Adesanya, O., Abdullateef, R. O., Babatunde, A. O., Ajibade, A., Onyeka, S., Cai, Z. R., Daneshjou, R., & Olatunji, T. (2024). A scoping review of reporting gaps in FDA-approved AI medical devices. *Npj Digital Medicine*, *7*(1), 273. https://doi.org/10.1038/s41746-024-01270-x
61. Nasir, M., Siddiqui, K., & Ahmed, S. (2025). Ethical-legal implications of AI-powered healthcare in critical perspective. *Frontiers in Artificial Intelligence*, *8*, 1619463. https://doi.org/10.3389/frai.2025.1619463
62. Navarro, A. (2024, September 25). EU MDR and IVDR: Classifying Medical Device Software (MDSW). *NAMSA*. https://namsa.com/resources/blog/eu-mdr-and-ivdr-classifying-medical-device-software-mdsw/
63. Nazi, Z. A., & Peng, W. (2024). Large Language Models in Healthcare and Medical Domain: A Review. *Informatics*, *11*(3), 57. https://doi.org/10.3390/informatics11030057
64. Nicholson, N., Giusti, F., & Martos, C. (2023). Ontology-Based AI Design Patterns and Constraints in Cancer Registry Data Validation. *Cancers*, *15*(24), 5812. https://doi.org/10.3390/cancers15245812
65. Onitiu, D., Wachter, S., & Mittelstadt, B. (2024). How AI challenges the medical device regulation: Patient safety, benefits, and intended uses. *Journal of Law and the Biosciences*, lsae007. https://doi.org/10.1093/jlb/lsae007
66. Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., & Nascimento, E. G. S. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing*, *7*(1), 15. https://doi.org/10.3390/bdcc7010015
67. Pantanowitz, L., Hanna, M., Pantanowitz, J., Lennerz, J., Henricks, W. H., Shen, P., Quinn, B., Bennet, S., & Rashidi, H. H. (2024). Regulatory Aspects of Artificial Intelligence and Machine Learning. *Modern Pathology*, *37*(12). https://doi.org/10.1016/j.modpat.2024.100609

68. Park, H. J. (2024). Patient perspectives on informed consent for medical AI: A web-based experiment. *DIGITAL HEALTH*, *10*, 20552076241247938. https://doi.org/10.1177/20552076241247938
69. Park, M. K., Ashwood, N., Capes, N., Park, M. K., Ashwood, N., & Capes, N. (2025). Ethics of Artificial Intelligence in Medicine. *Cureus*, *17*(5). https://doi.org/10.7759/cureus.83567
70. Patil, S. V., Myers, C. G., & Lu-Myers, Y. (2025). Calibrating AI Reliance—A Physician's Superhuman Dilemma. *JAMA Health Forum*, *6*(3), e250106. https://doi.org/10.1001/jamahealthforum.2025.0106
71. Pham, T. (2025). Ethical and legal considerations in healthcare AI: Innovation and policy for safe and fair use. *Royal Society Open Science*, *12*(5), 241873. https://doi.org/10.1098/rsos.241873
72. Ploug, T., Jørgensen, R. F., Motzfeldt, H. M., Ploug, N., & Holm, S. (2025). The need for patient rights in AI-driven healthcare – risk-based regulation is not enough. *Journal of the Royal Society of Medicine*, 01410768251344707. https://doi.org/10.1177/01410768251344707
73. Rao, A., & Aalami, O. (2023). Towards improving the visual explainability of artificial intelligence in the clinical setting. *BMC Digital Health*, *1*(1), 23. https://doi.org/10.1186/s44247-023-00022-3
74. Rezaeian, O., Bayrak, A. E., & Asan, O. (2025). *Explainability and AI Confidence in Clinical Decision Support Systems: Effects on Trust, Diagnostic Performance, and Cognitive Load in Breast Cancer Care* (No. arXiv:2501.16693). arXiv. https://doi.org/10.48550/arXiv.2501.16693
75. Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. *JMIR AI*, *3*(1), e53207. https://doi.org/10.2196/53207
76. Saadeh, M. I., Janhonen, J., Beer, E., Castelyn, C., & Hoffman, D. N. (2025). Automation complacency: Risks of abdicating medical decision making. *AI and Ethics*. https://doi.org/10.1007/s43681-025-00825-2
77. Saarela, M., & Podgorelec, V. (2024). Recent Applications of Explainable AI (XAI): A Systematic Literature Review. *Applied Sciences*, *14*(19), 8884. https://doi.org/10.3390/app14198884
78. Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S., & Pardalos, P. M. (2024). A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, *118*, 109370. https://doi.org/10.1016/j.compeleceng.2024.109370
79. Sadr, H., Nazari, M., Khodaverdian, Z., Farzan, R., Yousefzadeh-Chabok, S., Ashoobi, M. T., Hemmati, H., Hendi, A., Ashraf, A., Pedram, M. M., Hasannejad-Bibalan, M., & Yamaghani, M. R. (2025). Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: A comprehensive review of machine learning and deep learning approaches. *European Journal of Medical Research*, *30*(1), 418. https://doi.org/10.1186/s40001-025-02680-7
80. Saenz, A. D., Harned, Z., Banerjee, O., Abràmoff, M. D., & Rajpurkar, P. (2023). Autonomous AI systems in the face of liability, regulations and costs. *NPJ Digital Medicine*, *6*(1), 185. https://doi.org/10.1038/s41746-023-00929-1
81. Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q., & Tsaftaris, S. A. (2022). Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, *9*(8), 220638. https://doi.org/10.1098/rsos.220638
82. Santra, S., Kukreja, P., Saxena, K., Gandhi, S., & Singh, O. V. (2024). Navigating regulatory and policy challenges for AI enabled combination devices. *Frontiers in Medical Technology*, *6*, 1473350. https://doi.org/10.3389/fmedt.2024.1473350
83. Savage, T., Nayak, A., Gallo, R., Rangan, E., & Chen, J. H. (2024). Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *Npj Digital Medicine*, *7*(1), 20. https://doi.org/10.1038/s41746-024-01010-1
84. Shaw, D., Lorenzini, G., Ossa, L. A., Eckstein, J., Steiner, L., & Elger, B. S. (2025). When and what patients need to know about AI in clinical care. *Swiss Medical Weekly*, *155*(1), 4013–4013. https://doi.org/10.57187/s.4013
85. Shen, J., DiPaola, D., Ali, S., Sap, M., Park, H. W., & Breazeal, C. (2024). Empathy Toward Artificial Intelligence Versus Human Experiences and the Role of Transparency in Mental Health and Social Support Chatbot Design: Comparative Study. *JMIR Mental Health*, *11*, e62679. https://doi.org/10.2196/62679

86. Shin, H. S. (2019). Reasoning processes in clinical reasoning: From the perspective of cognitive psychology. *Korean Journal of Medical Education*, *31*(4), 299–308. https://doi.org/10.3946/kjme.2019.140
87. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2
88. Sinha, R. (2024). The role and impact of new technologies on healthcare systems. *Discover Health Systems*, *3*(1), 96. https://doi.org/10.1007/s44250-024-00163-w
89. Sirgiovanni, E. (2025). Should Doctor Robot possess moral empathy? *Bioethics*, *39*(1), 98–107. https://doi.org/10.1111/bioe.13345
90. Solutions, M. (2024, June 6). Evolution of AI in Healthcare and Clinical Research. *Medidata Solutions*. https://www.medidata.com/en/life-science-resources/medidata-blog/evolution-of-ai-in-healthcare-and-clinical-research/
91. Sourlos, N., Vliegenthart, R., Santinha, J., Klontzas, M. E., Cuocolo, R., Huisman, M., & van Ooijen, P. (2024). Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology. *Insights into Imaging*, *15*(1), 248. https://doi.org/10.1186/s13244-024-01833-2
92. Stinson, C. (2022). Algorithms are not neutral. *AI and Ethics*, *2*(4), 763–770. https://doi.org/10.1007/s43681-022-00136-w
93. Sukhera, J. (2022). Narrative Reviews: Flexible, Rigorous, and Practical. *Journal of Graduate Medical Education*, *14*(4), 414–417. https://doi.org/10.4300/JGME-D-22-00480.1
94. Tejani, A. S., Ng, Y. S., Xi, Y., & Rayan, J. C. (2024). Understanding and Mitigating Bias in Imaging Artificial Intelligence. *RadioGraphics*, *44*(5), e230067. https://doi.org/10.1148/rg.230067
95. Tikhomirov, L., Semmler, C., McCradden, M., Searston, R., Ghassemi, M., & Oakden-Rayner, L. (2024). Medical artificial intelligence for clinicians: The lost cognitive perspective. *The Lancet Digital Health*, *6*(8), e589–e594. https://doi.org/10.1016/S2589-7500(24)00095-5
96. Tun, H. M., Rahman, H. A., Naing, L., & Malik, O. A. (2025). Trust in Artificial Intelligence–Based Clinical Decision Support Systems Among Health Care Workers: Systematic Review. *Journal of Medical Internet Research*, *27*, e69678. https://doi.org/10.2196/69678
97. Ullah, E., Parwani, A., Baig, M. M., & Singh, R. (2024). Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagnostic Pathology*, *19*(1), 43. https://doi.org/10.1186/s13000-024-01464-7
98. van Diest, P. J., Flach, R. N., van Dooijeweert, C., Makineli, S., Breimer, G. E., Stathonikos, N., Pham, P., Nguyen, T. Q., & Veta, M. (2024). Pros and cons of artificial intelligence implementation in diagnostic pathology. *Histopathology*, *84*(6), 924–934. https://doi.org/10.1111/his.15153
99. Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M., & Božić, J. (2025). A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. *Healthcare*, *13*(6), 603. https://doi.org/10.3390/healthcare13060603
100. Vrudhula, A., Kwan, A. C., Ouyang, D., & Cheng, S. (2024). Machine Learning and Bias in Medical Imaging: Opportunities and Challenges. *Circulation: Cardiovascular Imaging*, *17*(2), e015495. https://doi.org/10.1161/CIRCIMAGING.123.015495
101. Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields: Applications, advances, and challenges. *Artificial Intelligence Review*, *57*(11), 299. https://doi.org/10.1007/s10462-024-10921-0
102. Waqas, A., Bui, M. M., Glassy, E. F., Naqa, I. E., Borkowski, P., Borkowski, A. A., & Rasool, G. (2023). Revolutionizing Digital Pathology With the Power of Generative Artificial Intelligence and Foundation Models. *Laboratory Investigation*, *103*(11). https://doi.org/10.1016/j.labinv.2023.100255
103. Weidener, L., & Fischer, M. (2024). Role of Ethics in Developing AI-Based Applications in Medicine: Insights From Expert Interviews and Discussion of Implications. *JMIR AI*, *3*(1), e51204. https://doi.org/10.2196/51204

104. Wellnhofer, E. (2022). Real-World and Regulatory Perspectives of Artificial Intelligence in Cardiovascular Imaging. *Frontiers in Cardiovascular Medicine*, *9*. https://doi.org/10.3389/fcvm.2022.890809
105. Wu, K., Wu, E., Theodorou, B., Liang, W., Mack, C., Glass, L., Sun, J., & Zou, J. (2024). Characterizing the Clinical Adoption of Medical AI Devices through U.S. Insurance Claims. *NEJM AI*, *1*(1), AIoa2300030. https://doi.org/10.1056/AIoa2300030
106. Xu, H., & Shuttleworth, K. M. J. (2024). Medical artificial intelligence and the black box problem: A view based on the ethical principle of "do no harm." *Intelligent Medicine*, *4*(1), 52–57. https://doi.org/10.1016/j.imed.2023.08.001
107. Xue, C., Kowshik, S. S., Lteif, D., Puducheri, S., Jasodanand, V. H., Zhou, O. T., Walia, A. S., Guney, O. B., Zhang, J. D., Poésy, S., Kaliaev, A., Andreu-Arasa, V. C., Dwyer, B. C., Farris, C. W., Hao, H., Kedar, S., Mian, A. Z., Murman, D. L., O'Shea, S. A., … Kolachalama, V. B. (2024). AI-based differential diagnosis of dementia etiologies on multimodal data. *Nature Medicine*, *30*(10), 2977–2989. https://doi.org/10.1038/s41591-024-03118-z
108. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D., & Ghassemi, M. (2024). The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, *30*(10), 2838–2848. https://doi.org/10.1038/s41591-024-03113-4
109. Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, *16*(3), e12962. https://doi.org/10.1111/soc4.12962
110. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.*, *15*(2), 20:1-20:38. https://doi.org/10.1145/3639372
111. Zhou, S., Xu, Z., Zhang, M., Xu, C., Guo, Y., Zhan, Z., Fang, Y., Ding, S., Wang, J., Xu, K., Xia, L., Yeung, J., Zha, D., Cai, D., Melton, G. B., Lin, M., & Zhang, R. (2025). Large language models for disease diagnosis: A scoping review. *Npj Artificial Intelligence*, *1*(1), 9. https://doi.org/10.1038/s44387-025-00011-z