International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	ALGORITHMIC AUTHORITY VS. HUMAN TOUCH: A NARRATIVE REVIEW OF PATIENT TRUST AND CLINICAL AUTONOMY IN AI-ASSISTED DIAGNOSTICS
DOI	https://doi.org/10.31435/ijitss.1(49).2026.4963
RECEIVED	05 February 2026
ACCEPTED	20 March 2026
PUBLISHED	25 March 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

ALGORITHMIC AUTHORITY VS. HUMAN TOUCH: A NARRATIVE REVIEW OF PATIENT TRUST AND CLINICAL AUTONOMY IN AI-ASSISTED DIAGNOSTICS

Maciej Kokoszka *(Corresponding Author, Email: kokoszkamaciejj@gmail.com)*
7th Navy Hospital, Gdańsk, Poland
ORCID ID: 0009-0009-7682-9767

Michalina Chodór
7th Navy Hospital, Gdańsk, Poland
ORCID ID: 0009-0000-2541-3686

Julia Maria Kuczkowska
7th Navy Hospital, Gdańsk, Poland
ORCID ID: 0009-0004-8320-876X

Judyta Bordakiewicz
7th Navy Hospital, Gdańsk, Poland
ORCID ID: 0009-0001-7286-0662

Zuzanna Michalska
Clinical Hospital of the Ministry of Internal Affairs and Administration with the Warmia-Mazury Oncology Centre, Olsztyn, Poland
ORCID ID: 0009-0005-8775-102X

Donata Pokorska
The Voivodal Specialist Hospital, Olsztyn, Poland
ORCID ID: 0009-0005-6271-3465

Julia Świechowska
Uniwersyteckie Centrum Kliniczne w Gdańsku, Gdańsk, Poland
ORCID ID: 0000-0001-6492-6863

Zuzanna Zarzycka
7th Navy Hospital, Gdańsk, Poland
ORCID ID: 0009-0008-1374-5876

Ingrid Samberger
7th Navy Hospital, Gdańsk, Poland
ORCID ID: 0009-0002-6906-6852

Magdalena Wiciak
Uniwersyteckie Centrum Kliniczne w Gdańsku, Gdańsk, Poland
ORCID ID: 0009-0007-9221-4063

ABSTRACT

Introduction: Contemporary medicine is undergoing an unprecedented transformation driven by the integration of advanced artificial intelligence (AI) and large language models (LLMs) into clinical workflows. While these technologies objectively enhance diagnostic precision, their implementation creates a fundamental paradox: the increase in technological efficacy often correlates with a decline in patient trust, known as the "AI trust gap." This review examines the tension between algorithmic authority and the necessity of the "human touch," analyzing the impact of digital innovations on clinical autonomy and the patient-physician-AI triad.
Materials and Methods: This study presents a detailed analysis of 44 peer-reviewed scientific articles published between 2022 and 2026. The review focuses on Clinical Decision Support Systems (CDSS) across key diagnostic areas, including radiology and pathology. The analysis encompasses the psychological mechanisms of AI acceptance, the risks of automation bias, and the potential of Explainable AI (XAI) to restore clinical transparency.
Key Findings: Research reveals that the mere disclosure of AI involvement can reduce patient trust (dropping from 0.50 to 0.30–0.34 in experimental settings). A "paradox of knowledge" was identified, where higher patient literacy regarding AI correlates with increased skepticism. Regarding clinical autonomy, a dichotomy exists: junior clinicians are prone to automation bias, while experts face the risk of "deskilling." The review also discusses the "Algorithmic Consultant" role and the necessity of "Triadic Decision-Making," where AI serves as a transparent partner rather than a black-box authority.
Conclusions: The integration of AI requires a reconfiguration of medical practice from a technology-first approach to a human-centered design. Preserving clinical autonomy depends on adopting a "trust but verify" model and implementing XAI strategies to mitigate transparency barriers. The success of algorithmic medicine relies on maintaining the physician's judgment as the cornerstone of care, ensuring that AI functions as a supportive co-pilot.

KEYWORDS

Artificial Intelligence in Medicine, Patient Trust, Clinical Autonomy, Automation Bias, Explainable AI (XAI), Shared Decision-Making

CITATION

Maciej Kokoszka, Michalina Chodór, Julia Maria Kuczkowska, Judyta Bordakiewicz, Zuzanna Michalska, Donata Pokorska, Julia Świechowska, Zuzanna Zarzycka, Ingrid Samberger, Magdalena Wiciak. (2026) Algorithmic Authority vs. Human Touch: A Narrative Review of Patient Trust and Clinical Autonomy in AI-Assisted Diagnostics. *International Journal of Innovative Technologies in Social Science*. 1(49). doi: 10.31435/ijitss.1(49).2026.4963

Introduction

Over the past three years, contemporary medicine has undergone a massive shift driven by artificial intelligence (AI), which has come to dominate both scientific discourse and clinical practice[1]. Integrating advanced machine learning algorithms into daily workflows is no longer just a theory; it is an operational reality approved by major regulatory bodies like the FDA. We have moved rapidly from simple analytical tools to complex generative models, such as ChatGPT, which demonstrate the ability to perform serious clinical reasoning[2].

The potential here is enormous, ranging from discovering new therapies to tailoring precision medicine for individual patients[1]. In fields like radiology and pathology, AI is already making image interpretation significantly more accurate, reducing errors and speeding up diagnoses[1,2]. These tools synthesize vast amounts of data from electronic records, turning raw numbers into useful insights. Moreover, AI assistants promise to automate documentation, theoretically freeing up clinicians to focus on direct patient interaction. This scale is reflected in the market, which is expected to jump from $27 billion in 2024 to over $613 billion within a decade[1].

Consequently, Clinical Decision Support Systems (CDSS) now allow doctors to combine specific patient data with evidence-based knowledge for personalized recommendations[3]. This capability helps catch rare diseases earlier and optimize treatment in critical conditions, leading directly to better health outcomes[4].

Ultimately, this transformation redefines the professional role—making AI an essential part of the healthcare ecosystem rather than just a support tool[1].

However, despite these technological benefits, implementing AI creates a fundamental paradox. While AI objectively improves accuracy, its overt use often leads to a decline in patient trust. Experiments show that knowing a doctor is supported by AI decreases a patient's willingness to seek help and lowers their perception of the physician's professionalism. This "AI trust gap" is especially pronounced when algorithms are used extensively.

Patients often fear that AI cannot account for their unique, individual health factors[2]. Simultaneously, clinicians face "automation bias" and the risk of "deskilling"—losing their professional edge by relying too heavily on CDSS recommendations[3,4]. There is a valid concern that doctors might unthinkingly accept machine-generated results, leading to errors if the algorithm fails or "hallucinates." Trading clinical intuition for blind trust in technology poses a serious threat to physician autonomy[4]. Furthermore, patients can feel objectified in a system where key decisions are made by a "black box" devoid of empathy[3,5]. This conflict between algorithmic authority and the need for human contact is a major barrier to adoption. The lack of transparency, combined with the risk of generative models presenting false facts convincingly, further deepens public skepticism[2].

An analysis of the current literature reveals a significant gap. Most publications focus on technical metrics like sensitivity and specificity, paying far less attention to patient acceptance or the psychology of the doctor-patient relationship. We lack a holistic approach that views the patient, physician, and AI as a cooperative triad. This article aims to fill that gap by analyzing how CDSS impacts trust in terms of competence, fairness, and benevolence[3]. We examine how AI integration affects clinical autonomy and whether it is possible to maintain the "human touch" in an automated environment[5]. Key challenges include trust calibration and understanding how knowledge of technology shapes patient attitudes[2]. We also emphasize participatory medicine, where technology empowers the patient rather than just automating procedures[1], and the ethical verification required for vulnerable groups[4]. Our goal is to help build communication strategies that minimize aversion to AI, as understanding these barriers is key to improving health outcomes[2].

Materials and Methods

This review is based on a detailed analysis of 44 peer-reviewed scientific articles published between 2022 and 2026, capturing the most recent trends following the generative AI revolution. Our scope centers on Clinical Decision Support Systems (CDSS) within key diagnostic fields, placing particular emphasis on image-based specialties like radiology and pathology, while also exploring emerging impacts in general practice[1,2]. Beyond standard diagnostics, the analysis extends to AI applications in screening protocols and complex case management.

To ensure a comprehensive perspective, we examined acceptance levels across diverse demographic groups, investigating how factors such as age, gender, and education influence patient attitudes toward medical technology[2]. The review also documents the shifting dynamics of patient-physician communication, including the growing role of chatbots in mental health and chronic care[1]. We distinguish between fully automated systems and the AI-assisted "hybrid" approach that currently dominates clinical practice[2]. Finally, the paper considers the broader societal impact of digital innovation, assessing whether AI bridges existing healthcare inequalities or widens the technology gap[1,5]. This review is intended as a resource for practitioners and researchers seeking to integrate artificial intelligence into diagnostic processes safely and ethically[5].

Section 1. Patient trust in AI-assisted diagnostics: The paradox of knowledge and certainty

The trust gap in the age of algorithmic medicine: Contemporary medicine faces a growing disconnect: while technology is objectively getting better, patients' sense of security isn't keeping pace. Specifically, randomized controlled trials on American adults reveal an alarming trend—simply mentioning that AI is part of the diagnostic process significantly lowers trust in the physician. Trust scores and willingness to seek help drop drastically, falling from around 0.50 in control groups to 0.30–0.34 when algorithms are heavily involved. This decline isn't just about professional competence; it impacts how patients perceive the doctor as a person, suggesting a deep psychological barrier to automated care.

Interestingly, this negative sentiment persists regardless of a patient's political views or education level. It becomes particularly pronounced in high-risk clinical situations, where the intervention of Clinical Decision Support Systems (CDSS) is often seen as undermining the staff's integrity[2]. For instance, female patients tend to be significantly more skeptical, rating doctors who use CDSS lower on professionalism scales[3].

In the context of medical imaging, patients prioritize three pillars: efficiency, responsibility, and understanding the process[6]. Even if AI systems technically outperform human experts, patients still place higher value on human reliability[2]. A major issue is responsibility—patients are unclear about who is legally liable if an autonomous machine makes a mistake[6]. The lack of transparency in AI models exacerbates this "black box" fear, hindering acceptance even when accuracy is proven[3,6].

The knowledge and trust paradox: One might assume that democratizing access to complex medical data via generative AI would improve trust. After all, these tools translate jargon into understandable language, which should strengthen the doctor-patient bond[5]. However, empirical research reveals a profound paradox: higher declared knowledge about AI often correlates with *less* trust, not more.

While daily use of AI breeds familiarity, understanding how algorithms actually work is negatively associated with the intention to undergo treatment. Surprisingly, those with the highest AI literacy are often the most skeptical. This is likely due to a heightened awareness of limitations, such as hallucinations and ethical risks. Knowledgeable patients may fear that AI cannot account for their unique, individual health factors.

This skepticism can be compounded by overconfidence, where patients overestimate their own knowledge relative to the system[2]. Consequently, patients armed with AI-derived information may start questioning their doctors, leading to conflict rather than cooperation. If the human diagnosis differs from the machine's, the patient becomes confused and loses confidence in the staff[5].

Demographics and psychological mechanisms of AI acceptance: Attitudes toward AI are strongly shaped by demographics and personality. Women and patients in life-threatening situations show the strongest aversion to CDSS, likely due to a higher sensitivity to potential errors and a fear of losing control[2,3]. Older age groups, however, demonstrate a unique habituation mechanism: after initial skepticism, regular contact can lead to "blind trust," creating a risk of uncritically accepting algorithmic suggestions[2,4]. Personality also plays a role; proactive individuals demand to verify data, while skeptics base their trust on negative past experiences[4].

We are witnessing a shift from the "e-patient" to the "AI-patient." These individuals don't just want information; they want a voice in the treatment process[1,2]. They view AI as a "co-pilot" that enables an equal partnership in the doctor's office[1]. The cornerstone of this shift is informed consent. The vast majority of respondents believe information about AI use should be mandatory[2,7]. Over 90% of patients feel disclosure should be standard practice, and they want the right to opt out of machine assistance[8]. This preference is particularly strong among those with higher socioeconomic status, who expect full transparency regarding the system's methodology[7,8].

Clinical implications: The widespread integration of AI is forcing a redefinition of the medical hierarchy, shifting emphasis from physician authority to patient autonomy. The challenge is to calibrate trust so that AI is seen as a tool that supports—not replaces—human empathy[2,5]. Research suggests satisfaction is highest when the physician openly communicates about AI use or precisely explains its benefits[2].

Transparency builds credibility, especially when disclosing discrepancies between a radiologist's opinion and an algorithm's output[8]. In this new model, the triad consists of the patient, physician, and AI, with the doctor evolving into a guide who interprets digital data[5]. Crucially, we must avoid "deskilling"—patients view the loss of a doctor's clinical intuition as a significant threat[3,4]. Paradoxically, offering a choice between purely human and hybrid diagnostics may actually increase overall acceptance by respecting self-determination. Communication strategies must also be tailored; younger and wealthier individuals often require more in-depth arguments to be convinced[7]. Ultimately, success depends not on technical perfection, but on integrating AI in an ethical, transparent, and patient-centered manner[2,9].

Section 2: Physician Autonomy Under Siege: From Clinical Authority to Algorithmic Consultation

Bringing artificial intelligence (AI) into the clinical domain marks a fundamental transformation of the medical profession. We are moving from a model of individual clinical freedom toward a structured, algorithmically mediated environment. Physician autonomy—defined as the freedom to control diagnosis, treatment, and professional tasks—is currently being redefined by the rapid spread of clinical decision support systems (CDSS)[10]. While AI is often framed simply as a tool for precision, its implementation frequently feels like a "siege" on traditional authority. The physician shifts from being the sole expert to a participant in a complex triadic relationship involving the machine and the patient[10,11]. Research suggests this isn't merely a technical upgrade; it is a profound socio-technical realignment that shakes the foundations of professional identity[12]. Consequently, the profession faces a landscape where the promise of efficiency comes paired with significant threats to the relational core of medicine[11,12].

The Autonomy Paradox: The impact of AI creates a distinct paradox. For junior clinicians and students, these tools offer a form of "augmented autonomy." They provide decision certainty, suggest care possibilities, and help navigate complex guidelines[10]. However, this reliance introduces the risk of "automation bias." Practitioners may favor algorithmic suggestions even when they are inaccurate or based on flawed data. In experimental settings, medical students have shown a marked tendency to be "swayed" by incorrect tool outputs, even when their initial, unassisted estimations were closer to the truth. This suggests that for those in the learning phase, AI doesn't just support decision-making; it fundamentally shapes the cognitive process, potentially creating a dependency that hinders independent reasoning[13].

In contrast, senior pathologists and specialists often take a more cautious stance. They view AI as a potential threat to their sovereignty and professional identity[12,13]. The "siege" here manifests as a narrowing of inter-observer variability. While AI improves agreement across the board, it also forces a convergence toward an "algorithmic mean"[13]. This convergence risks erasing the "nuanced expertise" that distinguishes human clinicians in atypical cases[11,12]. Furthermore, experts worry that if they are required to justify every deviation from an AI protocol, their autonomy is essentially subordinated to statistical inference[12,14]. Thus, empowering juniors might come at the steep cost of expert deskilling[15].

The Emergence of the Algorithmic Consultant: As medical systems become saturated with complex models, the traditional physician role is evolving into something new: the "Physician-Algorithm Specialist" or "Algorithmic Consultant"[10,16]. This shift is necessary because direct interaction between a doctor and an AI can be underwhelming if the human lacks training in data science. Analogous to clinical pharmacy, this specialist would act as a steward of the hospital's ecosystem, guiding tool selection and interpretation. The Algorithmic Consultant fills the "interpretation gap," adjusting for biases, model limitations, and data drift that a general practitioner might miss. This reconfigures authority from "making the primary diagnosis" to "validating the system's safety"[16].

This transformation reshapes the triad, especially in primary care where continuity is paramount[11]. AI tools are increasingly integrated as "silent partners" handling admin tasks, theoretically freeing up time for human interaction[11,14]. However, physicians remain wary that this leads to depersonalization—where technology becomes an intruder rather than a co-pilot[11]. To maintain autonomy, physicians are demanding involvement in the design of these systems to ensure they align with clinical reality, not just technical benchmarks[10,12]. The future clinician must possess high AI literacy to act as the final authority, integrating data-driven insights with the context that AI lacks[10–12].

Negotiation with the System - "Trust but Verify": The professional response to AI isn't a binary choice between acceptance and rejection; it is a sophisticated negotiation[12]. For instance, family physicians in Lithuania generally accept AI for routine tasks but remain cautious with complex scenarios. This "negotiated trust" requires clinicians to use the algorithm as a "second look," double-checking recommendations against their own expertise[11]. However, the "black box" nature of deep learning makes this verification difficult[11,12]. The lack of transparency creates a psychological burden known as "technostress"—the feeling that you are working *for* the system, rather than the system working for you[12,14].

Moreover, integrating AI introduces a new layer of medico-legal tension[10]. Here is the trap: if a physician ignores an AI recommendation and an error occurs, they face intense scrutiny. But if they follow a flawed AI suggestion, the liability remains ambiguous[10,15]. This creates an environment of "defensive practice," where autonomy is sacrificed to adhere to digital protocols just to be safe[12,14]. To mitigate this, practitioners advocate for incremental implementation—starting with low-risk tasks—to build trust and clear

boundaries[14]. The goal is to ensure AI augments clinical judgment rather than replacing the physician's ultimate responsibility[11,12].

The Risk of Deskilling and Self-Referential Loops: A critical long-term concern is deskilling. As LLMs like ChatGPT become ubiquitous for drafting notes and synthesizing literature, there is a risk that future professionals will lose the diagnostic acumen developed through deep engagement with primary material. This trend signals an inverse correlation: as trust in AI escalates, the preservation of clinical skills may decline. If clinicians habitually delegate analysis to AI, their ability to perform manual reviews during system failures could be severely impaired. This is particularly acute in environments where productivity pressures discourage "deliberate practice."

Compounding this is the "Paradox of the Self-Referential Learning Loop." LLMs are increasingly trained on data they generated themselves. As AI content floods medical records, models begin to learn from their own prior outputs, creating an "echo chamber"[15]. This loop threatens data diversity and may entrench clinical biases present in initial training sets[10,15]. For physician autonomy, this means clinical reasoning could be confined within a narrowing cycle of algorithmic reproduction. To prevent this, we need a symbiotic relationship where AI augments human intellect, maintaining the necessity for critical thinking even when supported by accurate models[15].

In conclusion, the current trajectory suggests physician autonomy is undergoing a significant reconfiguration rather than simple erosion. The profession must navigate the paradoxes of empowerment and atrophy, transition toward specialized roles, and negotiate new boundaries of trust[10,12]. Success in this era of "algorithmic medicine" depends on maintaining a human-centered design where the physician's emotional intelligence remains the cornerstone of care[12,14]. Only through proactive involvement can the medical profession survive this "siege" and ensure AI functions as a reliable assistant[11,12].

Section 3: Explainable AI - Making Decisions Transparent

The Black Box Problem - Why Precision Needs Explanation: The shift from hypothesis-driven discovery to a data-driven research paradigm positions artificial intelligence as a cornerstone of modern diagnostic logic. However, advocates for these systems bear an urgent responsibility to demonstrate that this guidance can be meaningfully trusted[17]. Systematic meta-analyses confirm a difficult reality: while AI integration significantly enhances diagnostic precision, the inherent lack of transparency and interpretability remains a critical barrier to widespread clinical adoption[18]. High-accuracy AI systems have demonstrated a clear ability to improve clinician performance, but it is equally established that incorrect or poorly justified AI advice can substantially decrease diagnostic accuracy[19]. Consequently, stakeholders widely expect explainable AI (XAI) methods to provide relevant information regarding a model's inner workings to facilitate safe medical decision-making[20].

Beyond Didactic Logic: In reality, physicians do not seek didactic, patronizing justifications of internal computational logic; they require confidence in the reliability of recommendations to support a professionally and legally justifiable rationale for their decisions[21]. The potential for XAI to improve interpretability is clear, but this is only realized when practitioners can comprehend the reasoning behind diagnostics to build necessary confidence[22]. One of the most significant challenges identified in the literature is that AI models often produce outputs without stating the logic behind them, making it difficult for healthcare professionals to identify errors or systematic biases[23]. The undisclosed nature of these complex systems creates fundamental obstacles, as practitioners inherently need to understand the underlying reasoning behind the outcome of a particular diagnostic process[24]. Ultimately, a doctor's acceptance and willingness to use AI are deeply influenced by the extent to which the system provides human-understandable representations of its process[25]. Earned trust must be built upon disciplined testing similar to evidence-based medicine[17]. Therefore, earnable accountability must be proactively pursued by developers to ensure systems combine high predictive performance with strong explanation fidelity, which has been shown to increase clinician trust scores by 12 to 18 percentage points[18].

Current Approaches - From Global Prototypes to Local Features: A systematic approach to transparency requires detailed disclosure of the "who, what, when, where, and how" of application development[17]. The landscape of interpretability is currently dominated by techniques like SHAP, Grad-CAM, and LIME[18]. Empirical evaluations have demonstrated that local feature-based explanations yield significantly better physician diagnostic accuracy when AI advice is correct compared to global prototype-based explanations[19]. However, caution is required because many post-hoc XAI methods prioritize human-readable simplicity over a faithful description of predictions[20]. To move beyond static justifications, XAI

should be reframed as "eXtended and eXplorable AI," emphasizing interactive uncertainty transparency[21]. In practical applications, LIME excels at providing local interpretability by emphasizing specific medical terms like "opacity"[22]. Visual tools like heatmaps allow clinicians to verify whether the AI focuses on anatomically relevant areas[23]. Health care professionals predominantly prioritize this local explainability as an enabler of trust[25]. Standardization of these evaluation practices is essential to bridge the divide between research prototypes and actionable clinical tools[18].

The Causality Gap and the Risk of Overreliance: Despite the promise of transparency, corroborating a causal link between data and clinical action remains a core tension[17]. Many current XAI methods function primarily as descriptive tools rather than explanations of biological cause-and-effect[20]. A critical pitfall emerges with local explanations, which can foster "reliance without verification," increasing overreliance even when the advice is empirically incorrect[19]. This risk is compounded by the fact that prediction models learn statistical associations, which may not match underlying causal mechanisms[20]. Furthermore, a constant stream of rationalizations for routine tasks risks fostering the very automation bias developers seek to mitigate[21]. Limitations in feature attribution can bias perceived depth[22]. Recent reviews highlight key gaps, including limited dataset diversity, which reduces generalizability[23]. Many high-performance models prioritize accuracy over interoperability, leaving practitioners in the dark regarding decision logic[24]. Ethical concerns also arise when systems are poorly integrated into workflows[25]. Consequently, using model explanations as the sole basis for decisions can be dangerous if correlations result from collider bias[20].

Toward Reliability Metadata and Uncertainty Quantification: To resolve these challenges, it is essential that XAI systems generate reliability metadata—marginal and instance-specific indicators—rather than didactic explanations[21]. Such metadata should disclose limitations so stakeholders can judge face validity[17]. Reliability-aware systems enrich outputs with actionable cues, such as calibrated confidence scores, that signal when a patient state differs from the training set[21]. Statistical validation using robust tests is crucial to ensuring AI solutions are reliable[22]. Since clinicians frequently report challenges in understanding complex models, there is an urgent need for approachable tools[23]. Design strategies must prioritize uncertainty quantification, enabling clinicians to apply heightened scrutiny. These interactive processes ensure AI functions as a "thought partner" rather than a detached instructor[21]. Utilizing techniques to handle data imbalance enhances fairness[24]. A framework centering on HCP needs is vital for adoption[25]. Multi-level interfaces can accommodate varying expertise, allowing for a "calibrated trust" mechanism[18,25].

AI as a Silent Partner: In practical applications, AI presents a unique ability to dynamically assemble patient cohorts for complex conditions[17]. CDSS applications in oncology and cardiology have already begun utilizing XAI to interpret malignancy and arrhythmia[18]. Optimal human-machine teamwork is essential for improved outcomes[19]. Sophisticated hybrid frameworks can achieve 99.2% accuracy while providing specific attributes responsible for the diagnosis[24]. For instance, advanced language models like LLaMA-3.1 achieve high accuracy in radiology reports when paired with LIME and SHAP[22]. SHAP values calculate the additive contribution of each feature to the final risk estimate[20,23]. Deep Ensembles allow for the visualization of temporal dependencies, aligning AI suggestions with dynamic clinical reasoning[18,21]. Furthermore, integrating multi-modal inputs provides the comprehensive picture required by specialists[23]. Testing how best to integrate this flow is vital for preventing burnout. Ultimately, earning sustained trust requires the same diligence used in evidence-based medicine[17]. By following user-centered design, AI-CDSS can transition to indispensable tools[25]. As these frameworks mature, they will continue to surface successful strategies, acting as a "silent partner" in the patient's journey[17].

Section 4: Automation Bias in Medical AI: When Physicians Stop Thinking and Trust the Machine

The Blind Spot of Certainty: Integrating artificial intelligence (AI) into clinical workflows is currently stalled by three fundamental barriers: a lack of interpretability, difficulties in generalizability, and the significant risk of automation bias[26]. This phenomenon is defined as the human tendency to over-rely on automated outputs, often at the expense of actively seeking information or evaluating uncertainty[26,27]. It differs fundamentally from confirmation bias; while the latter involves favoring results that align with an initial hypothesis, automation bias is a systematic and often blind assumption of the machine's superiority[15,28]. In high-stakes environments like the Intensive Care Unit (ICU), failing to quantify this uncertainty can lead to unexpected model failures and catastrophic errors[26]. Empirical data from medical imaging suggests this is a pervasive issue, with errors in AI-assisted rounds directly linked to misleading algorithmic suggestions in 45.5% of cases[29].

The Authority Effect and the Novice Trap: The "authority effect" of AI has been rigorously demonstrated through experimental studies using fictitious support systems. Research involving the diagnosis of otitis media showed that while a fictitious AI improved overall accuracy from 64% to 75%, it simultaneously misled less experienced practitioners into accepting incorrect diagnoses. The highest levels of automation bias and "detrimental algorithmic aversion" were observed among medical students and residents, whereas experienced physicians demonstrated greater resilience[30]. This susceptibility is often intrinsic to the decision-maker; essentially, the more a user perceives the system as beneficial, the more likely they are to agree with false recommendations. These findings confirm a paradox: non-specialists, who theoretically stand to gain the most from AI assistance, are the most vulnerable to its inaccuracies[27]. The psychological impact is profound, as clinicians frequently move toward inaccurate tool recommendations even when their initial, unaided assessments were closer to the ground truth[13].

Strategies for Mitigation: To mitigate these risks, several technical and educational strategies must be implemented. One promising approach is the "AI suppression strategy," which involves retracting algorithmic recommendations in high-risk zones where the probability of misleading the clinician outweighs the probability of correction. It is estimated that such a strategy could decrease automation bias events by approximately 41.7% while maintaining the benefits of AI for high-confidence predictions[29]. Furthermore, transitioning from binary point predictions to conformal prediction regions allows systems to flag uncertain cases for human intervention, effectively saying "I don’t know"[26]. Educationally, using Kern’s six-step approach to curriculum development is essential to foster AI literacy, ensuring future physicians understand the ethical, regulatory, and technical limitations of the tools they employ[31]. Limiting the use of generative tools like ChatGPT to specific administrative tasks can also prevent students from developing an "AI solutionism" mentality—the belief that AI has an answer for every clinical problem[28].

Clinical Consequences and Long-Term Risks: The clinical consequences of automation bias are already evident across multiple specialties. In neuroradiology, false-positive AI findings in detecting cerebral aneurysms lead to significantly higher suspicion levels, causing inexperienced readers to recommend unnecessary, aggressive follow-up procedures[32]. In digital pathology, students and residents estimating tumor cell fractions (TCF) have been observed following inaccurate CAD suggestions despite visible cues, such as the misclassification of intraglandular debris as non-tumor cells[13]. This behavior is exacerbated in fast-paced environments where "omission errors" (failing to notice AI mistakes) and "commission errors" (following AI despite conflicting evidence) become more frequent[33]. Even when clinicians are explicitly warned of low AI confidence, they often remain vulnerable once trust has been established, particularly when the AI provides a faulty conclusion with high apparent certainty[29].

The most concerning long-term implication is the progressive deskilling of the medical workforce. Habitual delegation of diagnostic tasks to AI leads to an unintended erosion of clinical skills, a risk particularly acute for "digital natives" who enter medicine with insufficient technical skepticism[15,28]. This is compounded by the "self-referential learning loop," where AI-generated content increasingly feeds back into the training data for future models. This loop threatens the diversity of clinical data, entrenching existing biases and degrading both algorithmic efficacy and human expertise. Furthermore, the lack of inherent understanding in Large Language Models (LLMs) can lead to "hallucinations"—plausible but entirely fabricated information—that clinicians may accept as fact if they lack specialized knowledge[15].

Maintaining the Human in the Loop: To ensure a safe future for medical AI, practitioners must remain the final arbiter of biological plausibility through "Human-in-the-loop" (HITL) strategies[33]. Transparency in how AI models are developed and validated is critical for building a "calibrated trust" that allows clinicians

to verify errors effectively. Ultimately, the responsibility for mitigating bias and preventing deskilling rests with both educational institutions and the clinicians themselves. They must maintain a balance of trust and skepticism to ensure AI augments rather than undermines patient care[15]. Addressing these systemic issues requires a paradigm shift: the medical profession must evolve to prioritize unique human skills—such as critical reasoning, empathy, and complex problem-solving—over the traditional reliance on memory-based diagnostic tasks[15,31]. Only through unstructured, ethically grounded, and evidence-based AI training programs can the healthcare system avoid the pitfalls of a future where physicians stop thinking and start blindly following the machine[28,31].

Section 5: Triadic Decision-Making: Integrating AI into Shared Decisions

The Structural Shift: Moving toward inclusive artificial intelligence (AI) requires admitting that measurement processes are rarely neutral; instead, they often introduce a "noise-bias cascade" where human interpretation and selection bias are inextricably linked[34]. This evolution necessitates shifting from the traditional clinician-patient dyad to an artificial intelligence–supported shared decision-making (AI-SDM) model, which serves as a structural framework for triadic deliberation[35]. However, the current landscape suggests that evidence regarding AI usage at the actual point of decision-making remains in its infancy, often lacking a comprehensive map of how these technologies facilitate clinical contact[36]. To address these gaps, modern protocols are investigating the impact of AI-based risk prediction—such as for individual graft loss—to determine if such data can increase the frequency of preference-sensitive conversations[37]. While healthcare professionals acknowledge that AI-driven support holds promise for enhancing triage, its integration into mental health raises significant concerns regarding the preservation of the therapeutic relationship and clinical autonomy[38]. This anxiety is echoed by patients undergoing cross-sectional imaging, who express a strong desire for transparency: 90.3% believe they should be explicitly informed when AI is utilized in their care[8].

Phases of Deliberation and Development Gaps: To maintain objectivity, system development must span the entire AI lifecycle, incorporating uncertainty measures during deployment to build trust before clinical validation occurs[34]. The triadic deliberation process operates through four integrated phases, beginning with structured data acquisition that consolidates medical history, patient values, and AI-derived evidence[35]. Scoping reviews reveal a critical flaw: while most studies utilize machine learning, almost none included patients or providers in the initial design stages[36]. Consequently, prospective trials like PRIMA-AI seek to bridge this gap by randomizing patients to AI-supported care, where physicians are assisted by validated systems to guide treatment options[37]. In practice, only 15 out of 23 established SDM elements were identified as relevant by professionals, suggesting that current AI functionalities do not yet align with theoretical components of shared decision-making[38]. This misalignment is critical; patients indicate they would largely reject screening exams if interpreted exclusively by an algorithm without human oversight[8].

Mitigating Bias and Generating Reasoning: Deep-seated biases, such as those in devices calibrated only for lighter skin tones, can be mitigated through strategies like oversampling underrepresented groups and federated learning[34]. The second phase of triadic deliberation involves reasoning generation, where models transform outputs into adaptive narratives addressing the "why" behind suggestions[35]. Research highlights that while AI discovers correlations, the final decision must be informed by what matters most to the patient, preserving the personal aspect of the encounter[36]. In kidney transplantation, for instance, systems provide color-coded predictions to ensure risk is easily interpretable for the physician[37]. Providers emphasize that AI should act as a tool for "holistic understanding" rather than a rigid question-and-answer form[38]. To align with these preferences, radiology practices must prioritize disclosure, ensuring patients can opt out of AI-assisted practices[8].

Interactive Deliberation and the Human Connection: The "noise-bias cascade" can be further addressed by involving diverse stakeholders to identify disparate impacts[34]. During the third phase, interactive deliberation occurs, allowing patients to interrogate recovery timelines in real-time[35]. Unfortunately, evidence shows a lack of emphasis on patient values in most AI interventions, which focus primarily on clinical recommendations[36]. For this reason, semistructured interviews are being used to assess how AI influences the normative foundations of trust[37]. Clinicians frequently raise concerns about AI's inability to interpret non-verbal cues, which could weaken the interpersonal connection[38]. From a patient standpoint, 89.1% believe that clarifying any interpretation discrepancies should be a standard of care[8].

Implementation, Safety, and Fairness: Standard performance measures like accuracy must be interpreted based on domain-specific factors, such as the severity of false negatives[34]. The final phase involves implementation and documentation, capturing AI reasoning in the medical record for accountability[35]. Despite the promise of structured workflows, routine use is often limited by time pressure[36]. Systems optimizing these workflows must undergo rigorous safety assessments, including monitoring for adverse events[37]. Professionals suggest AI could enhance patient education by delivering tailored psychoeducational content[38]. However, socioeconomic status influences these preferences; individuals in high SES groups show more willingness to pay for AI-enhanced services[8].

The Redefined Physician Authority: To reduce measurement bias, utilizing multicenter data and annotated inputs from multiple operators is recommended[34]. AI reasoning differentiates itself from traditional XAI by focusing on semantic clarity and interpersonal trust[35]. While AI's potential to empower patients and reduce waiting times is a major benefit, it must be balanced against "black-box" risks[36]. Success is often measured by tools like the CollaboRATE score to evaluate perceived involvement[37]. There is a noted gap in addressing "mutual agreement," which is essential for clinical decisions[38]. Patients prefer radiologists who explain why their human interpretation might differ from an automated output[8].

Finally, continuous monitoring is essential to rectify bias as conditions evolve[34]. Without communicative layers facilitating narrative synthesis, AI suggestions may appear arbitrary, eroding trust[35]. Evidence suggests technology can improve satisfaction if the sociotechnical environment is considered[37]. Integrating "human-centered AI" is advocated to preserve patient agency[38]. Ultimately, transparency must be prioritized to ensure AI does not exacerbate disparities[8]. The physician's authority is thus redefined: not as an absolute decision-maker, but as a facilitator of "justifiable clinical reasoning"[35]. By automating routine screenings, AI frees up energy for deeper connection[34,36]. This requires rethinking fairness beyond technical fixes[34]. Future research must prioritize AI that supports the full spectrum of SDM elements[38]. The integrity of the triadic model rests on the clinician's ability to correct AI outputs, protecting the autonomy of patient care[34,35].

Section 6: Discussion & Synthesis: Toward a Framework

The Trust and Autonomy Paradoxes: Integrating artificial intelligence (AI) into the clinical environment represents a paradigm shift that demands we fundamentally rethink the relationship between patient, clinician, and technology. Empirical evidence suggests a double-edged sword: while AI facilitates processes by cutting errors and costs, it faces profound psychological barriers. Experiments demonstrate that the mere mention of AI assistance consistently decreases patient trust, revealing a persistent preference for human expertise[2]. This creates a "Trust Paradox"—patients stand to benefit the most from these innovations, yet they are the most skeptical because their perspective is often under-researched[6].

Despite this hesitation, the transition to data-driven medicine is inevitable. AI is essential for interpreting complex "omics" data that bypass traditional biological hypotheses[39]. However, the "Autonomy Paradox" persists. Physicians identify the loss of clinical and professional freedom as a critical barrier to accepting these systems[10]. To bridge this gap, advocates must demonstrate that AI guidance can be trusted by adhering to precedents like evidence-based medicine (EBM)[17].

Barriers to Adoption - Dependency and Bias: A robust framework for this transition must prioritize explainable AI (XAI) to foster collaboration[18]. We need reliable uncertainty quantification, such as conformal prediction, to flag cases that require human intervention[26]. Without such safeguards, there is a legitimate risk of "ChatGPT Dependency Disorder," where over-reliance on generative tools degrades independent clinical reasoning[40]. Furthermore, achieving fairness requires confronting the "Trust-Bias Paradox." Measurement processes often introduce a "noise-bias cascade" due to unrepresentative sampling[34]. This necessitates policy responses that ensure equitable access, particularly in resource-constrained environments where infrastructure is lacking[41]. Practical adoption in surgery, for instance, remains limited by fears of over-reliance despite the potential for real-time analytics[43]. Finally, sustainability depends on adhering to standardized reporting guidelines like AGREE II and FAIR principles[44].

Proposed Three-Layer Framework: To address these challenges, we propose a comprehensive framework for trustworthy adoption:

1. **The Patient Layer - Consent and Empowerment**: Acceptance is contingent upon democratizing knowledge[5]. Transparency must move beyond simple disclosure to a patient-centric approach where AI usage is clearly communicated[6]. Recommendations should be presented as options, not directives, preserving the patient's right to weigh advice against personal values. This layer ensures interaction is culturally nuanced to bridge healthcare disparities[5].
2. **The Clinician Layer - Critical Thinking as the Fail-Safe**: Clinicians must be repositioned as the final authority, using AI as a support, not a replacement[10,17]. To mitigate automation bias, systems should incorporate "cognitive forcing"—withholding information until the doctor has performed independent reasoning[40]. Integrating XAI methods like SHAP allows for transparent reasoning pathways[18]. Advanced models like CPMORS demonstrate how uncertainty estimation can mandate clinician review for ambiguous cases[26]. Even in high-precision fields like neurosurgery, hands-on training is required to build interactive competence[43].
3. **The System Layer - Governance and Fairness**: Technology is never neutral[40]. Robust governance must address the "noise-bias cascade" through rigorous validation[34]. Developing nations need sector-specific ethical guidelines to prevent discrimination[41]. Developers must ensure transparency in model construction to earn accountability[17,42]. High-quality reporting guidelines like TRIPOD+AI must be enforced to ensure scientific rigor[44].

Conclusions

Transparency alone is insufficient; the translation of AI requires a symbiotic interaction where the system identifies high-reliability cases while clinicians focus on complex anomalies[26]. Automation bias is an inherent challenge requiring continuous monitoring[6]. Fairness is not a static state but requires active supervision involving diverse stakeholders[34,41].

For practitioners, this implies a shift from memory-based heuristics to AI literacy[39]. Physicians should prioritize open communication to preserve the therapeutic relationship[5]. Documentation must explicitly reflect where AI was used and how it was verified, especially in hybrid scenarios where liability is complex[42].

For developers, the directive is to engineer "explainable and fair" systems[18]. Investment in digital infrastructure is essential to prevent exacerbating health disparities[41]. Furthermore, medical curricula must include AI-specific ethical modules to build a proficient workforce[18,43].

Looking toward the future, AI in medicine is unpreventable, but its success depends on transitioning from "hyped" performance to "designed" trust[44]. Future research must prioritize longitudinal studies on patient outcomes and systemic bias[5]. Cross-national patterns of acceptance should be examined to ensure implementation strategies accommodate diverse cultural contexts[2,42]. Ultimately, trust in AI cannot be assumed based on sophistication; it must be cultivated through transparency, accountability, and a persistent focus on the human connection[5,18].

Author's contribution: Maciej Kokoszka
Conceptualization: Maciej Kokoszka, Michalina Chodór
Methodology: Julia Maria Kuczkowska, Maciej Kokoszka, Judyta Bordakiewicz
Software: Zuzanna Michalska, Donata Pokorska, Julia Świechowska
Check: Zuzanna Zarzycka, Ingrid Samberger, Magdalena Wiciak
Formal analysis: Judyta Bordakiewicz, Julia Świechowska
Investigation: Michalina Chodór, Maciej Kokoszka
Resources: Zuzanna Zarzycka, Ingrid Samberger
Data curation: Zuzanna Michalska, Julia Maria Kuczkowska, Ingrid Samberger
Writing-Rough Preparation: Maciej Kokoszka, Donata Pokorska, Zuzanna Zarzycka,
Writing-Reviews and Editing: Michalina Chodór, Maciej Kokoszka, Zuzanna Michalska
Visualization: Judyta Bordakiewicz, Julia Świechowska, Magdalena Wiciak
Supervision: Maciej Kokoszka, Zuzanna Michalska
Project administration: Michalina Chodór, Maciej Kokoszka
Funding statement: The study did not receive special funding.
Institutional review board statement: Not applicable.
Informed consent statement: Not applicable.
Data availability statement: Not applicable.
Conflict of interest: The authors declare no conflict of interest.

REFERENCES

1. Woods, S. S., Greene, S. M., Adams, L., Cordovano, G., & Hudson, M. F. (2025). From e-patients to AI patients: The tidal wave empowering patients, redefining clinical relationships, and transforming care. *Journal of Participatory Medicine, 17*, e75794. https://doi.org/10.2196/75794
2. Chen, C., & Cui, Z. (2025). Impact of AI-assisted diagnosis on American patients' trust in and intention to seek help from health care professionals: Randomized, web-based survey experiment. *Journal of Medical Internet Research, 27*. https://doi.org/10.2196/66083
3. Zondag, A. G. M., Rozestraten, R., Grimmelikhuijsen, S. G., Jongsma, K. R., van Solinge, W. W., Bots, M. L., Vernooij, R. W. M., & Haitjema, S. (2024). The effect of artificial intelligence on patient-physician trust: Cross-sectional vignette study. *Journal of Medical Internet Research, 26*. https://doi.org/10.2196/50853
4. Parchmann, N., Orzechowski, M., Brefka, S., & Steger, F. (2025). Evaluation of an AI-based clinical decision support system for perioperative care of older patients: Ethical analysis of focus groups with older adults. *JMIR Aging, 8*. https://doi.org/10.2196/71568
5. Traylor, D. O., Kern, K. V., Anderson, E. E., & Henderson, R. (2025). Beyond the screen: The impact of generative artificial intelligence (AI) on patient learning and the patient-physician relationship. *Cureus*. https://doi.org/10.7759/cureus.76825
6. Postle, R. D., & Forster, B. B. (2025). Patient perspectives of artificial intelligence in medical imaging. *SAGE Preprints*. https://doi.org/10.1177/08465371241298597
7. Park, H. J. (2024). Patient perspectives on informed consent for medical AI: A web-based experiment. *Digital Health, 10*. https://doi.org/10.1177/20552076241247938
8. McGhee, K. N., Barrett, D. J., Safarini, O., Elkassem, A. A., Eddins, J. T., Smith, A. D., & Rothenberg, S. A. (2025). Patient preferences for artificial intelligence in medical imaging: A single-center cross-sectional survey. *Journal of Imaging Informatics in Medicine*. https://doi.org/10.1007/s10278-025-01629-w
9. Foresman, G., Biro, J., Tran, A., MacRae, K., Kazi, S., Schubel, L., Visconti, A., Gallagher, W., Smith, K. M., Giardina, T., et al. (2025). Patient perspectives on artificial intelligence in health care: Focus group study for diagnostic communication and tool implementation. *Journal of Participatory Medicine, 17*. https://doi.org/10.2196/69564
10. Grosser, J., Düvel, J., Hasemann, L., Schneider, E., & Greiner, W. (2025). Studying the potential effects of artificial intelligence on physician autonomy: Scoping review. *JMIR Preprints*. https://doi.org/10.2196/59295
11. Ratkevičiūtė, K., & Aliukonis, V. (2025). Exploring opportunities and challenges of AI in primary healthcare: A qualitative study with family doctors in Lithuania. *Healthcare, 13*. https://doi.org/10.3390/healthcare13121429
12. Mache, S., Bernburg, M., Würtenberger, A., & Groneberg, D. A. (2025). Artificial intelligence in primary care: Support or additional burden on physicians' healthcare work?—A qualitative study. *Clinical Practice, 15*. https://doi.org/10.3390/clinpract15080138
13. Frei, A. L., Khan, A., Oberson, R., Reinhard, S., Banz, Y., Meeuwsen, F., Janowczyk, A., Grobholz, R., Dawson, H. E., Lugli, A., et al. (2025). Computer-aided tumor cell fraction (TCF) estimation by medical students, residents, and pathologists improves inter-observer agreement while highlighting the risk of automation bias. *Virchows Archiv*. https://doi.org/10.1007/s00428-025-04163-w
14. Agur Cohen, D., Heymann, A. D., & Levkovich, I. (2025). Partners in practice: Primary care physicians define the role of artificial intelligence. *Healthcare, 13*. https://doi.org/10.3390/healthcare13161972
15. Choudhury, A., & Chaudhry, Z. (2024). Large language models and user trust: Consequence of self-referential learning loop and the deskilling of health care professionals. *JMIR Preprints*. https://doi.org/10.2196/56764
16. Marwaha, J. S., Yuan, W., Poddar, M., Elsamadisi, P., & Brat, G. A. (2025). The algorithmic consultant: A new era of clinical AI calls for a new workforce of physician-algorithm specialists. *NPJ Digital Medicine*. https://doi.org/10.1038/s41746-025-01960-0
17. Wallace, P. J. (2024). Gaining trust: Lessons and opportunities for artificial intelligence in health care. *The Permanente Journal, 28*(3), 168–171. https://doi.org/10.7812/TPP/24.064
18. Abbas, Q., Jeong, W., & Lee, S. W. (2025). Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. *Healthcare, 13*. https://doi.org/10.3390/healthcare13172154
19. Prinster, D., Mahmood, A., Saria, S., Jeudy, J., Lin, C. T., Yi, P. H., & Huang, C. M. (2024). Care to explain? AI explanation types differentially impact chest radiograph diagnostic performance and physician trust in AI. *Radiology, 313*. https://doi.org/10.1148/radiol.233261
20. Carriero, A., de Hond, A., Cappers, B., Paulovich, F., Abeln, S., Moons, K. G., & van Smeden, M. (2025). Explainable AI in healthcare: To explain, to predict, or to describe? *Diagnostic and Prognostic Research, 9*, 29. https://doi.org/10.1186/s41512-025-00213-8
21. Cabitza, F., & Parimbelli, E. (2026). Let XAI generate reliability metadata, not medical explanations. *Computer Methods and Programs in Biomedicine, 273*, 109090. https://doi.org/10.1016/j.cmpb.2025.109090
22. Zamir, M. T., Khan, S. U., Gelbukh, A., Felipe Riverón, E. M., & Gelbukh, I. (2025). Explainable AI-driven analysis of radiology reports using text and image data: Experimental study. *JMIR Formative Research, 9*. https://doi.org/10.2196/77482

23. Alkhanbouli, R., Matar Abdulla Almadhaani, H., Alhosani, F., & Simsekler, M. C. E. (2025). The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions. *BMC Medical Informatics and Decision Making*. https://doi.org/10.1186/s12911-025-02944-6
24. Agrawal, R., Gupta, T., Gupta, S., Chauhan, S., Patel, P., & Hamdare, S. (2025). Fostering trust and interpretability: Integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency. *Diagnostic Pathology, 20*. https://doi.org/10.1186/s13000-025-01686-3
25. Liu, Y., Liu, C., Zheng, J., Xu, C., & Wang, D. (2025). Improving explainability and integrability of medical AI to promote health care professional acceptance and use: Mixed systematic review. *JMIR Preprints*. https://doi.org/10.2196/73374
26. Yang, M., Chen, H., Hu, W., Mischi, M., Shan, C., Li, J., Long, X., & Liu, C. (2024). Development and validation of an interpretable conformal predictor to predict sepsis mortality risk: Retrospective cohort study. *Journal of Medical Internet Research, 26*. https://doi.org/10.2196/50369
27. Kücking, F., Hübner, U., Przysucha, M., Hannemann, N., Kutza, J. O., Moelleken, M., Erfurt-Berge, C., Dissemond, J., Babitsch, B., & Busch, D. (2024). Automation bias in AI-decision support: Results from an empirical study. In *Studies in Health Technology and Informatics* (pp. 298–304). IOS Press. https://doi.org/10.3233/SHTI240871
28. Nguyen, T. (2024). ChatGPT in medical education: A precursor for automation bias? *JMIR Preprints*. https://doi.org/10.2196/50174
29. Wang, D. Y., Ding, J., Sun, A. L., Liu, S. G., Jiang, D., Li, N., & Yu, J. K. (2023). Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. *Journal of the American Medical Informatics Association, 30*, 1684–1692. https://doi.org/10.1093/jamia/ocad118
30. Hedman, M., Kosuta, V., Lindmark, M., Sandström, J., Trinh, B., Sundvall, P. D., Rystedt, K., Werner, M., Öhberg, F., & Lundberg, T. (2025). Diagnostic accuracy of otitis media with and without a fictitious AI support among physicians in primary care and medical students. *Scandinavian Journal of Primary Health Care*. https://doi.org/10.1080/02813432.2025.2571936
31. Ahsan, Z. (2025). Integrating artificial intelligence into medical education: A narrative systematic review of current applications, challenges, and future directions. *BMC Medical Education, 25*. https://doi.org/10.1186/s12909-025-07744-0
32. Kim, S. H., Schramm, S., Riedel, E. O., Schmitzer, L., Rosenkranz, E., Kertels, O., Bodden, J., Paprottka, K., Sepp, D., Renz, M., et al. (2025). Automation bias in AI-assisted detection of cerebral aneurysms on time-of-flight MR angiography. *La Radiologia Medica, 130*, 555–566. https://doi.org/10.1007/s11547-025-01964-6
33. 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*. https://doi.org/10.1038/s41746-025-01503-7
34. Faust, O., Salvi, M., Barua, P. D., Chakraborty, S., Molinari, F., & Acharya, U. R. (2025). Issues and limitations on the road to fair and inclusive AI solutions for biomedical challenges. *Sensors, 25*. https://doi.org/10.3390/s25010205
35. As'ad, M., Faran, N., & Joharji, H. (2025). AI-supported shared decision-making (AI-SDM): Conceptual framework. *JMIR AI, 4*. https://doi.org/10.2196/75866
36. Abbasgholizadeh Rahimi, S., Cwintal, M., Huang, Y., Ghadiri, P., Grad, R., Poenaru, D., Gore, G., Zomahoun, H. T. V., Légaré, F., & Pluye, P. (2022). Application of artificial intelligence in shared decision making: Scoping review. *JMIR Medical Informatics, 10*, e36199. https://doi.org/10.2196/36199
37. Osmanodja, B., Sassi, Z., Eickmann, S., Hansen, C. M., Roller, R., Burchardt, A., Samhammer, D., Dabrock, P., Möller, S., Budde, K., et al. (2024). Investigating the impact of AI on shared decision-making in post-kidney transplant care (PRIMA-AI): Protocol for a randomized controlled trial. *JMIR Research Protocols, 13*. https://doi.org/10.2196/54857
38. Auf, H., Nygren, J., Lundgren, L. E., Petersson, L., & Svedberg, P. (2025). Healthcare professionals' perspectives on AI-driven decision support in young adult mental health: An analysis through the lens of a shared decision-making framework. *Frontiers in Digital Health, 7*. https://doi.org/10.3389/fdgth.2025.1588759
39. Medical AI and AI for medical sciences: An editorial. (2025). *JMA Journal, 8*, 38–39. https://doi.org/10.31662/jmaj.2024-0355
40. Aljuraid, R. (2025). The illusion of control: AI chatbot dependency and the threat to clinical autonomy. *Studies in Health Technology and Informatics, 332*, 211–215. https://doi.org/10.3233/SHTI251529
41. Mwogosi, A. (2025). Ethical and privacy challenges of integrating generative AI into EHR systems in Tanzania: A scoping review with a policy perspective. *Digital Health, 11*. https://doi.org/10.1177/20552076251344385
42. Pham, T. (2025). Ethical and legal considerations in healthcare AI: Innovation and policy for safe and fair use. *Royal Society Open Science*. https://doi.org/10.1098/rsos.241873
43. Xie, H., Dai, X., Xie, J., Lei, S., Zeng, J., Yang, J., & Zhou, Y. (2025). Artificial intelligence adoption in surgery: Cognition, usage patterns and implementation barriers of DeepSeek among healthcare professionals in China's tertiary hospitals. *Journal of Multidisciplinary Healthcare, 18*, 7719–7737. https://doi.org/10.2147/JMDH.S538723
44. Shiferaw, K. B., Roloff, M., Waltemath, D., & Zeleke, A. A. (2023). Guidelines and standard frameworks for AI in medicine: Protocol for a systematic literature review. *JMIR Preprints*. https://doi.org/10.2196/47105