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	EXPLAINABLE ARTIFICIAL INTELLIGENCE IN HEALTHCARE: FROM ALGORITHMIC TRANSPARENCY TO TRUST AND SOCIAL ACCEPTANCE IN CLINICAL PRACTICE
DOI	https://doi.org/10.31435/ijitss.1(49).2026.4676
RECEIVED	11 November 2025
ACCEPTED	27 January 2026
PUBLISHED	02 February 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

EXPLAINABLE ARTIFICIAL INTELLIGENCE IN HEALTHCARE: FROM ALGORITHMIC TRANSPARENCY TO TRUST AND SOCIAL ACCEPTANCE IN CLINICAL PRACTICE

Erwin Grzegorzak *(Corresponding Author, Email: erwingrzegorzak@gmail.com)*
CM Łańcut Sp. z o.o., Łańcut, Poland
ORCID ID: 0009-0004-8720-0503

Rafał Pelczar
CM Łańcut Sp. z o.o., Łańcut, Poland
ORCID ID: 0000-0002-9510-6056

Maciej Zachara
Fryderyk Chopin University Hospital, Rzeszów, Poland
ORCID ID: 0009-0008-9994-9182

Mateusz Bartoszek
Fryderyk Chopin University Hospital, Rzeszów, Poland
ORCID ID: 0000-0003-4354-2566

Patryk Harnicki
Fryderyk Chopin University Hospital, Rzeszów, Poland
ORCID ID: 0009-0008-2689-6366

Mikołaj Grodzki
Fryderyk Chopin University Hospital, Rzeszów, Poland
ORCID ID: 0009-0003-3368-3119

Jakub Minas
Regional Specialized Hospital No. 4, Bytom, Poland
ORCID ID: 0009-0008-6190-5942

Paulina Dybiak
Ludwik Rydygier Memorial Specialized Hospital, Cracow, Poland
ORCID ID: 0009-0002-4825-183X

Adrian Morawiec
Ludwik Rydygier Memorial Specialized Hospital, Cracow, Poland
ORCID ID: 0000-0001-7083-7656

Paweł Słoma
Ludwik Rydygier Memorial Specialized Hospital, Cracow, Poland
ORCID ID: 0000-0002-9856-0553

Oliwia Krawczyk
LUX MED Sp. z o.o., Cracow, Poland
ORCID ID: 0009-0005-7578-0202

ABSTRACT

Background and objective: The rapid expansion of artificial intelligence (AI) in healthcare has resulted in substantial advances in diagnostics, prognostics, clinical decision support systems, and patient monitoring. Despite promising performance, many AI-based systems remain insufficiently understood by clinicians and patients due to their "black-box" nature. This lack of transparency may undermine trust, hinder acceptance, and limit safe integration into routine clinical practice. Explainable artificial intelligence (XAI) has emerged as a response to these challenges by enabling human-interpretable explanations of algorithmic decisions. The objective of this narrative review is to synthesize current evidence on XAI in healthcare, with particular emphasis on its technical foundations, clinical applications, influence on trust and decision-making, and broader social and ethical implications.

Scope of review: This review synthesizes literature published between 2019 and 2025 addressing explainability in medical AI. The analysis includes methodological studies, clinical evaluations, human–computer interaction research, and social science investigations related to transparency, trust, accountability, and acceptance of AI systems. Relevant publications were identified through structured searches of PubMed, MEDLINE, Scopus, and Google Scholar using keywords related to explainable AI, interpretability, ethics, trust, and clinical decision support.

Findings: XAI methods—including feature attribution, model simplification, counterfactual explanations, and visualization techniques—demonstrate potential to enhance clinician understanding of AI outputs and increase confidence in algorithm-assisted decisions. Evidence suggests that explainability may support diagnostic accuracy, reduce automation bias, and facilitate error detection. However, explainability alone does not ensure trust. Clinical context, user expertise, organizational culture, and regulatory frameworks play critical roles in shaping the adoption and appropriate use of explainable systems. Empirical research addressing patient perspectives remains limited.

Conclusions: Explainable AI constitutes an important step toward the responsible and socially acceptable integration of intelligent systems in healthcare. While XAI can enhance transparency and trust, its effectiveness depends on thoughtful design, contextual adaptation to clinical workflows, and alignment with user needs. Further interdisciplinary research is required to standardize explainability approaches, evaluate their real-world impact on clinical outcomes, and address the ethical, legal, and societal challenges associated with medical AI.

KEYWORDS

Explainable Artificial Intelligence, XAI, Healthcare, Clinical Decision Support, Trust, Social Acceptance

CITATION

Erwin Grzegorzak, Rafał Pelczar, Maciej Zachara, Mateusz Bartoszek, Patryk Harnicki, Mikołaj Grodzki, Jakub Minas, Paulina Dybiak, Adrian Morawiec, Paweł Słoma, Oliwia Krawczyk (2026) Explainable Artificial Intelligence in Healthcare: From Algorithmic Transparency to Trust and Social Acceptance in Clinical Practice. *International Journal of Innovative Technologies in Social Science*. 1(49). doi: 10.31435/ijitss.1(49).2026.4676

1. Introduction

Artificial intelligence (AI) has rapidly transformed multiple domains of healthcare, including diagnostic imaging, laboratory medicine, clinical decision support systems, and population-level health analytics [1,9,11]. The increasing availability of large-scale medical datasets, advances in computational power, and the development of sophisticated machine-learning algorithms—particularly deep learning—have enabled the creation of models capable of identifying complex patterns in clinical data [1,16]. These systems have demonstrated high performance in tasks such as image interpretation, outcome prediction, and risk stratification, often matching or exceeding human-level accuracy under controlled experimental conditions [11,16].

Despite these advances, the widespread adoption of AI in routine clinical practice remains limited [9,12]. One of the principal barriers is the lack of transparency inherent in many high-performing AI models [2,13]. Deep neural networks, while powerful, are frequently described as "black boxes," providing little insight into how specific inputs are transformed into outputs [13,14]. In healthcare, where decisions directly affect patient safety, clinical outcomes, and public trust, such opacity poses substantial challenges [3,12]. Clinicians remain ethically and legally accountable for decisions made under their supervision, and reliance on systems whose

reasoning cannot be adequately explained may undermine professional confidence and increase medico-legal risk [9,10].

Explainable artificial intelligence (XAI) has emerged as a response to these concerns [2,4,6]. XAI encompasses a set of techniques and design principles aimed at rendering AI systems more transparent, interpretable, and understandable to human users [2,6,7]. In medical contexts, explainability serves multiple functions: it supports clinical reasoning, facilitates model validation and error detection, enhances trust among clinicians and patients, and enables compliance with ethical and regulatory requirements [4,5,8]. Importantly, explainability is not solely a technical property of algorithms but a socio-technical concept that intersects with cognitive psychology, human–computer interaction, ethics, and social sciences [5,8,14].

The relevance of XAI has increased alongside growing societal awareness of algorithmic decision-making and its consequences [19]. Public discourse surrounding data privacy, algorithmic bias, and accountability has underscored the need for responsible and transparent AI systems, particularly in high-stakes domains such as healthcare [10,19]. In parallel, regulatory bodies and policy initiatives have begun to emphasize transparency and explainability as core components of trustworthy AI, further reinforcing the importance of this field [15].

Against this background, the present review aims to provide a comprehensive synthesis of current knowledge on explainable AI in healthcare. By integrating evidence from technical, clinical, and social science literature, the article examines how explainability is conceptualized, implemented, and evaluated in medical AI systems [2,4,5]. Particular attention is given to the role of explainability in fostering trust, shaping user perceptions, and influencing the social acceptance of AI-driven technologies in clinical practice [8,10,18].

2. Materials and methods

This narrative review was designed to comprehensively summarize and contextualize current knowledge on explainable artificial intelligence in healthcare, with particular emphasis on clinical applicability, trust, and social acceptance [5,8,10]. Given the interdisciplinary nature of the topic, a broad and inclusive methodological approach was adopted [5].

A systematic literature search was conducted in major biomedical and multidisciplinary databases, including PubMed, MEDLINE, Scopus, and Google Scholar [11,12]. The search covered publications from January 2019 to June 2025, a period characterized by rapid methodological development and increasing clinical interest in explainable AI [1,12,16]. Additional sources were identified through manual screening of reference lists and citation tracking of key publications [6,7].

Search strategies employed combinations of keywords related to artificial intelligence and explainability—such as “explainable artificial intelligence,” “XAI,” “interpretability,” “transparency,” and “model explanation”—together with healthcare-specific terms including “medicine,” “clinical decision support,” “diagnostics,” “trust,” “ethics,” and “patient perception” [2,4,6,14]. Boolean operators and database-specific filters were applied to refine search results.

Inclusion criteria comprised peer-reviewed articles published in English that addressed explainable AI within a healthcare context [2,4,5]. Eligible studies included methodological papers describing explainability techniques with clinical relevance [6,7], empirical studies evaluating explainable AI systems in real or simulated clinical settings [8,11], qualitative and quantitative research on clinician or patient perceptions of AI explainability [8,18], and ethical, legal, or policy analyses related to transparency in medical AI [10,15,19].

Exclusion criteria included studies unrelated to healthcare, purely technical papers lacking discussion of explainability or user impact [3,13], editorials without substantive analysis, conference abstracts without full text, and studies focused exclusively on veterinary or non-human applications.

For each included publication, key information was extracted, including study design, clinical domain, type of AI model, explainability method employed, target user group, and principal findings [6–8,11]. The extracted data were synthesized narratively, with attention to consistency of results, methodological limitations, and implications for clinical practice, policy, and future research [12,18,20]. This descriptive synthesis was used to identify key trends, challenges, and gaps in the existing literature [5,12,20].

3. Definition and Conceptual Foundations of Explainable Artificial Intelligence

Explainable artificial intelligence (XAI) encompasses a broad set of concepts, methods, and design principles aimed at making AI systems understandable to human users [2,4,5]. Although the terms transparency, interpretability, and explainability are frequently used interchangeably, they denote related yet distinct constructs [2,14]. Transparency generally refers to the openness of a system's structure, assumptions, and logic; interpretability to the extent to which a human can comprehend the relationship between inputs and outputs; and explainability to the capacity of a system to provide meaningful, user-oriented explanations of its decisions [2,4,14].

In healthcare, explainability serves both epistemic and pragmatic functions [4,8]. Epistemically, it enables clinicians to assess whether an AI system's reasoning is consistent with established medical knowledge, pathophysiological understanding, and clinical intuition [8,9]. Pragmatically, explainability supports accountability, facilitates communication with patients and colleagues, and enables compliance with ethical and regulatory standards [5,10,15]. Crucially, explainability is not an intrinsic or absolute property of a model but a relational attribute that depends on the user's expertise, the clinical context, and the stakes of the decision being made [4,8,14].

3.1. Historical Development of Explainability in Medical AI

The demand for explainability in medical decision-support systems predates contemporary machine-learning approaches [9]. Early expert systems in medicine, such as rule-based diagnostic tools, were explicitly designed to provide traceable reasoning paths that clinicians could inspect, question, and override [9]. These systems prioritized transparency and justifiability, often at the expense of predictive performance.

With the emergence of statistical learning methods and, subsequently, deep learning, predictive accuracy improved substantially, but interpretability was frequently sacrificed [1,11,16]. As increasingly opaque models began to influence high-stakes clinical decisions, concerns regarding safety, bias, accountability, and trust intensified [3,10,12,13]. High-profile debates surrounding algorithmic bias, adverse events, and regulatory oversight accelerated renewed interest in explainability [10,13,19].

In parallel, the development of human-centered AI paradigms reframed explainability as a socio-technical requirement rather than a purely computational challenge [5,8,14]. Within this perspective, explanations are understood as tools for communication and sense-making, embedded within broader clinical, organizational, and societal contexts [8,19].

3.2. Taxonomy of Explainability Methods

Explainability approaches can be broadly categorized into intrinsic and post-hoc methods [2,6,7,14]. Intrinsic explainability is achieved through models that are interpretable by design, such as linear regression, generalized additive models, decision trees, and rule-based systems [4,14]. These models allow users to directly inspect relationships between inputs and outputs and are particularly well suited to tabular clinical data, risk scores, and guideline-based reasoning [9,11].

Post-hoc explainability methods are applied after model training and are especially relevant for complex architectures, including deep neural networks [2,6,7]. Feature attribution techniques, such as SHAP and LIME, estimate the contribution of individual variables to model predictions [6,7]. Visualization methods, including saliency maps and activation mappings, are widely used in image-based diagnostics to highlight regions associated with predictions [1,16]. Counterfactual explanations describe how minimal changes to inputs could alter model outputs, thereby supporting clinical reasoning and shared decision-making [4,8].

Each category involves trade-offs between fidelity to the underlying model, comprehensibility for users, and computational cost [2,14]. In healthcare, the clinical relevance, cognitive usability, and contextual appropriateness of explanations are often more important than exhaustive technical detail [8,12].

3.3. Human Factors and Cognitive Considerations

Explainability must be interpreted in light of human cognitive processes and clinical working conditions [8]. Clinicians routinely operate under time pressure, uncertainty, and information overload [8,9]. Explanations that are overly complex, poorly timed, or misaligned with clinical reasoning may hinder rather than support decision-making [8,12]. Research in human–computer interaction highlights the importance of concise, context-aware explanations that facilitate rapid sense-making and support clinical workflows [8,14].

Furthermore, different stakeholders require different forms of explanation [5,8,10]. Clinicians, patients, administrators, and regulators have distinct informational needs, goals, and levels of expertise [10,15,18].

Effective XAI systems therefore require adaptive explanation strategies that can be tailored to user roles, clinical tasks, and decision contexts [5,8].

These considerations underscore that explainability is not solely a property of algorithms but a broader design and integration challenge, involving user interfaces, workflows, and communication practices embedded within healthcare environments [5,8,19].

4. Applications of Explainable Artificial Intelligence in Healthcare

Assessing the effectiveness of explainable artificial intelligence (XAI) in healthcare requires evaluation frameworks that extend beyond conventional measures of predictive performance [8,12]. Traditional metrics, such as sensitivity, specificity, and the area under the receiver operating characteristic curve, do not capture whether explanations are understandable, clinically meaningful, or safe for real-world decision-making [3,12]. Consequently, increasing attention has been directed toward multidimensional evaluation strategies tailored to the specific cognitive, organizational, and ethical demands of clinical environments [8,18].

Key dimensions of explainability evaluation include fidelity, defined as the extent to which explanations accurately reflect underlying model behavior [2,14]; interpretability, referring to the ease with which users can comprehend explanations [2,8]; usefulness, indicating whether explanations meaningfully support clinical decision-making [8,9]; and impact on outcomes, encompassing changes in clinician behavior, diagnostic accuracy, error rates, or patient outcomes [11,12,18]. Evaluation methods range from controlled user studies and simulation-based experiments to observational analyses conducted in real-world clinical settings [8,20].

Human-centered evaluation is particularly critical [8,14]. Surveys, interviews, and task-based assessments provide insight into how clinicians perceive, interpret, and use explanations under realistic conditions characterized by time pressure, uncertainty, and information overload [8,9]. Available evidence suggests that explanations perceived as clinically relevant, concise, and aligned with medical reasoning are more likely to be adopted, whereas overly verbose or technically dense explanations are often ignored [8,12]. Despite growing interest, the absence of standardized evaluation methodologies remains a major limitation, hindering cross-study comparability and complicating regulatory assessment [12,15].

4.1. Diagnostic Support Systems

Diagnostic support represents one of the most prominent application areas of explainable AI in healthcare [1,11,16]. Deep learning models have demonstrated high performance in radiology, pathology, ophthalmology, cardiology, and dermatology, particularly in tasks involving pattern recognition in complex visual data [1,16]. Nevertheless, clinical deployment has been constrained by persistent concerns regarding interpretability, reproducibility, and accountability [12,13].

Explainability techniques such as saliency maps, class activation mappings (e.g., Grad-CAM), and attention-based visualizations are commonly employed to highlight image regions that contribute most strongly to model predictions [6,7,16]. These visual explanations may assist clinicians in validating algorithmic outputs, identifying potential artifacts, and understanding model failure modes [8,12]. Several studies suggest that explainable diagnostic outputs can enhance diagnostic confidence and support learning, particularly among junior clinicians and trainees [8].

However, the clinical value of visual explanations remains contested [3,13]. Empirical analyses indicate that saliency maps may reflect statistical correlations rather than causally relevant features, potentially leading to misleading interpretations [3,14]. Furthermore, variability in explanations across model architectures or minor input perturbations raises concerns regarding robustness and reliability [6,7]. As a result, current consensus emphasizes that explainable diagnostic systems should function as decision-support tools rather than autonomous decision-makers, with human oversight remaining essential [9,12].

4.2. Predictive Modeling and Risk Stratification

Predictive modeling plays a central role in contemporary healthcare by estimating the likelihood of adverse outcomes such as disease progression, hospital readmission, complications, or mortality [11,12]. In these contexts, explainability is particularly important to ensure that predictions are clinically meaningful, ethically acceptable, and aligned with established medical knowledge [4,10].

Feature attribution methods, including SHAP and LIME, are widely used to quantify the contribution of individual variables to model outputs [6,7]. These techniques enable clinicians to assess whether predictions are driven by clinically relevant factors consistent with pathophysiological understanding [8,9]. For example,

explainable models predicting sepsis or cardiovascular risk may reveal whether key variables—such as vital signs, laboratory values, or comorbidities—appropriately influence risk estimates [11].

Evidence suggests that explainable predictive models can enhance clinician trust and facilitate shared decision-making [8,18]. Nevertheless, challenges persist in balancing explanatory detail with cognitive usability [8,12]. Overly detailed explanations may overwhelm users, whereas excessively simplified summaries risk obscuring uncertainty or model limitations [12,14]. User-centered design principles and iterative evaluation are therefore essential to ensure that explanations remain both accurate and cognitively manageable [8,14].

4.3. Clinical Decision Support Systems

Clinical decision support systems (CDSS) integrate patient-specific data, evidence-based guidelines, and predictive analytics to generate recommendations for diagnosis, treatment, or monitoring [9,11]. Explainability within CDSS is critical to mitigate automation bias, preserve clinician autonomy, and maintain accountability for clinical decisions [9,10,13].

Explainable CDSS can provide rationale-based recommendations, highlight key factors influencing suggested actions, and present alternative options when appropriate [8,9]. Empirical studies indicate that such systems may improve adherence to clinical guidelines and reduce unwarranted variability in care [9,11]. However, successful adoption depends on seamless integration into clinical workflows and alignment with professional norms and expectations [12,20].

Barriers to implementation include concerns regarding increased cognitive load, workflow disruption, and medico-legal liability [12,13]. Importantly, explainability alone does not guarantee acceptance [10,12]. Organizational support, targeted training, and clearly defined governance structures are equally necessary to ensure the responsible and effective use of explainable CDSS in clinical practice [10,15,20].

4.4. Public Health and Population-Level Applications

Beyond individual patient care, explainable AI has gained increasing relevance in public health surveillance, epidemiological modeling, and health-system planning [17,19]. In these contexts, algorithmic decisions often affect large populations and involve the allocation of limited resources, rendering transparency and accountability particularly critical [10,15,19].

Explainable models have been used to forecast disease incidence, predict hospital and intensive care unit demand, and optimize the distribution of medical resources [17]. During infectious disease outbreaks, including the COVID-19 pandemic, explainable predictive models supported scenario analysis and policy decision-making by making assumptions and key drivers of forecasts explicit [17,18]. Such transparency enabled policymakers and clinicians to scrutinize model behavior, assess uncertainty, and communicate decision rationales to the public [15,17].

At the health-system level, explainable AI supports quality improvement initiatives, risk adjustment, and performance benchmarking [12,20]. Transparent models facilitate auditing processes and enable stakeholders to understand how metrics are generated, thereby reducing skepticism and resistance [19,20]. Importantly, explainability contributes to public trust by demonstrating that algorithmic decisions remain subject to human oversight and rational justification [10,15].

4.5. Integration of Explainable AI into Clinical Workflows

A critical determinant of the success of explainable AI in healthcare is its integration into existing clinical workflows [12,20]. Even highly accurate and interpretable models may fail to achieve adoption if they disrupt established practices or increase cognitive burden [8,12]. Effective workflow integration therefore requires careful consideration of technical, organizational, and human factors [8,20].

Explainable AI systems must deliver explanations at the appropriate time, in a suitable format, and to the intended user [8,14]. For instance, real-time decision support in emergency settings requires concise and rapidly interpretable explanations, whereas longitudinal care planning may benefit from more detailed and exploratory insights [8]. User interface design plays a central role in ensuring that explanations are accessible, actionable, and aligned with clinical reasoning [14].

Organizational alignment is equally essential [10,20]. Healthcare institutions must define clear use cases, responsibilities, and escalation pathways for AI-supported decisions [10,15]. Explainability facilitates this process by clarifying how recommendations are generated and identifying situations in which human intervention is required [5,15]. Successful integration typically involves iterative testing, continuous feedback from end users, and ongoing system refinement [20].

5. Trust, Perception, and Social Acceptance

Trust constitutes a fundamental prerequisite for the effective and responsible integration of artificial intelligence into healthcare [9,10,12]. It influences whether clinicians adopt AI-based tools, how they rely on algorithmic recommendations, and whether patients accept AI-supported care [8,18]. Trust in medical AI is a multidimensional construct encompassing perceptions of technical competence, reliability, transparency, fairness, and alignment with professional norms and societal values [5,10,19]. Explainability plays a central—though not exclusive—role in shaping these perceptions [2,4,8].

5.1. Conceptual Models of Trust in Medical AI

Trust in medical AI is commonly conceptualized within socio-technical frameworks that integrate human, technological, and organizational dimensions [5,8,19]. These models typically distinguish between dispositional trust, reflecting a general attitude toward technology; situational trust, which is context-dependent and varies with task criticality and uncertainty; and learned trust, shaped by prior interactions with AI systems and observed outcomes [8]. Explainability primarily influences situational and learned trust by providing users with cues about system reasoning, reliability, and limitations [2,8,14].

In healthcare, calibrated trust—defined as appropriate reliance that avoids both blind acceptance and unwarranted skepticism—is essential [13]. Explainable AI can support trust calibration by revealing uncertainty, highlighting relevant decision factors, and encouraging critical appraisal of algorithmic outputs [4,8]. However, explainability alone is insufficient [3,12]. Effective trust calibration also depends on consistent system performance, transparency regarding performance boundaries, and alignment with clinical goals and professional standards [9,10,12].

5.2. Clinician Perspectives on Explainable AI

From the clinician's perspective, explainable AI is primarily valued as a tool that supports, rather than replaces, clinical reasoning [8,9]. Physicians consistently report that understanding why an AI system produces a particular recommendation is crucial for integrating algorithmic outputs into their own decision-making processes [8,9]. Explanations enable clinicians to assess plausibility, detect potential errors, and determine whether AI-generated insights are applicable to a specific clinical context [8,12].

Empirical studies across multiple specialties indicate that explainability is particularly important in complex or high-risk decisions, such as oncology diagnostics, intensive care management, and treatment prioritization [8,11]. In such scenarios, clinicians are less willing to rely on opaque systems, even when predictive accuracy is high [13]. Explainable models may also reduce resistance to adoption by addressing concerns related to loss of professional autonomy, deskilling, and medico-legal responsibility [9,10,13].

Nevertheless, clinician expectations regarding explainability remain heterogeneous [8]. Senior physicians often prefer high-level, concept-based explanations aligned with pathophysiological reasoning, whereas trainees and junior clinicians may benefit from more detailed, feature-level explanations [8,14]. This variability underscores the need for adaptable explanation interfaces that can be tailored to user expertise, clinical role, and decision context [5,8].

5.3. Patient Perspectives and Communication

Patients represent a critical yet comparatively understudied stakeholder group in the implementation of explainable AI in healthcare [18]. Although patients are rarely direct users of AI systems, they are directly affected by AI-supported clinical decisions [10,18]. Trust in healthcare is closely linked to perceived transparency, empathy, and accountability, all of which may be influenced by how AI is incorporated into care delivery [10,19].

Available evidence suggests that patients generally do not require detailed technical explanations of AI algorithms [18]. Instead, they value clear communication regarding the role of AI in decision-making, assurances of human oversight, and an understanding of how AI contributes to improved care [10,18]. Explainability therefore plays an indirect but important role by enabling clinicians to communicate AI-supported decisions in a comprehensible and reassuring manner [8,10].

Common patient concerns include data privacy, algorithmic bias, and the potential depersonalization of care [10,15,19]. Addressing these concerns requires transparent governance, clear consent processes, and clinician-mediated explanations that emphasize human responsibility and oversight in AI-assisted decision-making [10,15].

5.4. Communication Strategies and Explanation Design

The effectiveness of explainable AI depends not only on the availability of explanations but also on how they are communicated [8,14]. Explanation design should consider language, visual representation, timing, and relevance to the clinical task [8,14]. Narrative explanations, graphical summaries, and comparative baselines have been proposed as strategies to enhance comprehension and usability [14].

Clinicians generally prefer concise explanations that map onto familiar clinical concepts, such as risk factors, pathophysiology, or guideline-based reasoning [8,9]. Layered explanation approaches, which allow users to access increasing levels of detail on demand, have demonstrated particular promise in accommodating diverse expertise levels without increasing cognitive load [14]. By contrast, poorly designed explanations may increase workload, be ignored, or even undermine trust [8,12].

5.5. Cultural and Institutional Influences on Acceptance

Acceptance of explainable AI is shaped by broader cultural norms, professional identities, and institutional trust [10,15,19]. Healthcare systems characterized by strong traditions of evidence-based practice, transparent governance, and organizational learning may be more receptive to explainable technologies [12,20]. Conversely, environments marked by high workload, limited resources, or low institutional trust may exhibit greater resistance, regardless of technical performance [12,19].

International perspectives further reveal substantial variability in expectations regarding transparency, accountability, and human oversight [15,19]. These differences highlight the importance of context-sensitive implementation strategies and caution against one-size-fits-all approaches to explainability in healthcare [5,15].

5.6. Measuring Trust and Social Acceptance

Assessing trust and social acceptance of explainable AI requires validated instruments and mixed-methods approaches [8,18]. Quantitative surveys measuring perceived usefulness, ease of use, and trustworthiness are commonly employed, while qualitative methods—such as interviews, focus groups, and ethnographic observation—provide deeper insight into real-world interpretation and use [8,18].

A key research priority remains the linkage of subjective trust measures to objective outcomes, including adherence to AI recommendations, diagnostic accuracy, error rates, and patient outcomes [12,18]. Establishing such links is essential for determining whether explainability meaningfully improves clinical practice rather than merely shaping user perceptions [3,12].

6. Ethical, legal, and organizational challenges

The ethical, legal, and organizational dimensions of explainable artificial intelligence constitute a decisive layer determining whether XAI can be safely and effectively integrated into healthcare systems [10,12,15]. Although explainability is frequently proposed as a solution to the ethical risks posed by opaque algorithms, it should be understood as a facilitating condition rather than a standalone safeguard [3,13].

Ethically, explainable AI is closely linked to the principle of respect for autonomy [10,19]. In clinical practice, informed consent relies on the ability of patients to understand the rationale behind diagnostic and therapeutic decisions [10]. When AI systems contribute to these decisions, explainability enables clinicians to articulate the role of algorithms in care pathways [5,8]. This is particularly relevant in personalized medicine, oncology, and critical care, where AI-driven risk stratification may influence treatment intensity or eligibility for advanced therapies [11,18].

Non-maleficence and beneficence are also directly implicated [10,18]. Explainable systems may support early identification of erroneous predictions, inappropriate correlations, or unsafe recommendations, thereby reducing the risk of harm [4,12]. At the same time, there is a danger that superficial or misleading explanations may create an illusion of understanding, potentially obscuring uncertainty and limitations [3,13]. Ethical deployment of XAI therefore requires careful validation of explanation methods and explicit communication of uncertainty [4,14].

Justice and fairness represent additional ethical challenges [10,19]. Algorithmic bias arising from unrepresentative training data or structural inequities can disproportionately affect vulnerable populations [10,19]. Explainable AI can assist in identifying such biases by revealing which variables drive predictions across demographic groups [4,18]. However, bias mitigation demands systemic interventions, including diverse datasets, inclusive design processes, and continuous auditing [10,12,19]. Without these measures, explainability risks functioning as a symbolic gesture rather than an effective ethical control [3,19].

From a legal perspective, explainability is increasingly recognized as a prerequisite for accountability in automated decision-making [13,15]. In most jurisdictions, clinicians remain legally responsible for patient care, even when AI systems are used as decision-support tools [9,10]. The inability to explain algorithmic outputs complicates liability attribution and may deter adoption [12,13]. Regulatory initiatives, particularly within the European Union, emphasize transparency, traceability, and human oversight for high-risk AI applications in healthcare [15].

Despite these developments, legal standards for explainability remain heterogeneous and evolving [13,15]. Ambiguity regarding what constitutes a sufficient explanation poses challenges for developers and healthcare organizations [3,13]. Clearer guidance is needed to balance innovation with legal certainty and patient protection [15].

Organizational readiness is a critical determinant of successful XAI implementation [12,20]. Healthcare institutions face constraints related to infrastructure, interoperability, financial resources, and workforce expertise [12,20]. Explainable systems must be integrated into existing clinical workflows without increasing cognitive burden or disrupting care delivery [8,12]. Failure to address organizational factors may undermine even technically robust solutions [20].

Education and training represent key organizational enablers [12,20]. Clinicians require foundational knowledge of AI principles, strengths, and limitations, as well as practical guidance on interpreting explanations [8,11]. Interdisciplinary collaboration among clinicians, data scientists, ethicists, and administrators is essential to ensure that explainable AI systems are aligned with clinical realities, ethical norms, and institutional goals [5,10,20]. Education and training initiatives should be developed to enhance AI literacy among healthcare professionals, enabling informed and critical engagement with explainable technologies [11,12].

7. Discussion

The findings synthesized in this narrative review demonstrate that explainable artificial intelligence occupies a pivotal position in the ongoing digital transformation of healthcare. As AI systems increasingly influence diagnostic reasoning, prognostic assessment, and therapeutic decision-making, expectations regarding transparency, accountability, and legitimacy become more pronounced. Explainability emerges not merely as a technical enhancement but as a foundational requirement for aligning algorithmic systems with clinical practice, ethical principles, and broader social expectations.

From a clinical perspective, explainable AI has the potential to enhance human–AI collaboration by supporting reflective reasoning rather than automation-driven substitution. Explanations act as cognitive scaffolds that enable clinicians to interrogate algorithmic outputs, contextualize recommendations within patient-specific circumstances, and retain responsibility for final decisions. This collaborative paradigm is consistent with prevailing ethical norms in medicine, which emphasize professional autonomy, accountability, and patient-centered care.

At the same time, the reviewed literature cautions against the uncritical adoption of explainability techniques. Not all explanations are equally meaningful, and poorly designed explanations may mislead users, increase cognitive load, or foster inappropriate reliance on AI systems. In some contexts, overly persuasive or overly simplified explanations may undermine trust calibration by creating a false sense of certainty. These risks underscore the need for rigorous evaluation of explanation fidelity, usability, and clinical impact, as well as for user-centered design approaches tailored to diverse clinical roles and levels of expertise.

From a broader social and organizational perspective, explainable AI contributes to the legitimacy of digital healthcare innovations. Transparent decision-making processes facilitate accountability, support regulatory oversight, and enable informed public discourse regarding the role of AI in medicine. However, explainability alone is insufficient to secure social acceptance. Its benefits must be embedded within comprehensive frameworks of ethical governance, data protection, equity, and institutional responsibility. Without such integration, explainability risks functioning as a symbolic feature rather than a substantive safeguard.

8. Limitations and Future Directions

Despite the rapidly expanding literature on explainable artificial intelligence in healthcare, several limitations constrain current understanding and practical implementation [3,12]. First, a substantial proportion of existing evidence is derived from laboratory-based experiments, simulations, or retrospective analyses [8,12]. Although these approaches offer valuable insights into technical feasibility and user perceptions, they may not adequately capture the complexity of real-world clinical environments, where time pressure, competing priorities, and workflow constraints strongly shape technology use [8,20].

Second, there is considerable heterogeneity in how explainability is conceptualized, implemented, and evaluated across studies [2,4,14]. Researchers employ diverse explanation methods, user populations, and outcome measures, limiting the comparability and generalizability of findings [12,14]. The absence of standardized definitions and evaluation frameworks hampers evidence synthesis and complicates regulatory assessment of explainable AI systems [12,15].

Third, empirical research has focused predominantly on clinician users, with comparatively limited attention to patient-centered outcomes, long-term safety, and system-level effects [8,18]. Patient perspectives—particularly those of vulnerable or marginalized populations—remain underrepresented [10,18]. Moreover, relatively few studies have examined whether explainable AI leads to measurable improvements in clinical outcomes, reductions in error rates, or gains in healthcare efficiency when compared with non-explainable systems [3,12].

Future research should prioritize prospective, real-world evaluations of explainable AI systems integrated into routine clinical practice [12,20]. Such studies should assess not only predictive performance but also impacts on clinician behavior, workflow integration, patient experience, and health outcomes [8,18,20]. Controlled implementation studies may be particularly valuable in clarifying causal relationships between explainability and clinical impact [12].

Another key research direction involves the development of standardized frameworks and metrics for evaluating explainability in healthcare [2,14]. These frameworks should encompass technical fidelity, usability, cognitive load, trust calibration, and ethical implications [4,8,14]. Consensus-driven standards would facilitate cross-study comparison, support regulatory approval processes, and guide developers in designing clinically meaningful explanations [15].

Interdisciplinary collaboration will be essential to advance the field [5,19]. Integrating perspectives from medicine, computer science, psychology, ethics, law, and social sciences can inform more holistic and socially responsive approaches to explainable AI [5,19]. In parallel, education and training initiatives should be expanded to improve AI literacy among healthcare professionals, enabling informed and critical engagement with explainable systems [11,12].

Finally, regulatory and policy frameworks must evolve alongside technological innovation [10,15]. Clear guidance on transparency requirements, accountability, validation, and post-deployment monitoring will be critical to ensuring that explainable AI contributes to equitable, safe, and effective healthcare [15,18]. Sustained dialogue among clinicians, patients, developers, and policymakers will be necessary to align explainable AI with societal values and expectations [10,19].

9. Conclusions

Explainable artificial intelligence represents a critical pillar of responsible, sustainable, and socially acceptable deployment of AI in healthcare. By enhancing transparency, supporting calibrated trust, and enabling accountability, XAI provides a pathway for reconciling the growing capabilities of machine-learning systems with the ethical, legal, and social demands of clinical practice.

The evidence reviewed in this article indicates that explainable AI can strengthen clinician confidence, improve human–AI collaboration, and support safer, more reflective decision-making. By rendering algorithmic reasoning accessible, explanations allow clinicians to critically appraise AI outputs, integrate them with contextual clinical knowledge, and retain responsibility for final decisions. At the same time, this review underscores that explainability should not be regarded as a universal solution to all challenges associated with medical AI. Its effectiveness depends on thoughtful design, rigorous evaluation, appropriate contextualization, and meaningful integration within clinical workflows and organizational structures.

To fully realize the potential of explainable AI, sustained interdisciplinary research, standardized evaluation approaches, and robust governance frameworks are required. When implemented responsibly and embedded within broader ethical and institutional safeguards, explainable AI systems may contribute to improved quality of care, enhanced patient safety, and increased public trust in digital healthcare technologies.

9.1. Implications for Clinical Practice

The practical implications of explainable artificial intelligence for everyday clinical practice are substantial. When appropriately implemented, explainable systems can support safer decision-making, enhance interdisciplinary communication, and strengthen clinician confidence in AI-assisted tools. Explanations aligned with established clinical reasoning enable physicians to contextualize algorithmic outputs, reconcile them with patient-specific factors, and communicate decisions effectively to patients and colleagues.

Explainable AI may also facilitate documentation and medico-legal defensibility by providing traceable rationales for AI-supported recommendations. In academic and teaching hospitals, explainable systems can serve as educational resources, promoting reflective learning and discussion among trainees. These benefits, however, depend on adequate training, institutional support, and a clear delineation of responsibility between human decision-makers and algorithmic systems.

9.2. Implications for Health Policy and Regulation

At the policy level, explainable AI has important implications for governance, regulation, and public accountability. Policymakers increasingly recognize the need for transparency in high-risk AI applications, particularly those deployed in healthcare. Explainability can support regulatory oversight by enabling auditing, validation, and post-deployment monitoring of AI systems.

Health authorities may also leverage explainable models to inform policy decisions, allocate resources, and communicate decision rationales to the public. To achieve these objectives, regulatory frameworks must provide clear guidance on transparency requirements, documentation standards, and the respective responsibilities of developers and healthcare providers. Greater international harmonization of standards may further facilitate innovation while ensuring patient safety and protection.

9.3. Implications for Future Research and Innovation

Future research and innovation in explainable AI should prioritize user-centered design, interdisciplinary collaboration, and real-world validation. Advances in interactive and adaptive explanation systems may allow explanations to evolve in response to user feedback, clinical context, and changing decision needs. Further investigation is also required to clarify how explainability interacts with other trust-enhancing mechanisms, such as performance guarantees, certification processes, and ethical governance structures.

Long-term impacts of explainable AI—including effects on clinical outcomes, professional roles, workflow efficiency, and health-system performance—remain insufficiently explored. Addressing these gaps will be essential for developing a comprehensive understanding of both the value and limitations of explainable AI in healthcare and for guiding its responsible and socially beneficial adoption.

10. List of Abbreviations

AI – Artificial Intelligence
XAI – Explainable Artificial Intelligence
ML – Machine Learning
CDSS – Clinical Decision Support System
HCI – Human–Computer Interaction
GDPR – General Data Protection Regulation
SHAP – Shapley Additive Explanations
LIME – Local Interpretable Model-Agnostic Explanations

Author's contribution:
Conceptualization and methodology – Erwin Grzegorzak
Check - Paweł Słoma, Erwin Grzegorzak
Resources – Erwin Grzegorzak
Investigation - Maciej Zachara, Erwin Grzegorzak
Form analysis - Adrian Morawiec, Erwin Grzegorzak
Writing - rough preparation - Patryk Harnicki, Mateusz Bartoszek, Erwin Grzegorzak
Writing - review and editing – Paulina Dybiak Supervision - Oliwia Krawczyk, Erwin Grzegorzak
Visualization - Rafał Pelczar, Erwin Grzegorzak
Project administration - Mikołaj Grodzki, Jakub Minas, Erwin Grzegorzak

All authors have read and agreed with the published version of the manuscript.
Funding statement: The study did not receive special funding.
Conflict of interest statement: The authors declare that they have no conflict of interest.

REFERENCES

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019;25(1):44–56. doi:10.1038/s41591-018-0300-7.
2. Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. ITU Journal. 2019;1:39–48.
3. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digital Health. 2021;3(11):e745–e750. doi:10.1016/S2589-7500(21)00208-9.
4. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery. 2019;9(4):e1312.
5. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making. 2020;20:310.
6. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. Proc KDD. 2016:1135–1144.
7. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. NeurIPS. 2017;30:4765–4774.
8. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. MLHC. 2019.
9. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA. 2018;320(21):2199–2200.
10. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Medicine. 2018;15(11):e1002689.
11. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019;380(14):1347–1358.
12. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine. 2019;17:195.
13. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Center Report. 2019;49(1):15–21.
14. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv. 2017.
15. European Commission. Ethics guidelines for trustworthy AI. 2019.
16. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nature Medicine. 2019;25(1):24–29.
17. van der Schaar M, et al. How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. Machine Learning. 2021;110:1–14.
18. McCradden MD, Joshi S, Anderson JA, Goldenberg A. Patient safety and quality improvement: ethical principles for a regulatory approach to AI in healthcare. PLOS Medicine. 2020;17(7):e1003242.
19. Mittelstadt BD, et al. The ethics of algorithms: Mapping the debate. Big Data & Society. 2016;3(2):1–21.
20. Sendak MP, D'Arcy J, Kashyap S, Gao M, Nichols M, Corey K, Ratliff W, Balu S. A path for translation of machine learning products into healthcare delivery. EMJ Innovations. 2020;4:19–29.