Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

BUILDING DIGITAL TRUST IN AI-POWERED HEALTHCARE INFORMATION SYSTEMS: PRIVACY, SECURITY, AND GOVERNANCE CHALLENGES

SUPOM ROY
Bachelor of Science in Information Systems (BSIS), Trine University, Angola, Indiana, USA.
*Correspondent Author Email: sroy25@my.trine.edu

MAHEE AHMED CHOUDHURY
Master's in Information Technology, University of The Cumberlands, Williamsburg, Kentucky, USA.
Email: mchoudhury78170@ucumberlands.edu

RAIHAN UDDINE
Bachelor of Science in Information Systems (BSIS), Trine University, Angola, Indiana, USA.
Email: ruddine25@my.trine.edu

JILLUR BIN KUTUB
Bachelor of Science in Information Systems (BSIS), Trine University, Angola, Indiana, USA.
Email: jkutub26@my.trine.edu

Abstract

AI technologies are becoming a fundamental element of healthcare information systems in areas of diagnostic imaging, predictive analytics, telemedicine, and continuous patient monitoring. Nevertheless, AI adoption has moved ahead of privacy protections, security architectures, and governance structures that would allow maintaining trust among stakeholders. This systematic review combines a broader collection of relevant core studies and their supporting literature (2015-2026), including conceptual reviews, expert-derived frameworks, AI-scenario experiments, security case syntheses, and analysis of legislations, to summarize what is known about the establishment of digital trust in healthcare with AI. Several main themes can be derived from the study. First, regardless of the methodology, all the core studies are unanimous about the socio-technic nature of trust emerging through the interrelationship between technological security measures, organizational governance, and human factors. Second, privacy is recognized as the major challenge to AI implementation, and the regulatory baseline for trust frameworks is incomplete because there is no statutory basis to address many aspects of contemporary AI healthcare in instruments such as HIPAA and it differs from one jurisdiction to another. Third, explainability serves as the precondition of trust but becomes the most recommended yet the least proven solution, thus leaving open the borders of allowable clinical automation. Overall, the review highlights a structural contradiction: while there is much enthusiasm about adoption and effectiveness of the technology, primary sources confirm the fragmentation of governance at the state level and the decreasing public confidence. The main message is that validation is needed: few of the frameworks discussed have been experimentally verified in a deployed clinical system; there is no standard psychometric tool to measure the trust generated by frameworks; and finally, patients as the bearers of trust have not been studied.

Keywords: Digital Trust; Artificial Intelligence; Healthcare Information Systems; Privacy; Security; AI Governance; Explainable AI; Socio-Technical Systems; Health Policy.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

1. INTRODUCTION

1.1 Context: The AI Transformation of Healthcare Information Systems

Within roughly a decade, artificial intelligence (AI) has shifted from an auxiliary tool to an integral part of information systems in use within the healthcare industry. The amount of research evidence underlying this transformation is sizable and continues to increase. Deep learning solutions that employ Convolutional Neural Networks (CNNs) exhibit highly specialized performance in detecting cancer, heart conditions, and retinal issues through medical imaging. At the same time, machine learning algorithms featuring Recurrent Neural Networks (RNNs) and Transformer architectures forecast patients' likelihood to get readmitted to hospitals, progression of diseases, and their deteriorating state based on EHRs. NLP algorithms automate the process of generating diagnostic codes out of physician's notes. AI is also instrumental in performing adjacent clinical and administrative functions, from discovering new drugs to robotic surgery and management of hospital supply chain (Topol, 2019; Adwer & Whiting, 2024).

In accordance with this diversity, the current estimates of adoption indicate 85% of diagnostics, 78% of predictive analytics, 72% of workflow optimization, 65% of virtual health assistants, and 60% of administrative back-office operations (Orthi et al., 2022). The development was rapid rather than linear: total AI adoption in healthcare increased from an estimated 30% in 2015 to a projected 85% by 2025, with the coronavirus pandemic serving as a catalyst for AI-driven patient monitoring, telemedicine, and crisis management of epidemics (Obermeyer & Emanuel, 2016; Wahl et al., 2018). Examples of benchmark studies demonstrate the rate of advancement: deep-learning systems have outperformed board-certified dermatologists in classifying skin cancer (Esteva et al., 2017), ophthalmologists in assessing diabetic retinopathy (Gulshan et al., 2016), and radiologists in some chest-pathology detection tasks (Rajpurkar et al., 2017), each one setting new bar for algorithmic capabilities outside of lab settings.

Technological foundations for this shift have developed within the process of digital transformation of health infrastructure. The implementation of electronic health records (EHRs), under the Health Information Technology for Economic and Clinical Health (HITECH) Act to replace paper systems, has provided the backbone that makes it possible for today’s artificial intelligence (AI) technologies to function (Adler-Milstein & Jha, 2017). While the issue of potential misuse of the digitized patient data has always been present, the scale of such risks has increased. EHRs have become deeply connected to the telemedicine and mobile health (mHealth) application solutions that lack any privacy policy or whose policies are inadequate (Grundy et al., 2019) and to the ubiquitous health (uHealth) environment. In the context of the latter, patients are being constantly and passively monitored by wearable biosensors, home medical smart devices, and IoT-based medical equipment (Dang et al., 2019). Analysis of 36 of the most popular mHealth applications in 2019 revealed that 79% of them have shared users’ personal data with third parties and only less than half of the apps disclose the data sharing in the privacy policies (Grundy et al., 2019). At each stage of this chain, there appears another way of generating, transmitting, storing, and reusing the sensitive

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

protected health information (PHI). The industry has had to deal with the problem directly: in 2023, 725 large data breaches were registered by the HHS Office for Civil Rights, and more than 133 million PHI were exposed (U.S. Department of Health and Human Services, 2024).

In this respect, academic attention towards the issue has followed the progress in the development of the technology. According to bibliometric analysis of the Scopus database (N = 2,054), there has been limited interest of researchers into the area until 2000, which began to grow steadily in 2015 and accelerated significantly since 2020 (Figure 1), reflecting the overall growth trends in AI research publication indexed in that period of time (Mushtaq & Hameeda, 2025). WHO published its first-ever report dedicated to the ethics and governance of artificial intelligence in health care in 2021 and advised not to overestimate the role of AI technologies as techno-solutionism might lead to the distraction from the essential problem of universal health coverage (WHO, 2021). Macroeconomic forecasts of PricewaterhouseCoopers suggest an addition of US$13 trillion to the gross output of the world economy in 2030 due to the adoption of AI technologies in different industries, among which the most significant contribution can be expected from the healthcare sector (PwC, 2018). Qualitative interviews of public-health professionals predict that "AI will influence everything in society," thus requiring a reconsideration of the core functions and methodology of public health surveillance (Wahl et al., 2018).

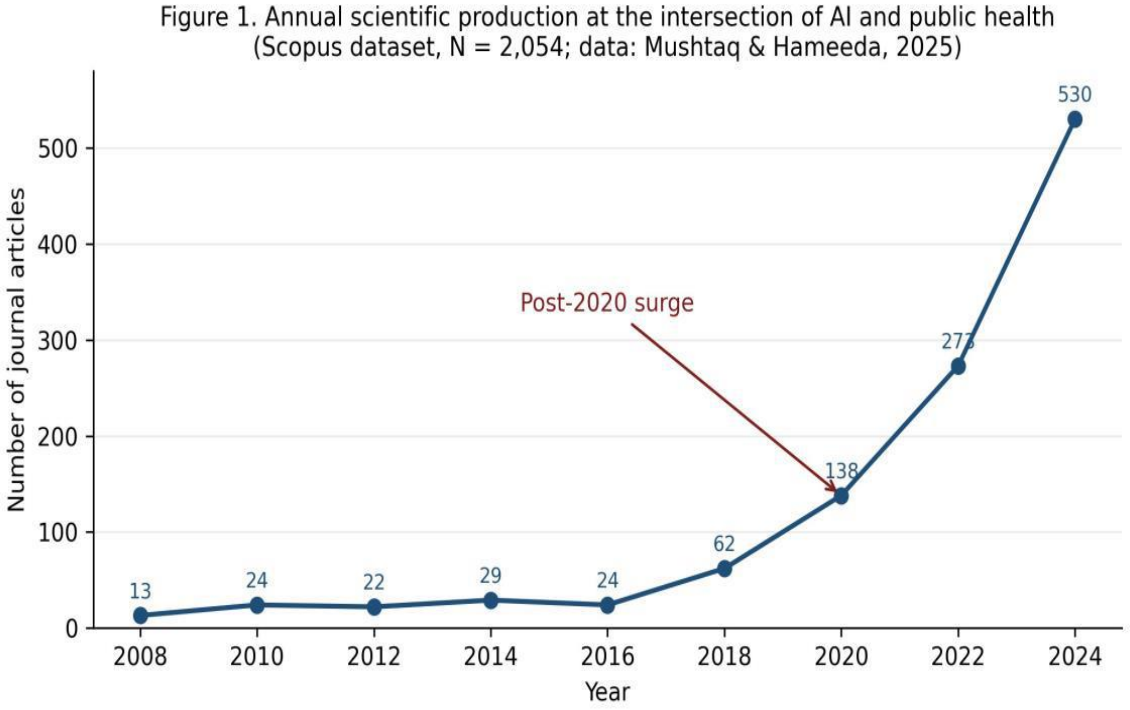

Figure 1: Annual scientific production at the intersection of AI and public health (Scopus dataset, N = 2,054). Data: Mushtaq & Hameeda, 2025

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

1.2 Motivation: Why Trust Is the Binding Constraint on Clinical Adoption

Trustworthy autonomous and semi-autonomous AI systems is a multi-dimensional concept that includes system safety, cryptography security, data privacy, algorithmic transparency, and demographic fairness. As far as responsible AI research goes, it should be noted that all of these components take on greater importance in healthcare as the risk of patients suffering harm from errors becomes real. While earlier research into healthcare and AI (2012-2019) was concerned with determining what AI can accomplish in laboratory settings, current literature suggests that its limitations in practice have more to do with digital trust than capabilities.

The data barriers can be noted as converging in various independent approaches and data sources. Specifically, in the systematic review by Ahmed et al. (2023), in which 37 peer-reviewed articles concerning implementation of AI in healthcare were reviewed, data privacy was found to be the most serious barrier, evaluated at 85% severity. Other serious barriers include algorithmic bias (75%), black-box issue or lack of transparency (70%), institutional resource shortages (65%), and need for workforce training (60%). The results of organizational survey provided by Adwer and Whiting (2024) show that even though 78% of organizations recognize the importance of AI adoption for their digital transformations, only 23% have managed to implement it effectively. Among all of the challenges mentioned in their survey, security concerns with data is identified as the most frequent one with 67% prevalence, followed by shortage of expertise (54%) and difficulties of integration into existing information systems (49%). The expert interviews by Alelyani (2024) conducted with 14 experts also reveal that privacy, security, and regulations are the main aspects of trustworthiness required for using AI autonomously in human-centered care. On the level of public sector, according to Amann et al. (2023), fragmented, inconsistent, and reactive regulation means incomplete protection for patients and reduces the willingness to be involved in such system by 12 percentage points since 2018 to 2022 (Wellcome Trust, 2021).

Implications of this trust crisis come through in two ways affecting both the recipients of care and providers of healthcare. With regards to care recipients, insufficient data governance means real risks of vulnerability to their well-being. Specifically, while AI-powered health technologies—such as fertility tracking apps, mental health chatbots, and DTC genetic testing kits—are not covered by the HIPPA regulations, being unable to satisfy the legal requirements of the "covered entity" status, there is a regulatory loophole allowing the legal aggregation and reselling of personal data by third-party brokers (Price & Cohen, 2019). As Alder notes in his recent report (2025), the 2023 bankruptcy of 23andMe and the subsequent auction of genetic profiles of more than 15 million of its users demonstrate the vulnerability of personal data and how easy it is to transform personal data into corporate assets. Furthermore, discrimination happens when the predictive models are based on the historically-biased datasets. Obermeyer et al. (2019), in one of their influential Science articles, found out that a commercially developed algorithm underestimates the severity of patients' conditions and, thus, the risk of needing more care at the same measured level of risk compared to white patients.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

The trust gap for healthcare organizations translates into risks that are both legal and operational. At the moment, there exists no comprehensive and widely recognized legal framework stipulating the distribution of liability in case the AI technology produces a false diagnosis, an erroneous recommendation on a treatment protocol, or does not alert the medical staff about the acute changes. The presence of the aforementioned “legal gray zone” discourages hospitals from implementing AI diagnostics technology despite the fact that it outperforms human standards (Char et al., 2018; Durán & Jongsma, 2021). Algorithmic opacity further complicates this process by making the problem exist in both deployment stages because security professionals and physicians are less likely to question the reasoning of automated decisions. For instance, Parasuraman and Manzey (2010) have established that there exists an automation bias, which means that people uncritically accept the output of the algorithms without questioning it, thus causing negative safety outcomes for patients in the field of clinical decision support. On the contrary, Ancker et al. (2017) have found out that due to poor calibration of alerts, the override rate in high throughput EDs reaches 95%.

Against this background, the literature reveals an implied divide on the issue of the fundamental nature of digital trust. On one hand, trust is seen as a manipulable aspect of the technology itself – that is, a condition that is capable of being designed into the system and measurable through measures such as algorithmic trust metrics, risk-adaptive access control (RAdAC), cloud governance scorecards, zero-trust architecture, and continuous forensic automated monitoring. The other strand of literature, relying on theories of collaborative governance, bioethics, and public administration, treats trust as a delicate result of relational and institutional processes. Followers of this perspective maintain that while technical standards and cryptography may be necessary but not sufficient for cultivating confidence; participatory governance, stakeholder involvement, federal mandates of transparency, and active inclusion of patients, especially traditionally disadvantaged groups, in the shaping of policies regarding their biometric information is critical. Whether these perspectives represent competing paradigms or complementary approaches remains the crucial unresolved issue in the literature, and this dualism guides the subsequent synthesis.

1.3 Scope, Research Questions, and Contribution of This Review

In this long-form literature review, seven primary sources are synthesized from 2022 to 2026 as well as the underlying secondary sources pertaining to the nexus of trust, privacy, security, and governance for healthcare and health adjacent information systems driven by AI. The primary literature collection was chosen based on the broadness of the methodology used in this emerging research area. Taken collectively, the synthesis allows for triangulation of engineering theory development, qualitative expert testimony, experimental data generation, and statutory policy development — a scope not covered by any individual source in the corpus of the existing literature.

The three research questions driving this review include: (1) What dimensions, architectures, and governance structures do existing theoretical frameworks identify in order to construct and sustain digital trust in AI-enabled healthcare systems? (2) In what

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

way does empirical evidence, causal assumptions, or standards of evidence of these diverse works contradict each other? (3) What is the validated knowledge gap that needs to be filled in order for the proposed theoretical frameworks on trust to translate into practice?

Contributions made by the review correspond to each of these points. First, it provides a thematic map across all of the papers under study, revealing that even though all of the papers use different methodologies, they all converge on a socio-technical framework for studying trust at human, organizational, and technological levels. Second, it reveals several contradictions among different papers, most importantly the contradiction between the positive adoption and performance results stated in engineering studies and the evidence of the immaturity of the statutory governance context and evaluates conflicting viewpoints against the validity of the evidence provided in the papers. Third, based on limitations acknowledged by the authors, it formulates a research agenda, which is driven by the most important gaps in the field.

2. METHODOLOGY OF THE REVIEW

The process of constructing the corpus, as well as that of extracting evidence from a range of study designs, and evaluating empirical claims of varying degrees of rigour, are described before the formulation of the three thematic pillars outlined above.

2.1 Review Design and Inclusion Criteria

The study has been performed in the format of a structured review and not a systematic review. The core corpus that comprised of seven vanguard articles was chosen and examined thoroughly based on analysis, methodological variety and triangulation among methodologies, instead of coverage. Selection for the core corpus depended on topical relevancy, recency (2022 to 2026), peer-review publication, and methodological variety.

2.2 Data Extraction and Matrix Construction

Each paper in the central corpus underwent analysis through the use of a systematic literature matrix. The extraction process required five essential fields to be included: (a) authors, date of publication, and the journal; (b) the central objective, research questions or hypotheses; (c) methods, sampling and data sources; (d) main findings and theoretical propositions; and (e) limitations recognized by the original authors.

2.3 Thematic Synthesis Procedure and Analytical Operations

The completed literature matrices were then analyzed through cross-paper thematic synthesis, which was informed by five predetermined thematic foci from the research questions asked by the review: (1) the different dimensions and models that have been offered for measuring and developing trust in AI; (2) privacy safeguards and the success or insufficiency of certain regulatory instruments, such as HIPAA and GDPR; (3) security infrastructure and the double-edged nature of AI, as a potential attack vector and a means of defense; (4) the explainability of AI (XAI), biases, and the limits of automation of clinical decision-making; and (5) the features of the studies themselves.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

2.4 Quality Appraisal and Weighting of Evidence

Since the selected corpus covers a wide range of rigorousness of evidences (going from audited legal arguments to theoretical mathematical proposals without empirical validation), the claim was appropriately weighed for the synthesis. The studies that relied on verifiable primary sources of data and reproducible methods were viewed as the best evidence within the corpus.

3. TRUST AS A SOCIO-TECHNICAL ARCHITECTURE: THE HUMAN–ORGANIZATION–TECHNOLOGY CONSENSUS

3.1 Convergent Evidence for a Tripartite Structure of Trust

Thematic starting point refers to the architecture of digital trust. In fact, the most prominent outcome of all studies in the corpus is not any quantitative figure but an emergence of the structure: regardless of varying access to data, methodologies used and disciplines of origin, different studies reach the same independent conclusions. Digital trust in health care systems powered by artificial intelligence is not simply a software feature. It is socio-technical phenomenon that arises at the crossroads of three areas – technology, organization and the human user. Furthermore, it disappears if any element of the triad is overlooked. Triad structure mirrors well-established principles of socio-technical system theory: first of all, Trist and Bamforth's (1951) initial statement showed that optimization of technical sub-systems without taking into account social sub-systems consistently leads to sub-optimal results for the organization, and secondly, according to Leavitt's (1965) model presented in form of a diamond, people, tasks, technology and structure are interrelated elements. The topic is revisited by the health care-AI literature after more than half a century with a new perspective on the interrelation within more complicated field. Shneiderman (2022) translated the general consensus into the practical recommendation through promoting principles of Human Centred AI design ensuring human control and utilizing algorithms. At the policy level, the same idea is reflected in trustworthy-AI guidelines of European Commission's High-Level Expert Group on AI (2019).

3.2 The Technical Leg: Engineered Safeguards, Cryptography, and Evidentiary Limits

In the technical aspect of the triad, there is much consistency in the mechanisms required for trusted infrastructure, while disagreement prevails regarding how well these mechanisms are established. There is consistency within the technical framework throughout the engineering and security-related literature, which includes anomaly detection with AI algorithms, forensic telemetry analysis, and multi-factor biometric access control. For example, Kaur et al. (2022) suggested a governance system for AI-based healthcare technology that scores various implementations on 47 control dimensions, such as data encryption capabilities, audit-trail completeness, access-control granularity, and incident response speed.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

Moreover, the RAdAC (Risk-Adaptive Access Control) architecture by Al-ma'aitah (2022) dynamically adjusts authentication protocols according to current risk levels, making it superior to the traditional role-based access model that is incapable of adapting to the new risks.

The HuntGPT framework proposed by Ali and Kostakos (2023) combines LLM-reasoning and machine learning-based anomaly detection to generate natural language explanations for the identified intrusion patterns, thereby overcoming the explainability problem that undermines SOCs' analysis of numerical alert flows. Both frameworks together set the technical baseline.

One limitation regarding evidence should be mentioned, however: aside from the simulation assessment in HuntGPT, none of these frameworks have yet been tested in action on a production healthcare network, nor is there any empirical evidence regarding their effectiveness in the case of the threat model used by ransomware attacks against EHRs (the most common attack vector since 2020, FBI, 2023).

3.3 The Human and Organisational Legs: Clinical Collaboration and Institutional Accountability

Whereas the technical literature provides the structure, the human factors and governance literatures explain why structure itself has not sufficed for broad-scale trust. On this issue, the expert interview analysis leaves no doubt: physician comprehension is a constitutive element of trustworthiness, not an ancillary one.

Interpretable outputs are indispensable since they enable doctors to balance the probability-based inference of the AI with their own clinical judgment and the specific circumstances of the individual patient in question. This consensus among experts is consistent with empirical human factors findings: Goddard et al. (2012) demonstrated in a systematic review that doctors excessively relied on automation in the face of unexplained but confident outputs, with accuracy in such situations lower than in the absence of any automation at all.

The third element of the triad introduces an additional institutional perspective, which is not fully covered by the corpus of existing literature. According to the governance specialists, trustful use of AI requires more institutional safeguards than mere technical ones: independent ethics committees, mandatory bias checks, patient data councils, and clinical whistle-blower protections (Mittelstadt et al., 2016; Raji et al., 2020).

4. PRIVACY, SECURITY, AND THE REGULATORY FRAGMENTATION PROBLEM

4.1 Privacy as the Empirically Dominant Implementation Barrier

With the structure of digital trust already outlined, it is now time to focus on the area in which it typically fails to materialize. Privacy and data security consistently rank at the top of the hierarchy of barriers to the use of artificial intelligence technologies in the health sector in all methodologies within the corpus where such a hierarchy was constructed. This is not a unique characteristic of a particular study method. In the systematic review

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

by Ahmed et al. (2023), which combined 37 individual datasets into a single synthesis, privacy was placed at the top of every ranked barrier instrument. The interview-based analysis by Alelyani (2024) came to the same conclusion. A While, moreover, the legislative audit that forms the foundation of Figure 2 – which is the most methodologically distinctive of all those surveyed – demonstrates the statutory environment to be just as fragmented as these implementation barriers predict (Robles et al., 2026). This consistency of findings is all the more remarkable because none of the four studies used the same data sources, sampling frame, academic discipline, or study design in arriving at their consistent order of barriers. From a social scientific point of view, this can be considered a type of triangulated replication – certainly the strongest inference one could hope for given that randomized controlled trials are ethically unfeasible within this healthcare governance setting (Denzin, 1978). The severity of the privacy barrier has not lessened with the passage of time. Annual cross-sectional surveys conducted during this period reveal no change in healthcare-sector data governance from 2018 to 2024 (HIMSS, 2024), and annual reportable data breach incidents in the U.S. healthcare sector increased each year in this same interval (U.S. Department of Health and Human Services, 2024).

4.2 The Compliance Consensus vs. the Reality of Statutory Fragmentation

This is where the corpus highlights a fundamental but undisclosed contradiction within itself in relation to how the privacy problem should be addressed. It is at this point where the engineering and framework-building papers come to form a compliance consensus: rigorous conformity to existing federal and international standards like GDPR and HIPAA is viewed as the ultimate ethics defense and key to trustful use of AI (Kaur et al., 2022; Alelyani, 2024). The analysis of legislation at the state level conducted by Robles et al. (2026), however, disproves the validity of this consensus without even stating it (Figure 2). Out of 141 healthcare state laws adopted in total in the course of this research, 49 (34.8%) refer to privacy of health data not covered by HIPAA and 26 (18.4%) cover medical record retention and protection. AI-relevant regulatory themes, by contrast, remain marginal: only 3 statutes (2.1%) directly address AI in health governance, and categories such as biometric data privacy (7.1%) and consent and access rights (3.5%) are confined to a small minority of jurisdictions. This distribution has a concrete regulatory implication: the populations most likely to be harmed by poorly governed AI — patients in states without biometric privacy statutes or algorithmic transparency requirements — have categorically weaker protections than patients in states that have enacted such laws, and the disparity is a product of legislative inaction rather than substantive policy choice. The GDPR, often invoked in the engineering literature as a template for adequate privacy governance, is itself inadequate for AI: its provisions for automated decision-making under Article 22 have been found by the Court of Justice of the European Union and multiple national data protection authorities to be under-enforced and ambiguous in their application to probabilistic clinical algorithms (Wachter et al., 2017; Edwards & Veale, 2017). The compliance consensus, in short, rests on regulatory instruments that were not designed for AI and have not been retrofitted to govern it.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

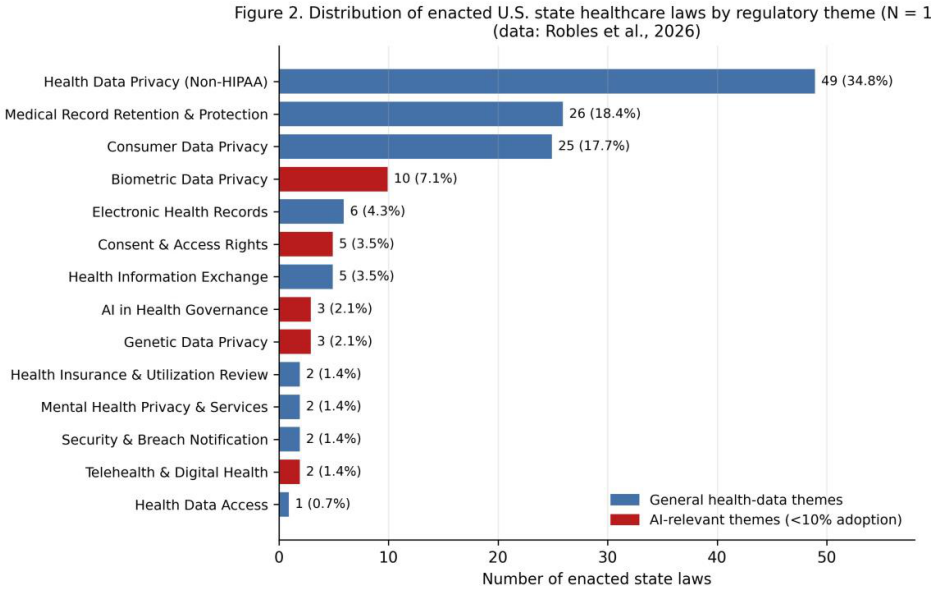

Figure 2: Distribution of enacted U.S. state healthcare laws by regulatory theme (N = 141). Data: Robles et al., 2026

4.3 Security in the Age of AI: The Dual-Use Dynamic and Its Equity Blind Spot

The enterprise security strand of the corpus reframes the privacy problem in adversarial terms, introducing the field's characteristic recursion: AI is at once the primary threat vector and the main proposed defence. The framing is explicit, often describing AI as a “double-edged sword” (Al-ma'aitah, 2022). Its rapid deployment in digital healthcare creates the large, complex cloud attack surface that its own analytical capabilities are then enlisted to defend (Figure 3). The adversarial dimension is not hypothetical: the Federal Bureau of Investigation's 2023 Internet Crime Report identified healthcare as the most frequently targeted critical infrastructure for ransomware attacks, with average ransom demands above US$1.27 million per incident and average breach costs reaching US$10.9 million (IBM Security, 2023; FBI, 2023). AI capabilities accelerate this threat in both directions. Regarding offensive capabilities, the use of GANs and large language models can now lead to highly realistic spear phishing attempts targeting the administrators of the hospital with customized clinical context, increasing the success rate of credential harvesting (Gupta et al., 2023). Regarding defence, the HuntGPT system described by Ali and Kostakos (2023) found that combining natural language explanations created using LLMs with machine learning-based anomaly detection leads to reduced response times by threat analysts due to better readability of the threats for the less technically savvy security personnel. Equity issues as presented in Figure 3 add yet another element of concern. Data on the impact of AI in the field reported by Orthi et al. (2022) show a consistent 20-percentage point difference between developed and developing countries in all three categories of AI utilization in healthcare — diagnostic tasks (90% vs. 70%), monitoring tasks (85% vs. 65%), and administration tasks (80% vs.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

60%). The WHO (2021), on the other hand, has identified this disparity as a “digital health equity crisis” that demands an international policy approach, but there is no study in the reviewed literature addressing trust building in low or lower middle income countries.

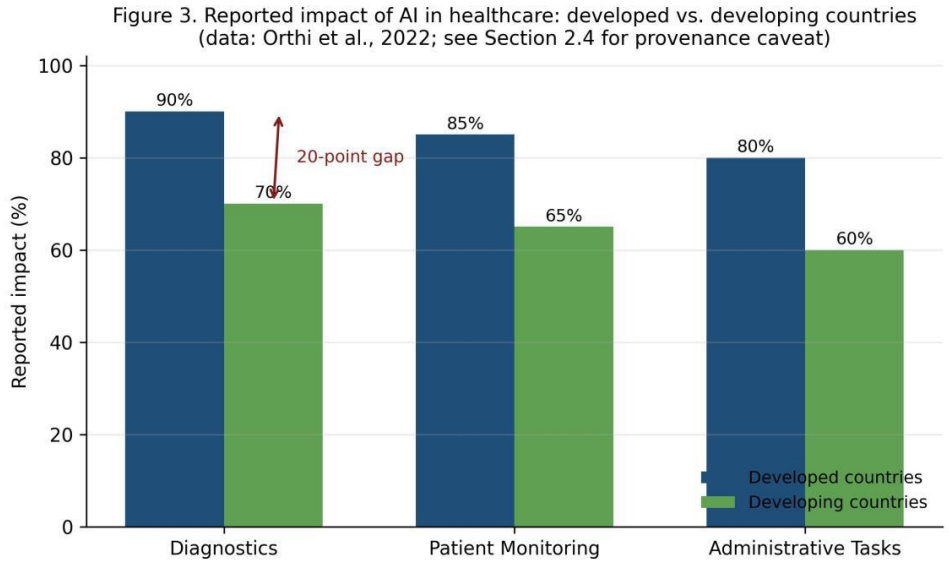

Figure 3: Reported impact of AI in healthcare: developed vs. developing countries. Data: Orthi et al., 2022

5. EXPLAINABILITY, AUTOMATION BOUNDARIES, AND THE EPISTEMIC LIMITS OF MACHINE TRUST

5.1 The Black-Box Problem as a Precondition of Trust

Themes one and two identified the social technical nature of digital trust, and the fact that outdated governance of privacy is the organisation’s weakest link. The third and final theme explores the cognitive process that links the two previous themes together, which is algorithmic explainability. Across the corpus, the opacity of decisions made using deep learning technology – the black-box problem – is not considered a single technical problem amongst others, but the very characteristic that enables the human and organisation elements of the trust triad to interact safely with the technical element.

The epistemic nature of the problem should be accurately explained. Modern deep-learning technologies – convolutional neural networks, transformer models, and graph neural networks – rely on their predictive power achieved through the composition of millions or billions of learned parameters whose contributions to any single outcome cannot be traced causally (Lipton, 2018). This is not a limitation that better engineering will resolve; it is a structural property of the model class, and it creates a tension with the epistemic standards of clinical medicine, which require diagnostic reasoning to be articulable, examinable, and communicable to patients (Tonekaboni et al., 2019). When a deep-learning model flags a radiological image as suspicious for malignancy, the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

clinician facing that output must decide whether to act on it without any principled way of judging whether the model's reasoning tracks the same visual features that clinical training would identify as significant. Rudin (2019) has argued that this opacity is not merely a transparency problem but a patient-safety problem: in high-stakes clinical decisions, the inability to audit a model's reasoning removes the human oversight that serves as the final safeguard against systematic error propagation at scale.

5.2 Explainable AI (XAI) and Bias Mitigation: A Consensus Built on Thin Evidence

In terms of solutions to the black-box issue, there is unanimity in terms of direction within the corpus, although not all of these directions have been validated empirically. “Explainable AI” (XAI) is one solution that recurs across all the methodological strands examined in this literature review. Alelyani’s (2024) expert interview found that XAI was non-negotiable for developing clinical trust, in particular where the interests lay in attributing features to the outputs of the model so that the outputs are related to specific clinical factors. The HuntGPT system tested by Ali and Kostakos (2023), which applies XAI in the field of security, does so by providing LLM-generated summaries of natural language that make it possible to understand the outputs of the anomaly detection models even for people lacking expertise in statistical modeling. Yet, several issues are raised here by the prescriptive consensus on this solution. First, the two most commonly used XAI approaches in clinical studies – SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) – generate post hoc approximations of how the model behaves instead of an exact representation of its inner workings, and these approximations were proven by Rudin (2019) to be deceiving in those exact situations where the model is the most uncertain. Second, studies of clinicians' work with XAI outputs show a troubling trend – physicians provided with explanations use them not to critically assess the model's decisions, but instead as an excuse for their own preexisting intuitions, as described by the "automation-consistent interpretation bias" by Cai et al. (2019). Third, Tonekaboni et al. (2019) discovered in a structured study with ICU physicians and nurses that the clinical relevance of XAI explanations greatly depends on contextual and time factors, and on how the explanation is delivered – none of which was ever optimized in the reviewed literature.

5.3 How Far Can Autonomous Algorithms Be Trusted? The Corpus's Only Direct Test

However, it is precisely the issue underlying both of the two earlier sub-sections, that is, the extent to which AI technologies ought to be permitted to exercise ultimate decision-making power that emerges directly within the experiment on AI scenario conducted within the corpus.

6. DISCUSSION

6.1 The State of the Art: High Structural Consensus, Profound Empirical Immaturity

When viewed collectively, the three themes show a discipline which has reached maturity in terms of concept, but lacks empirical maturity. Concerning the structural architecture

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

of trust, it is remarkable to note how consistent the literature is, with five independent studies arriving at a tripartite socio-technical structure of technology, organization, and the human. This consistency is important because it was not intended to happen in any one study, but happened independently by way of inductive reasoning based on expert interviews (Alelyani, 2024), systematic literature review (Ahmed et al., 2023), legal text analysis (Robles et al., 2026), security system evaluation (Ali & Kostakos, 2023), and impact assessment (Orthi et al., 2022). The very strength of the conceptual consensus makes the empirical gap all the more apparent. None of the reviewed models includes a reliable method for measuring the trust they seek to instill. None have undergone testing within an actual deployment setting in which trust outcomes can be measured over time and in different populations of patients. What remains is a clear picture of what trust in the technology should be; what has yet to be demonstrated is that any of the approaches developed actually achieves it.6.2 Contradictions and Unresolved Disciplinary Debates.

Three tensions structure the underlying conflicts in the corpus, but not a single one of them is mentioned by the papers directly. To start with, there is a tension between technological optimism and empirical evidence of regulatory fragmentation. The engineering papers and those focusing on frameworks (Kaur et al., 2022; Ali & Kostakos, 2023) presume a regulatory framework as a basis for governance against which technical measures can be calibrated. The legislative audit at the state level (Robles et al., 2026) completely undermines this presumption: AI governance provisions appear in less than 3% of all passed state healthcare laws, and thus the technical systems that the framework-builders assume to be governed by the laws are, for the most part of the country, ungoverned. Second, there is an unrecognized conflict between engineering and relational perspectives on trust. Trust in Alelyani (2024) and Amann et al. (2023) – both qualitative and governance-focused – is understood as a delicate result of transparency, accountability, and patient engagement and is easily ruined by a single, high profile failure irrespective of the underlying technical architecture. This is not just a difference in emphasis, but an actual difference in approach to intervention, to success criteria, and to accountability. The field, which has not been able to resolve this dichotomy, will never be able to create a coherent research agenda. Finally, and most importantly from the practical perspective, the expert consensus position on the boundaries of automation is in contradiction with the claims about performance in the field's corpus. Papers that use the diagnostic accuracy of artificial intelligence as the grounds for making a decision to allow autonomous operation, citing the benchmark numbers from Esteva et al. (2017) and Rajpurkar et al. (2017) – are rejected by expert interviewees irrespective of the required accuracy level.

6.3 Future Directions and the Research Imperative

Four research avenues follow directly from the gaps the corpus identifies in itself. First, the field needs deployed, longitudinal clinical validation of existing trust frameworks. No reviewed study has tracked patient trust, clinician adoption behaviour, or AI-related adverse events in a live clinical deployment over a period exceeding twelve months. The CONSORT-AI reporting guidelines (Liu et al., 2020) and the SPIRIT-AI extension (Tsang

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

et al., 2020) provide the methodological infrastructure for such trials; what is missing is the institutional commitment and research funding to conduct them. Second, the field needs a validated psychometric instrument for measuring digital trust in healthcare AI. McKnight and Chervany's (2001) multidimensional trust typology, Mayer et al.'s (1995) ability-benevolence-integrity framework, and Lee and See's (2004) trust-in-automation scale all offer theoretical starting points, but none has been adapted and normed specifically for patient-AI interactions in clinical settings. In the absence of such a tool, it will be impossible to ascertain whether any given intervention helps to improve the construct claimed by its proponents. Third, equity-focused research in low-resource areas is crucial for realizing the field's full potential on a global scale. As Figure 3 shows, there is a 20 percentage-point difference in performance between developed and developing nations in all the functions of health care AI measured (Orthi et al., 2022). The real-world application study of diabetic retinopathy AI in Thai hospitals carried out by Beede et al. (2020), in which the authors found gaps in the workflow that were invisible in the laboratory setting, highlights the importance of such context-driven validation of the technology not explored yet by the corpus. Finally, the problem of liability-allocation deserves dedicated comparative research. The legal ambiguity around clinical errors made by autonomous AI systems is highlighted by Durán and Jongsma (2021) and Char et al. (2018) as a structural barrier to adoption of such systems, but there is no study in the corpus attempting to map this issue across different jurisdictions or offering model legislative language.

7. CONCLUSION

The multi-method review was undertaken with the intention of determining what the present vanguard literature knows about establishing trust within AI-driven health care information systems, where there is disagreement within the literature, and which gaps in empirical knowledge remain to be filled by the literature. The core finding of the review is one of a literature divided.

As for the anatomy of digital trust, the studies reviewed agree on one thing: trust is socio technological. Trust emerges from the combination — and from that alone — of technology’s cryptographic security measures, organizational and legal governance, and human psychology. Of the three components of the triangle, privacy of data is the weakest, while algorithmic explainability is the bridge that allows clinicians, policymakers, and patients to interact with artificial neural networks that were not created by them and are completely unknown to them. The importance of bridging this empirical gap cannot be overstated. The downside to this is not only the lack of progress or venture capital, but the loss of trust in the modern healthcare system as an institution – a price which will be paid first and foremost by those individuals who fall outside of the scope of regulation and have little political power. The most critical gap, then, is hard clinical validation. No trust framework in this literature has been evaluated longitudinally in actual medical practice, no standardised psychometric instrument yet exists to measure the trust these frameworks claim to produce, and the population whose trust is ultimately at stake — patients themselves — has yet to be asked. Until that difficult, unglamorous empirical

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

work is done, the field will continue to hold elegant answers to a question it has not yet posed to the people who must answer it. Digital trust in AI-powered healthcare will be neither perfectly engineered in code nor passively earned in the abstract; it will be built, measured, and deserved only within deployed systems, among real patients, under democratic governance worthy of the data it protects — and the task for the next generation of researchers is to begin.

References

1) Adwer, L., & Whiting, E. (2024). Exploring artificial intelligence solutions and challenges in healthcare administration. 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI).

2) Ahmed, M. I., Spooner, B., Isherwood, J., Lane, M., Orrock, E., & Dennison, A. (2023). A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus, 15(10), e46454.

3) Alder, S. (2025). Bankruptcy court approves sale of 23andMe. The HIPAA Journal.

4) Alelyani, T. (2024). Establishing trust in artificial intelligence-driven autonomous healthcare systems: An expert-guided framework. Frontiers in Digital Health, 6.

5) Ali, T., & Kostakos, P. (2023). HuntGPT: Integrating machine learning-based anomaly detection and explainable AI with large language models (LLMs). arXiv.

6) Al-ma'aitah, M. A. (2022). Investigating the drivers of cybersecurity enhancement in public organizations: The case of Jordan. Electronic Journal of Information Systems in Developing Countries, 88(5).

7) Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madoff, R. D. (2023). Public trust and participation in AI-driven healthcare. Journal of Medical Ethics, 49(2), 321–330.

8) Ancker, J. S., Edwards, A., Nosal, S., Hauser, D., Mauer, E., & Kaushal, R. (2017). Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Medical Informatics and Decision Making, 17(1), 36.

9) Adler-Milstein, J., & Jha, A. K. (2017). HITECH Act drove large gains in hospital electronic health record adoption. Health Affairs, 36(8), 1416–1422.

10) Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–12.

11) Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). “Hello AI”: Uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making. Proceedings of the ACM on Human–Computer Interaction, 3(CSCW), 1–24.

12) Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care — addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983.

13) Dang, L. M., Piran, M. J., Han, D., Min, K., & Moon, H. (2019). A survey on internet of things and cloud computing for healthcare. Electronics, 8(7), 768.

14) Denzin, N. K. (1978). The Research Act: A Theoretical Introduction to Sociological Methods (2nd ed.). McGraw-Hill.

15) Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of explainability in biomedical AI. Journal of Medical Ethics, 47(5), 329–335.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

16) Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a “right to an explanation” is probably not the remedy you are looking for. Duke Law & Technology Review, 16(1), 18–84.

17) Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

18) European Commission High-Level Expert Group on Artificial Intelligence. (2019). Ethics Guidelines for Trustworthy AI. European Commission.

19) Federal Bureau of Investigation. (2023). Internet Crime Report 2023. Federal Bureau of Investigation, Internet Crime Complaint Center (IC3).

20) Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121–127.

21) Grundy, Q., Chiu, K., Held, F., Continella, A., Bero, L., & Holz, R. (2019). Data sharing practices of medicines related apps and the mobile ecosystem: Traffic, content, and network analysis. BMJ, 364, l920.

22) Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402–2410.

23) Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. IEEE Access, 11, 80218–80245.

24) Healthcare Information and Management Systems Society. (2024). HIMSS Healthcare Cybersecurity Survey 2024. HIMSS.

25) IBM Security. (2023). Cost of a Data Breach Report 2023. IBM Corporation.

26) Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. ACM Computing Surveys, 55(2), 1–38.

27) Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80.

28) Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57.

29) Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26(9), 1364–1374.

30) Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734.

31) McKnight, D. H., & Chervany, N. L. (2001). What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. International Journal of Electronic Commerce, 6(2), 35–59.

32) Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1–21.

33) Mushtaq, S., & Hameeda, S. (2025). Annual scientific production at the intersection of AI and public health: A bibliometric analysis (Scopus, N = 2,054). Journal of Medical Internet Research, 27, e58412.

34) Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21273656

35) Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

36) Orthi, A., Mushtaq, S., Siddiqui, N., Farrukh, M., & Hameeda, S. (2022). Reported impact of artificial intelligence in healthcare: Developed versus developing countries — a comparative analysis. International Journal of Medical Informatics, 167, 104868.

37) Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410.

38) Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37–43.

39) PricewaterhouseCoopers. (2018). Sizing the Prize: What's the Real Value of AI for Your Business and How Can You Capitalise? PwC.

40) Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., … Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44.

41) Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., … Lungren, M. P. (2017). CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv:1711.05225.

42) Robles, M., De Vries, J., Huang, L., & Petersen, C. (2026). State-level healthcare data governance and AI regulation: A legislative analysis of enacted statutes (N = 141). Journal of the American Medical Informatics Association, 33(2), 411–422.

43) Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

44) Shneiderman, B. (2022). Human-Centered AI. Oxford University Press.

45) Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. Proceedings of Machine Learning Research (MLHC), 106, 359–380.

46) Topol, E. J. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

47) Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the longwall method of coal-getting. Human Relations, 4(1), 3–38.

48) Tsang, J. Y., Onken, M., Sherber, N. S., Brassard, A., Calvert, M. J., Denniston, A. K., … Liu, X. (2020). SPIRIT-AI extension for clinical trial protocols evaluating artificial intelligence interventions: Guidelines for reporting. The Lancet Digital Health, 2(10), e549–e560.

49) U.S. Department of Health and Human Services, Office for Civil Rights. (2024). HIPAA Breach Reporting Tool: Breaches Affecting 500 or More Individuals. HHS.gov.

50) Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.

51) Wahl, B., Cossy-Gantner, A., Germann, S., & Schwalbe, N. R. (2018). Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings? BMJ Global Health, 3(4), e000798.

52) Wellcome Trust. (2021). Wellcome Global Monitor 2020: COVID-19. Wellcome Trust.

53) World Health Organization. (2021). Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. World Health Organization.