Published: 13 February 2026
Online AM: 16 January 2026
Accepted: 08 January 2026
Revised: 17 November 2025
Received: 12 December 2024

SURVEY

Building Trust in Artificial Intelligence: A Systematic Review through the Lens of Trust Theory

MASSIMO REGONA, Queensland University of Technology, Brisbane, QLD, Australia

TAN YIGITCANLAR, Queensland University of Technology, Brisbane, QLD, Australia

CAROL HON, Queensland University of Technology, Brisbane, QLD, Australia

MELISSA TEO, Queensland University of Technology, Brisbane, QLD, Australia

Open Access Support provided by:

Queensland University of Technology

Building Trust in Artificial Intelligence: A Systematic Review through the Lens of Trust Theory

MASSIMO REGONA, QUT Urban AI Hub, Queensland University of Technology, Brisbane, Australia
TAN YIGITCANLAR, QUT Urban AI Hub, Queensland University of Technology, Brisbane, Australia
CAROL HON, City 4.0 Lab, Queensland University of Technology, Brisbane, Australia
MELISSA TEO, City 4.0 Lab, Queensland University of Technology, Brisbane, Australia

Artificial intelligence (AI) is reshaping industries by enhancing efficiency and accuracy, yet its adoption remains contingent on user trust, which is frequently undermined by concerns over privacy, algorithmic bias, and security vulnerabilities. Trust in AI depends on principles such as transparency, accountability, safety, privacy, robustness, and reliability, all of which are central to user confidence. However, existing studies often overlook the interdependencies among these factors and their collective influence on user engagement. Guided by Trust Theory and a systematic literature review employing the PRISMA protocol, this study examines the trust indicators most relevant to high-stakes applications. The review reveals that transparency and communication are consistently prioritised, while adaptability and affordability remain underexplored, highlighting gaps in current scholarship. Trust in AI evolves as users gain experience with these systems, with reliability, predictability, and ethical alignment emerging as critical determinants. Addressing persistent challenges such as bias, data protection, and fairness is essential for reinforcing trust and enabling broader adoption of AI across industries.

CCS Concepts: • **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence;** • **Security and privacy → Trust frameworks**;

Additional Key Words and Phrases: Artificial Intelligence, Responsible AI, Algorithmic Bias, Trust Theory, User Trust, Trustworthiness in Technology

ACM Reference Format:
Massimo Regona, Tan Yigitcanlar, Carol Hon, and Melissa Teo. 2026. Building Trust in Artificial Intelligence: A Systematic Review through the Lens of Trust Theory. *ACM Comput. Surv.* 58, 9, Article 220 (February 2026), 39 pages. https://doi.org/10.1145/3789256

This research was funded by the Australian Research Council Discovery Grant Scheme, grant number DP220101255.
Authors' Contact Information: Massimo Regona, QUT Urban AI Hub, Queensland University of Technology, Brisbane, Queensland, Australia; e-mail: massimo.regona@hdr.qut.edu.au; Tan Yigitcanlar (Corresponding author), QUT Urban AI Hub, Queensland University of Technology, Brisbane, Queensland, Australia; e-mail: tan.yigitcanlar@qut.edu.au; Carol Hon, City 4.0 Lab, Queensland University of Technology, Brisbane, Queensland, Australia; e-mail: carol.hon@qut.edu.au; Melissa Teo, City 4.0 Lab, Queensland University of Technology, Brisbane, Queensland, Australia; e-mail: melissa.teo@qut.edu.au.

ACM 0360-0300/2026/02-ART220
https://doi.org/10.1145/3789256

1 Introduction and Background

Trust theory provides a structured lens for understanding how trust is established and sustained, encompassing psychological, social and institutional dimensions [5, 6]. Both interpersonal trust between users and AI systems, and institutional trust in the organisations that design and deploy them, are necessary for long-term adoption [7, 8]. While the literature on trustworthy AI is expanding, much of it examines single elements such as transparency, explainability or ethics in isolation [9], with limited integration into a holistic framework grounded in trust theory. Important constructs such as accountability, robustness, inclusiveness and adaptability remain underrepresented [9, 10], particularly in high-stakes contexts such as healthcare, finance and construction, where trust failures may have significant consequences [11, 12].

Several reviews have advanced the understanding of trust in AI, yet their scope and emphasis differ from the present study. One review offered a conceptual overview of trustworthy AI without quantifying the prevalence of trust indicators [9]. Another translated trustworthy AI principles into practice, although with a focus on governance frameworks rather than integration with trust theory [13, 14]. A separate study concentrated on explainability and transparency but devoted limited attention to indicators such as inclusiveness or adaptability [15]. Empirical findings on human trust in AI have also been synthesised, though without developing a unified indicator set applicable across sectors [8]. In contrast, the present review systematically examines peer-reviewed literature from 2020 to mid-2024, identifying 18 trust indicators and quantifying their representation across regions, sectors, and application domains. The analysis integrates trust theory with AI ethics and governance to provide a comprehensive, up-to-date evidence base for researchers and practitioners.

Building trust in AI requires embedding transparency, fairness and accountability throughout the system lifecycle [16, 17], supported by continuous stakeholder engagement, rigorous validation and ongoing monitoring [18, 19]. By combining conceptual integration with quantitative mapping, this study offers a theoretical contribution through the alignment of trust theory and trustworthy AI principles, and a practical framework to guide the design, governance and deployment of AI systems that are more likely to achieve and sustain user trust [20, 21].

1.1 Conceptualising Artificial Intelligence

Understanding the social and psychological mechanisms that underpin trust in human–AI interaction is critical as AI's role expands across public and private domains [22]. AI can be defined as a class of technologies that interact with their environment by collecting information, recognising patterns, predicting outcomes, generating responses, and refining their decision-making processes to achieve specific objectives [23]. These systems are distinct from traditional rule-based technologies because they possess the capacity to adapt their behaviour through continuous learning. The development of AI is underpinned by several core components that enable diverse applications across industries:

- Machine learning: Facilitates systems that learn from data and refine performance without explicit programming, supporting applications such as fraud detection and predictive analytics [15].
- Neural networks: Model brain-like processes to recognise complex patterns, driving advancements in image recognition, language translation, and speech processing [24].
- Natural language processing (NLP): Enables machines to interpret and generate human language, forming the basis of chatbots, digital assistants, and translation systems [25].
- Computer vision: Provides the capability to process and analyse visual inputs, underpinning progress in medical imaging and autonomous vehicles [26, 27].

- Reinforcement learning: Involves learning through trial-and-error interactions with the environment, a method central to robotics, adaptive control, and gaming [28].

The environments in which AI operates are often highly complex and uncertain, making outcomes difficult to predict. Decision-making processes are typically non-deterministic and not always transparent, creating challenges in understanding how outcomes are reached [29]. Although advances in algorithmic learning have improved accuracy and broadened AI's functionality, they have also increased autonomy and introduced new risks. These uncertainties reinforce the need for trust as a central factor in the adoption and acceptance of AI technologies [30, 31].

1.2 Conceptualising Trust

Trust is a dynamic construct that emerges through interaction and reflects the willingness to accept vulnerability in the expectation of positive outcomes. It is generally defined as a psychological state involving confidence in another entity's behaviour, based on positive expectations of reliability, ethical conduct, and predictability, particularly in contexts characterised by uncertainty [32, 33]. In the case of AI, trust is essential for user acceptance and the sustained integration of intelligent systems into everyday decision-making processes.

Trust can be examined at three interrelated levels. Interpersonal trust refers to confidence placed in other individuals and is essential for effective collaboration, cooperation, and communication [34]. Organisational trust is rooted in the credibility and reputation of institutions and is shaped by governance structures, ethical standards, policies, and transparency of operations, all of which influence stakeholders' willingness to engage [15]. Technological trust relates to confidence in systems or artefacts such as AI and is determined by attributes including reliability, transparency, security, and the responsible use of data. This level of trust is increasingly significant as users interact with autonomous and data-driven technologies, where deficiencies in these attributes can undermine confidence [35].

Building on these levels, trust theory further distinguishes between cognitive and affective dimensions. Cognitive trust is grounded in rational evaluations, such as system performance, consistency, and predictability, whereas affective trust is shaped by emotional assurance and perceptions of benevolence [12]. Together, these dimensions provide a complementary framework for understanding how trust in AI systems is developed, maintained, and potentially eroded over time.

1.3 Trust in Technology

Existing definitions of trust in AI reflect varied disciplinary perspectives, but their scope also reveals conceptual ambiguity that requires clarification and critical interpretation. [14] conceptualise trust as the willingness to take risks by balancing benefits and risks of AI performance, while [36] emphasise trust emerging from consistent and reliable behaviour in critical scenarios. [37] provide one of the most widely cited definitions, framing trust as the willingness of a party to be vulnerable to the actions of another based on the expectation that the other will perform an important action, irrespective of monitoring or control. These perspectives are not contradictory but complementary, together enriching the conceptualisation of trust in AI. However, adopting them uncritically risks perpetuating conceptual ambiguity. To advance clarity, this study interprets trust in AI as both a willingness to accept vulnerability and a confidence in consistent, reliable, and ethically aligned system performance, particularly important given the complexity, autonomy, and unpredictability of modern AI systems.

Table 1 represents the 18 key trust indicators in AI systems, which are crucial in shaping users' perceptions and influencing their intention to engage with the technology. Indicators such as transparency, accountability, reliability, affordability, safety, and ethics directly impact how users assess

Table 1. Critical Indicators of Trust in AI Systems [20]

Dimension	Indicator	Description
Integrity	Accuracy	AI systems must produce precise and dependable outputs, with minimal errors, to support decision-making and reduce costly mistakes [38].
	Experience	Practical exposure to AI in real-world tasks enhances user understanding and strengthens acceptance of the technology [39].
Acceptability	Equitability	AI systems must ensure fair treatment and equal access to benefits, eliminating biases and promoting inclusivity across user groups [18, 21]
	Ethics	Ethical standards must guide AI use, ensuring fairness, transparency, and non-discrimination in decision-making processes [35].
Accessibility	Inclusiveness	Systems should be designed to accommodate diverse users, providing benefits across different roles, abilities, and expertise levels [3].
	Affordability	Adoption and long-term operation of AI should be cost-effective, delivering measurable value and sustainable returns on investment [36].
	Adaptability	AI should remain effective across varying conditions and contexts, demonstrating resilience and flexibility [40].
Governance	Accountability	Clear assignment of responsibilities, with enforceable mechanisms for oversight, ensures that outcomes can be traced and obligations fulfilled [41].
	Regulatory	Compliance with legal, professional, and industry standards promotes responsible AI operation and safeguards users [38].
	Training	Providing structured training and resources increases user competence and confidence in engaging with AI systems [20].
Information Exchange	Communication	Intuitive, user-friendly interfaces and clear communication pathways enable efficient interactions between users and AI [42].
	Transparency	Processes and decision-making within AI should be explainable and accessible, enabling stakeholders to understand system logic [19]
	Security	Robust protection against cyber threats and unauthorised access is critical for maintaining system integrity and user trust [43].
	Safety	AI systems must safeguard human well-being by incorporating features that minimise physical and operational risks [3].
	Privacy	Strong data protection and confidentiality safeguards are necessary to ensure user trust in the secure handling of sensitive information [44].
Alignment	Robustness	Systems must maintain reliable performance across diverse scenarios, demonstrating resilience under stress or change [45].
	Knowledge	Users require an applied understanding of AI's capabilities and limitations to use the technology effectively [39].
	Reliability	Consistent and dependable system behaviour builds long-term confidence in AI applications [46].

AI's trustworthiness [20]. This table provides the foundation for the study, demonstrating how these trust elements interconnect to foster trust and acceptance, particularly in high-stakes applications. It highlights how trust can be built and maintained throughout the AI lifecycle, enhancing user perception, reducing risk, and encouraging adoption. This holistic approach ensures ethical principles, governance, and safeguards are embedded at every stage, supporting the study's goal of fostering greater user confidence in AI systems [35].

The indicators outlined in Table 1 provide a structured foundation for conceptualising and assessing trust in AI systems [42]. They demonstrate that trust extends beyond technical performance to encompass governance mechanisms, ethical safeguards and user-centred design features [41]. Importantly, these indicators are not isolated but operate in combination, shaping perceptions of reliability, safety and fairness across diverse contexts. Situating them within the broader

framework of technological trust highlights that user confidence is fostered through consistent, transparent and ethically accountable system performance [47].

1.4 Trust Theory

Trust theory provides a valuable framework for understanding trust in AI and automation as both a technical and social construct. Trust in AI is shaped by the system's performance, its interaction with users, and the broader societal and ethical implications [37]. A central principle in trust theory is that the willingness to be vulnerable to an AI system must be justified by 'good reasons'. Without these reasons, or positive expectations, trust is reduced to hope or faith. For AI, trust is built on system-oriented assessments that focus on functionality, reliability, predictability, and helpfulness [3]. Positive expectations from users are based on the anticipated utility of the AI system, such as the reliability and ethical behaviour expected from its operation [44].

Understanding the trustor, the individual or entity placing trust in the AI, is a crucial element. This understanding shapes the associated risks and vulnerabilities of trusting an AI system [48]. Trust is developed through the interaction between users and the AI, which highlights the need for ongoing, transparent communication, accountability, and ethical oversight. These elements help ensure that AI systems remain trusted and reliable as it evolves and becomes integrated into more complex and diverse applications [36, 46]. The trust relationship is further underscored by the following components; (a) The trustor refers to the individual or entity that places trust in an AI system; (b) The referent of trust is the AI system or technology being trusted, and (c) The trusting relationship encompasses both the opportunities and challenges that arise from the interaction between the trustor and the AI system [49].

1.5 User-centric Trust in Artificial Intelligence

As AI systems become integral to decision-making across industries, the role of user-centric involvement in fostering trust is increasingly recognised [23]. A user-focused approach ensures that AI systems are designed with human needs, values and oversight at the forefront, enhancing their reliability, ethical grounding, and alignment with societal expectations. The theory of trust suggests that trust is built through interactions that meet both functional and relational needs, which is why embedding user feedback and involvement throughout the AI lifecycle is critical for developing technologies that are trusted and effective [36].

To build and maintain trust, users must be engaged from the design phase to system refinements. This involvement includes participation in algorithm development, performance evaluation, and error correction, which ensures that AI systems meet practical demands and remain accountable [50]. In this context, trust is not static but evolves as systems develop, and user involvement ensures that trust remains aligned with the system's performance and ethical standards over time. By prioritising a user-centric approach, organisations ensure that AI systems are not only technically proficient but also trusted by those who interact with them [47, 51].

1.6 User Involvement in Artificial Intelligence

Human involvement in AI development is essential for building trust, with the degree of involvement varying based on the application and associated risks [12]. In high-stakes environments such as healthcare or construction, where AI assists human decision-making, trust is maintained through close human oversight. Professionals like doctors or project managers interpret AI outputs and make final decisions, ensuring AI enhances rather than replaces human judgment [20]. According to the theory of trust, this form of human oversight fosters the trust by maintaining control over critical decisions, reinforcing the system's reliability and ethical alignment.

In moderate-risk applications, such as loan approvals or document analysis, users review AI recommendations and have the authority to override decisions when necessary. This level of involvement mitigates the risk of bias and ensures that AI systems operate fairly and transparently [42]. For lower-risk applications, such as customer service chatbots, users may take a more passive role but still monitor the system to ensure it meets expectations, reinforcing trust through continued alignment with user needs [52]. By adjusting the level of human involvement according to risk and complexity, organisations can ensure that trust in AI systems is maintained. This approach aligns with trust theory, which emphasises transparency, control, and accountability as key to building and sustaining trust [53].

Building on this foundation, it is important to recognise that trust is not developed in isolation at the user level, but rather through the combined efforts of multiple actors involved in the AI lifecycle. Stakeholders play distinct roles in embedding transparency, accountability, and reliability into systems, ensuring that trust is established and sustained across different contexts.

- Data scientists and developers play a pivotal role in ensuring that AI systems meet technical, ethical, and legal standards, which is essential for building trust [54]. Their work on refining algorithms, correcting errors, and improving system performance is crucial for maintaining reliability and trustworthiness [13].
- End-users, such as professionals in fields like healthcare, business or law, rely on AI to support their *decision*-making processes. Their feedback is essential for identifying areas of improvement, ensuring AI systems align with practical needs and retain trust [12].
- Domain experts, who provide industry-specific insights, help tailor AI systems to the specific demands of fields like healthcare or finance, reinforcing trust by ensuring the systems are relevant and effective [53].
- Policymakers, regulatory and industry bodies are responsible for overseeing AI's adherence to ethical *standards* and legal requirements. Their role in monitoring compliance is essential for maintaining trust, as it ensures transparency and accountability [37].

These user roles play an active part across various stages of the AI lifecycle, contributing to the development, deployment, and ongoing maintenance of AI systems. By engaging at each phase, it helps ensure that AI technologies are not only created but also effectively implemented and continuously monitored to meet evolving standards and needs [54].

2 Research Design

This study aims at providing a comprehensive review of the trust indicators that shape users perceptions of AI systems. By examining various factors that influence stakeholders trust in AI, the study seeks to identify the key elements that either contribute to or detract from trust in AI. This analysis will not only highlight the critical areas for building and maintaining trust but also inform the development of AI systems that are more likely to be accepted and trusted by a broad range of users.

To achieve this objective, a systematic literature review was conducted. The **Preferred Reporting Items for Systematic Reviews and Meta-Analyses** (**PRISMA**) protocol was employed to ensure the replicability and rigour of the study. Relevant literature on AI was sourced from academic peer-reviewed journals, with search parameters derived from established trust indicators, such as accuracy, transparency, reliability, and accountability [12]. These indicators are primarily grounded in trust theory, which provides a foundational framework for evaluating trust in AI systems. However, the scope of the search also extended beyond traditional trust theory, incorporating broader considerations such as ethical implications, inclusivity, and adaptability. This approach ensured a comprehensive exploration of the factors influencing trust in AI, not only

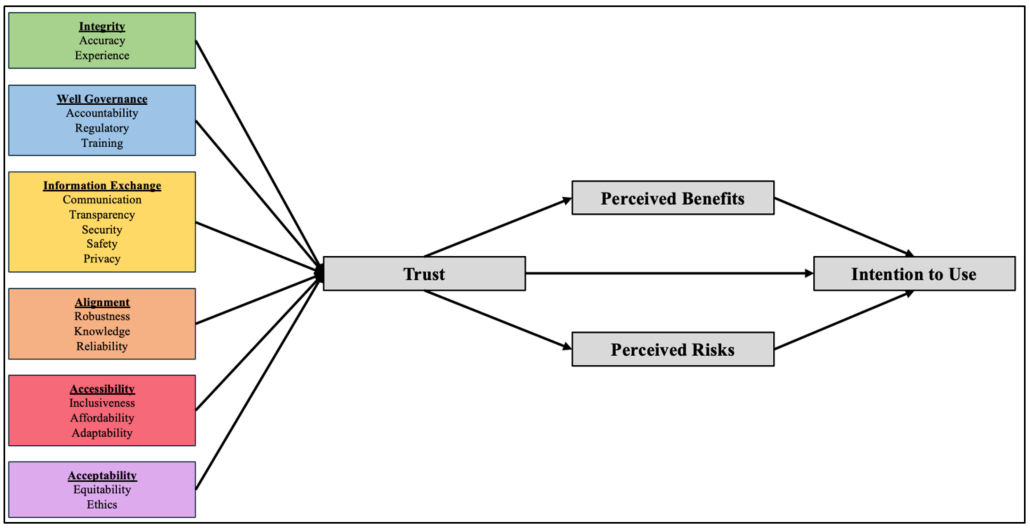

Fig. 1. Research model of trust building capabilities.

from a theoretical standpoint but also in practical, real-world applications [53]. Furthermore, the search process adopted a cross-disciplinary approach, integrating insights from fields such as psychology, sociology, computer science, built environment and engineering. This multidisciplinary perspective enriched our understanding of the complex factors influencing trust in AI. The analysis specifically focused on identifying, developing, and applying trust indicators, examining how these technologies impact user trust.

By analysing a wide range of trust indicators, the study will contribute much needed and holistic view of the AI trust landscape, identifying key trends and gaps in the existing literature. This comprehensive review will offer valuable insights into the factors that shape trust in AI, thereby informing the development of strategies to enhance trust in AI systems. The findings from this systematic review will guide the design and implementation of AI systems in ways that meet the diverse needs and expectations of users [55]. This analysis aims at contributing significantly to the broader discourse on trustworthy AI, offering invaluable insights for researchers, practitioners, and policymakers [4, 31].

The study has implemented a similar three-stage methodological approach by following the PRISMA protocol, as follows: Stage 1 (planning stage) includes research objectives that answer the research question, keywords, and a set of exclusion and inclusion criteria. The aim is framed at identifying predominant trust indicators and identify the challenges that users face when trying to trust AI technologies. While there are research articles that analyse trust towards AI, there is no research that provides an overview of all 18 trust indicators that influence the perception of users. While many research articles analyse trust in AI, particularly its link to user adoption, none provide a comprehensive overview of the 18 trust indicators that shape user perceptions. Incorporating these indicators into a trust model would offer a more complete understanding of the factors driving trust in AI technologies, which in turn influence perceived benefits, risks, and the intention to use as seen in Figure 1.

The initial search criteria for this study were based on “Trust and AI,” “Perception towards AI,” “Transparency in AI,” “Uncertainty in AI,” and “AI System Reliability.” To enhance the scope, the search was expanded to include "technological trust" and "trust in technology, " drawing from pre-AI research to explore relevant concepts that may inform and broaden the understanding of trust

Table 2. Exclusion and Inclusion Criteria, Derived from [2]

Primary data		Secondary data	
Inclusionary	Exclusionary	Inclusionary	Exclusionary
Peer-reviewed journal articles Conference papers Government reports Full-text available online Published in English Primary Data - Exclusionary:	Books and chapters Industry reports Non-peer-reviewed sources	Trust in AI applications Factors influencing trust in AI Studies on trust challenges specific to AI Empirical and theoretical studies Policy articles and guidelines on AI trust	Non-AI-related trust studies Irrelevant research objectives General trust literature not specific to AI Research not available in English Outdated studies (more than 4 years old)

in AI systems. These concepts formed the basis for identifying relevant literature and defining the scope of the review. From the literature review, a co-occurrence analysis of keywords related to the perception of trust towards AI identified 18 key trust indicators. Additional keywords were analysed, but did not reveal any new technology areas, overlapping with previously identified terms. Among the 18 indicators, transparency, communication, accountability, safety, and privacy were most frequently mentioned in journal articles. These keywords delineated the research boundaries and provided a comprehensive overview of the AI technologies currently employed in various industries. This approach ensured that the research covered a wide range of trust-related factors, offering a detailed understanding of how these indicators influence user's perceptions of AI. By focusing on these critical concepts, the study aimed at highlighting the most significant areas for building and maintaining trust in AI systems, ultimately contributing to the development of more trustworthy AI technologies.

The keyword search conducted in July 2024 obtained 354 results that satisfied the established search criteria. After removing duplicates, 327 articles were retained encompassing research beyond the university library databases. The search engine covered over 400 bibliographic repositories including Scopus, ScienceDirect, Web of Science, Directory of Open Access Journals, and Wiley Online Library. This initial search did not impose specific time period restrictions. Furthermore, as detailed in Table 2, criteria were developed to effectively reduce the number and complexity of articles for review facilitating a more streamlined screening process.

In Stage 2 (conducting the review), relevant articles were searched in July 2024. The use of AI has seen significant momentum over the past four years, with yearly changes in its use and adoption. Therefore, the literature was limited to publications from the last four years. The initial search, covering articles published from January 2020 to July 2024, reduced the number of articles from 327 to 298. These 298 articles were then assessed against the category formulation, as outlined in Table 3, further narrowing the selection to 138 relevant articles. Subsequently, the titles, abstracts, and keywords of these 138 articles were screened according to the exclusion criteria, resulting in a final set of 57 relevant articles.

In Stage 3 (reporting), 57 articles were analysed using descriptive techniques such as explanation building and pattern matching. The objective of these screening processes was to examine the selected articles according to predefined categories to assess similarities and differences [31]. A four-step process was then employed to classify the reviewed literature into specific themes [55].

Firstly, significant challenges and critiques related to user trust in AI were identified in the reviewed literature. Secondly, the most important themes were categorised and reviewed in relation to the research aims. The third step involved cross-checking these categories with other review

Table 3. Category Formulation Criteria, Derived from [55]

Selection criteria
• Identify key authors relevant to user trust in AI using qualitative data
• Determine the barriers to building trust in AI technologies
• Identify the challenges and opportunities that AI technologies may present in terms of user trust
• Categorise similar trust-related opportunities and challenges
• Group AI technologies based on their impact on trust at different stages and form categories
• Check the consistency of trust-related categories against other literature
• Shortlist categories and analyse recent literature reviews on trust in AI
• Verify, classify, and finalise the trust-related categories
• Distribute and select relevant categories based on their pertinence to user trust in AI

studies to identify additional challenges. Finally, the themes were organised and finalised under 18 common themes. Figure 2 provides an overview of the process used for selecting and categorising the articles, and for the salient characteristics of the reviewed literature (see Appendix A for salient characteristics of the reviewed literature).

This comprehensive analysis provided a clear picture of which trust indicators are most commonly addressed in the literature, highlighting areas of focus and potential gaps in current research. The frequency of mentions for each indicator underscores their perceived importance and the emphasis placed on them by various users in the context of AI trustworthiness.

3 Analysis and Results

3.1 General Observations

Publications from 2020 to 2024 were chosen to ensure the research reflects the most recent developments in AI and user trust. Given the rapid advancements in AI and its increasing integration into key sectors, recent literature provides the most relevant insights into emerging ethical concerns, regulatory frameworks, and technical innovations.

The publication dates demonstrate a sustained and growing emphasis on addressing contemporary challenges and devising solutions for building trustworthy AI systems, in line with evolving societal, industrial, and policy expectations. This consistent focus is particularly apparent in the publication trends, with a significant number of studies emerging in recent years: 2023 (n = 12), 2022 (n = 13), 2021 (n = 16) and 2020 (n = 16). The 2024 publication count is understated due to delays in the peer-review process, with many studies still pending release. These figures reflect the heightened priority placed on trustworthy AI as technological advancements accelerate, prompting greater attention from both academia and industry to meet the increasing demand for robust, ethical, and reliable AI systems.

The articles also highlight key themes in building trust in AI systems across various industries. AI's influence is expanding across key sectors, with significant focus on healthcare (n = 16), finance (n = 12), transportation (n = 10), and agriculture (n = 10). While construction (n = 6) and education (n = 3**)** receive fewer mentions, the integration of AI in these industries is steadily growing. This reflects the increasing adoption of AI to address industry-specific challenges, enhance operational efficiency, and foster innovation, as technological advancements continue to accelerate.

Many leading authors were affiliated with Europe (n = 27), highlighting the region's strong focus on trustworthy AI research, driven by its regulatory frameworks and commitment to ethical AI development. Furthermore, there is also strong interest in North America (n = 16). However, there are only limited studies from Asia (n = 5), others (n = 5) and Oceania (n = 4). As shown in Figure 3, user trust in AI has experienced consistent growth over recent years. This steady increase

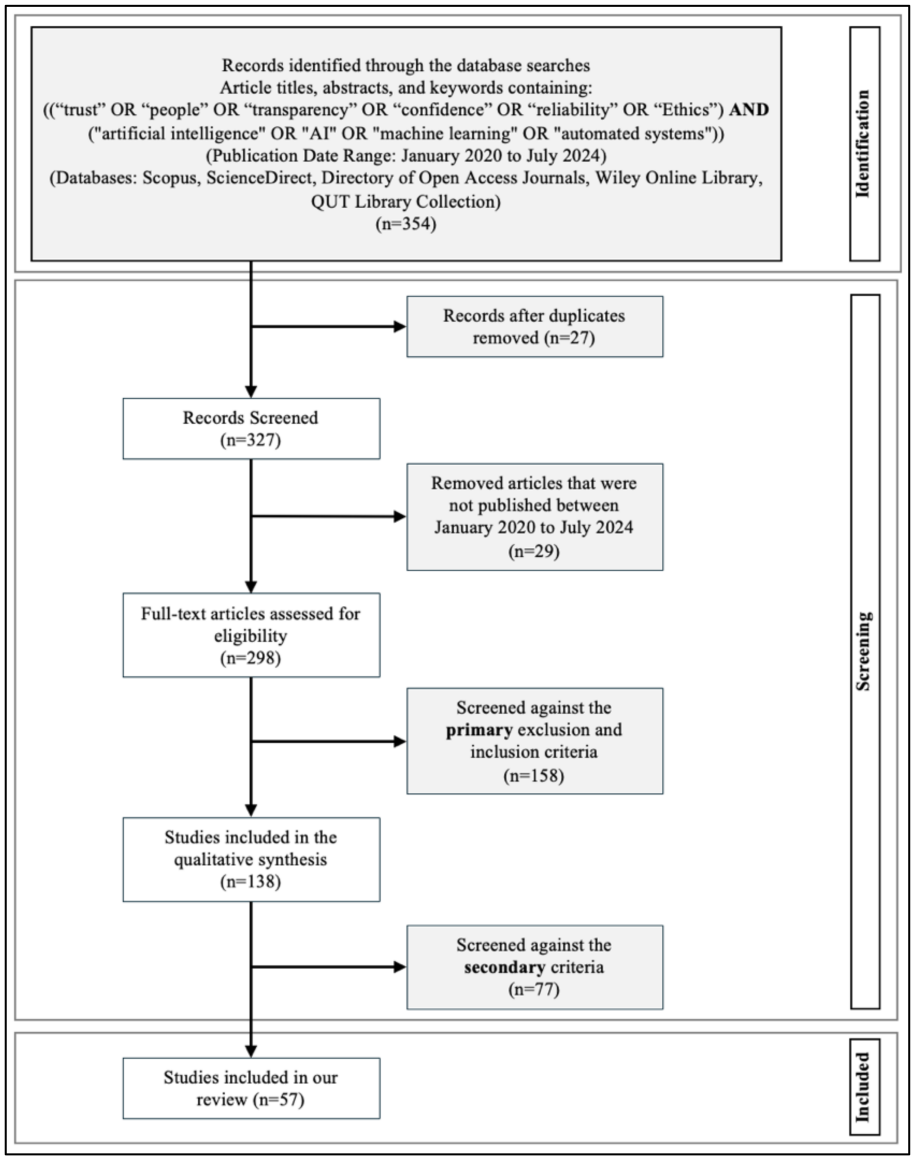

Fig. 2. The PRISMA selection process of relevant literature.

reflects the ongoing efforts in improving AI transparency, reliability, and ethical standards, which have gradually built user confidence in AI technologies.

Of the 57 articles reviewed, four authors contributed to multiple articles: Shyam Sundar (n = 3), Andreas Holzinger (n = 2), Vera Liao (n = 2), and Allan Dafoe (n = 2). No single author has dominated the literature on trust in AI, which is further reflected in the diversity of journals in which these articles were published. The majority of the articles focused on key trust-related factors, including transparency, accountability, reliability, and ethics. These works emphasised how AI systems can build user trust by ensuring fairness, security, and clear decision-making processes. In addition, it addresses significant challenges such as algorithmic bias and the importance

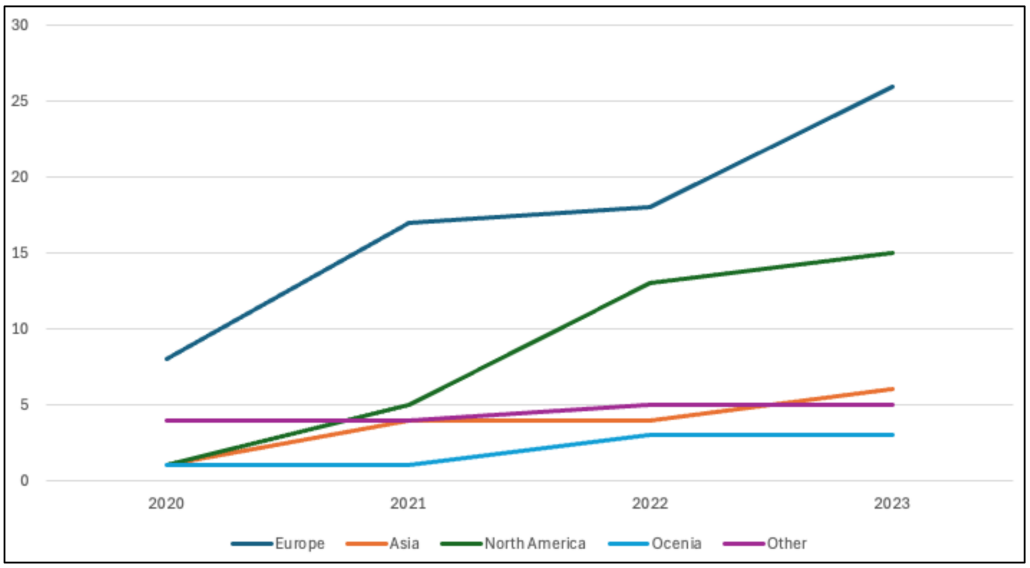

Fig. 3. Distributions of publication by region and year.

Table 4. Frequency of Trust Indicators Mentioned in the Analysed Articles

Trust indicators	**No. of mentions in** articles
Accountability	32
Accuracy	27
Adaptability	7
Affordability	15
Communication	48
Equitability	28
Ethics	27
Experience	18
Inclusiveness	13
Knowledge	22
Privacy	30
Regulatory	27
Reliability	29
Robustness	24
Safety	32
Security	29
Training	26
Transparency	51

of aligning AI systems with human values to foster greater confidence among users across various sectors. This broad approach indicates the multidisciplinary nature of trust in AI and highlights the diverse perspectives needed to address this evolving issue effectively.

The analysis of 57 articles highlights key trust indicators in AI systems. Table 4 summarises these indicators, showing the frequency with which each was mentioned in the reviewed articles.

Table 5. Process and Outcome Indicators of Trust in AI

Main category	Process indicators	Outcome indicators
Integrity	Accuracy, Experience	Trust is strengthened when systems deliver accurate results, remain usable, and uphold privacy protections [29].
Well Governance	Accountability, Regulatory, Training	Clear accountability, regulatory compliance, and appropriate training reinforce legitimacy. Weakness in these areas leads to ambiguity and lower institutional trust [56].
Information Exchange	Communication, Transparency, Security, Safety, Privacy	Effective communication, transparent data practices, strong privacy protections, and secure system operation help build user confidence. When communication is unclear, privacy is compromised, or safety and security measures fail, trust is quickly diminished [18, 57].
Alignment	Robustness, Knowledge, Reliability	Trust is supported when systems are technically robust, knowledge is shared appropriately, and outputs remain reliable [54, 58].
Accessibility	Inclusiveness, Affordability, Adaptability	Trust grows when AI is inclusive, affordable, and adaptable to different contexts. Exclusionary practices or high costs weaken acceptance [16, 21].
Acceptability	Equitability, Ethics	Long-term trust depends on equitable treatment and adherence to ethical standards. Perceived unfairness or unethical behaviour undermines acceptance [10, 43].

Transparency and communication were the most frequently discussed, while adaptability and affordability received less attention, indicating potential areas for further research.

The analyses revealed that transparency (10.5%) and communication (9.9%) emerge as the most commonly discussed elements, emphasising the crucial role of open communication and clear processes in fostering trust. This suggests that users place significant value on being able to understand how AI systems function and make decisions. Accountability and safety, both at 6.6%, also feature prominently. These findings highlight the importance of being able to hold AI systems accountable for their actions and ensuring that it operates safely, particularly in environments where human interaction is critical. The focus on privacy (6.2%) and security (6.0%) aligns with growing concerns over data protection and safeguarding sensitive information within AI systems, reflecting the increasing awareness of these issues as AI becomes more integrated into society.

Reliability (6.0%) and ethics (5.6%) highlights the importance of AI systems being dependable and adhering to ethical standards, ensuring effective functionality while aligning with broader societal values. Accuracy (5.6%) and regulatory compliance (5.6%) further emphasise the need for AI systems to deliver precise outcomes while adhering to legal and regulatory frameworks, ensuring trust through lawful operation and technical precision. Other factors such as experience (3.7%) and inclusiveness (2.7%) are mentioned less frequently, indicating that while important, these aspects receive comparatively less attention in the current body of literature. This may suggest a gap in addressing the user experience and ensuring AI systems are inclusive and accessible to a wider demographic. Interestingly, adaptability (1.4%) and affordability (3.1%) are among the least discussed factors. This could indicate a need for further exploration into how AI systems can be made more flexible to accommodate different use cases and more cost-effective to ensure widespread adoption across various sectors, including smaller organisations with limited resources.

3.2 Trust Indicators in Artificial Intelligence

Trust in AI can be conceptualised through a two-layered framework of process and outcome indicators (Table 5). Process indicators capture how AI systems are designed, governed, and managed, while outcome indicators reflect the observable effects of these processes on individuals, organisations, and society. This distinction provides a structured way of understanding trust as both

Table 6. Accountability Issues between Users and AI Technologies

Challenge	Reasoning
Blame Shifting	Organisations deflect responsibility for errors instead of accepting it [60].
Delayed Accountability	Slow response to AI errors or misconduct undermines trust [44].
Inadequate Redress	Users lack channels to challenge or appeal AI-driven outcomes [21].
Lack of Documentation	Incomplete records hinder oversight of AI decisions and actors [49].
Unclear Responsibility	Ambiguity over who is responsible for AI decisions reduces user confidence [14].

an embedded feature of system design and a manifested consequence of system operation, highlighting the importance of aligning governance, transparency, and accountability with user-facing measures of fairness, safety, and reliability [35, 54].

The categorisation of process and outcome indicators can be further reinforced by aligning it with internationally recognised AI governance frameworks. The OECD's AI Principles emphasise values such as transparency, accountability, fairness, and safety, while also distinguishing between enabling conditions, such as accessibility and governance, and outcomes, such as equitable treatment and reliable system performance. This distinction mirrors the separation adopted here, clarifying that fairness operates as an evaluative outcome while accessibility functions as a precondition for participation [16].

Similar emphases can be observed across other international initiatives and industry-led guidelines, which consistently highlight accountability, regulatory compliance, and transparency as foundational enablers that underpin the delivery of measurable outcomes in safety, reliability, and fairness [54, 56]. Positioning the framework in this way provides conceptual clarity, mitigates category overlap, and situates the indicators within a broader global landscape that has sought to balance ethical principles with practical mechanisms for trustworthy AI [46]. Collectively, this positioning provides a coherent foundation for examining how trust in AI can be conceptualised and assessed, ensuring that the indicators remain both theoretically robust and aligned with internationally recognised principles [10].

3.2.1 Accountability. Accountability in AI ensures that developers, operators, and stakeholders are held responsible for system outcomes, embedding ethical and societal obligations across the lifecycle [42]. It becomes actionable through governance frameworks, oversight structures, and feedback mechanisms that enable systems to be monitored and adjusted in response to risks or unintended consequences [59]. A key dimension is traceability, which links decisions to their sources and complements transparency by clarifying both processes and responsibilities [57].

Accountability also requires communication and governance mechanisms such as audits, regulation, and feedback channels, which strengthen trust by ensuring concerns are addressed [46]. It intersects with transparency and ethics, as openly sharing how systems are designed improves public understanding and mitigates misconceptions. However, distributed responsibilities across actors and jurisdictions create accountability gaps, underscoring the need for robust frameworks that integrate traceability, governance, and ethical practice, as outlined in Table 6.

3.2.2 Accuracy. Accuracy is central to trust in AI because it ensures outputs are reliable, consistent, and aligned with user expectations. When systems perform with precision and minimal error, users gain confidence in their capacity to operate effectively and autonomously, reducing the need for constant oversight [44]. Beyond technical performance, accuracy is closely tied to fairness and bias, as inaccurate data or flawed algorithms can generate skewed outcomes that undermine confidence. By contrast, highly accurate systems reinforce fairness and safety standards,

Table 7. Accuracy Issues Between Users and AI Technologies

Challenge	Reasoning
Bias in Data	Skewed datasets produce unfair and inaccurate results [44].
Context Ignorance	Failure to factor in context yields misleading predictions [40].
Data Quality	Poor-quality training data leads to unreliable outputs [49].
Edge-Case Failures	Poor handling of rare cases undermines confidence [45].
Model Drift	Lack of timely updates decreases accuracy over time [20].
Overfitting	Models perform on training data but fail in new contexts [27].
Weak Generalisation	Limited transfer to unseen scenarios reduces accuracy [62].

Table 8. Adaptability Issues Between Users and AI Technologies

Challenge	Reasoning
Resistance to Change	Reluctance to adjust workflows hinders adoption [19].
Inflexible Processes	Rigid business processes block integration [64].
Uneven Team Adoption	Inconsistent uptake creates operational disparities [43].
Unclear Value Proposition	Unseen benefits drive hesitation and mistrust [65].
Cultural Resistance	Anti-innovation cultures impede adaptation [40].
Weak Leadership Support	Limited sponsorship stalls change [43].
Slow Implementation	Delays frustrate users and reduce confidence [41].

positioning accuracy as both a technical benchmark and an ethical safeguard [54]. Accuracy also intersects with accountability and transparency, since reliable outcomes make it possible to trace decision-making and demonstrate compliance [20]. Continuous monitoring is essential, as even minor inaccuracies in dynamic environments can erode trust [61]. Accordingly, accuracy remains a core trust indicator within the broader framework summarised in Table 7.

3.2.3 Adaptability. User adaptability is essential for cultivating trust in AI, as it reflects the capacity of individuals to adjust their practices, workflows, and expectations when engaging with new technologies [43]. It encompasses openness to learning, readiness to modify routines, and confidence in adapting to evolving human–technology interactions. Trust is strengthened when users feel capable of adapting, since this reduces anxiety about disruption and enhances perceptions of control [63]. Adaptability also facilitates the integration of AI into processes in ways that maximise benefits such as efficiency, accuracy, and decision-making, aligning with indicators like perceived usefulness and ease of use [5]. Moreover, adaptability supports collaboration and feedback, as user adjustments can refine system performance. However, uneven resources and skills may create adaptability gaps if training and support are lacking, underscoring the need for sustained investment in user readiness, as reflected in Table 8.

3.2.4 Affordability. Affordability is a critical factor in building trust in AI, as it determines the extent to which systems are accessible to diverse users and organisations. High costs restrict adoption to well-resourced actors, while affordable solutions enable broader participation and reinforce confidence in AI's reliability and effectiveness [66]. Affordability also contributes to fairness and equity, since reasonably priced systems reduce technological exclusion and help close gaps between sectors, aligning with indicators such as inclusiveness and equitability [35]. Moreover, affordable AI signals a commitment to societal benefit rather than purely commercial gain, fostering goodwill and strengthening trust [60]. By broadening access, affordability supports sustainable

Table 9. Affordability Issues Between Users and AI Technologies

Challenge	Reasoning
High Upfront Costs	Expensive initial investment deters adoption [49]
Maintenance Burden	Ongoing update costs strain budgets [66].
Cost–Benefit Uncertainty	Unclear ROI weakens the business case [50].
Access Inequality	High costs limit access and widen gaps [14].
Opaque Subscriptions	Recurring fees without clear value erode trust [67].
Training/Integration Costs	Extra spend to train staff and integrate systems is a barrier [66].
Scaling Costs	Escalating costs at scale threaten viability [49].
Hidden Fees	Unexpected charges undermine provider trust [37].

Table 10. Communication Issues Between Users and AI Technologies

Challenge	Reasoning
Lack of Clarity	Vague explanations confuse users [57].
Over-Promising	Exaggerated claims lead to unmet expectations [3].
Inconsistent Messaging	Conflicting information creates confusion [21].
Inadequate Provision of Information	Insufficient detail on data use and decision processes [13].
Excessive Jargon	Technical language alienates non-experts [57].
Missing Feedback Channels	No routes to ask questions or give input [16].
Infrequent Updates	Users are not informed about changes or improvements [68].

growth, stimulates competition, and encourages innovation. The challenge lies in balancing cost with quality and functionality, making affordability a core dimension of trustworthy AI, as reflected in Table 9.

3.2.5 Communication. Communication is essential for establishing and sustaining trust in AI, as it ensures that complex technical information is conveyed clearly and accessibly. Acting as a bridge between developers, regulators, and non-technical users, communication prevents misunderstandings that could lead to confusion or mistrust [35]. Clear and ongoing communication demystifies AI by explaining system functions, decision-making processes, and the nature of underlying data and algorithms in formats that are widely understandable [38, 56]. It also manages expectations by clarifying capabilities and limitations, thereby reducing uncertainty and fostering resilience when systems encounter errors [39, 47]. Communication therefore overlaps strongly with transparency and accountability, as both depend on clarity in information exchange. Ensuring effectiveness across diverse audiences, however, remains a persistent challenge, underscoring its role as a key trust indicator within the framework summarised in Table 10.

3.2.6 Equitability. Equitability is critical to building and sustaining trust in AI, as it ensures that technologies are designed and implemented in ways that promote fairness and provide benefits to all users. By prioritising equitability, developers reduce the risk of bias and discriminatory outcomes, fostering confidence that AI serves diverse groups impartially [69]. Since AI systems are trained on datasets that may contain historical biases or reflect wider inequalities, embedding equitability as a design principle requires actively identifying and mitigating these risks to prevent unfair outcomes [45, 70]. Perceptions of fairness are central to user acceptance: when systems are equitable, users are more likely to trust outputs, integrate AI into their workflows, and support

Table 11. Equitability Issues Between Users and AI Technologies

Challenge	Reasoning
Bias in Algorithms	Model behaviour that perpetuates bias yields unfair outcomes [47].
Unequal Access	Resource gaps limit equitable use of AI [11].
Discriminatory Outcomes	Outputs favouring some groups erode fairness perceptions [70].
Non-Diverse Training Data	Homogeneous data produces biased models [34].
Neglect of Minority Groups	Design that overlooks minorities fails equity goals [3].
Disparate Impact of Errors	Mistakes disproportionately harm certain groups [71].

Table 12. Ethical Issues Between Users and AI Technologies

Challenge	Reasoning
Informed Consent Gaps	Data used without clear consent breaches norms [40]
Manipulative Design	Behaviour-shaping uses are viewed as unethical [69].
Inequity Risks	Systems that exacerbate social inequities breach ethical standards [18].

wider adoption [21]. Equitability also intersects with inclusiveness and accountability, but disparities in digital literacy and access to resources remain a challenge. Addressing these gaps through education, training, and policy support is essential, as reflected in Table 11.

3.2.7 Ethics. Ethics is a foundational element of trustworthy AI, ensuring that technologies are developed and deployed in alignment with moral principles and societal values. Ethical frameworks seek to minimise harm, protect rights, and promote user well-being, thereby reinforcing confidence in AI [69]. Without this foundation, systems risk perpetuating bias, deepening inequality, or causing unintended harm. When users perceive that ethical principles underpin design and deployment, they are more inclined to trust and engage with AI. Ethics intersects closely with other trust indicators, particularly fairness, accountability, and transparency, by providing mechanisms for addressing concerns and ensuring that users are both informed and empowered [54]. This demonstrates that ethics is not an isolated principle but one that integrates with the broader framework of trustworthy AI.

Ethical responsibility extends beyond developers to organisations and governments, which must oversee deployment in ways that serve the public interest. Long-term perspectives are essential, considering how systems may evolve and what consequences they may generate for future generations [10]. Collaborative approaches that involve diverse stakeholders further strengthen trust by embedding multiple perspectives in governance [12]. Organisations that embed ethics in their operations signal a commitment to responsible innovation, reassuring users that technological progress aligns with societal benefit. Maintaining consistent ethical standards, however, remains challenging where oversight gaps exist, highlighting the need for robust guidelines and enforcement mechanisms, as reflected in Table 12.

3.2.8 Experience. Experience is an important foundation for user trust in AI, as it reflects both the developer's practical expertise and the system's demonstrated ability to perform effectively in real-world contexts. AI designed with insights from prior deployments and user feedback is more likely to meet needs, address operational challenges, and deliver outcomes that extend beyond theoretical performance [72]. Iterative refinement through real-world interactions signals reliability and alignment with evolving requirements, reinforcing accountability and long-term trustworthiness [67]. Experience also supports ethical and regulatory compliance by reducing risks such as

Table 13. Experience Issues Between Users and AI Technologies

Challenge	Reasoning
Interface Complexity	Complicated UIs frustrate users [57].
Non-Intuitive Design	Confusing flows hinder effective use [60].
Poor Responsiveness	Slow or unresponsive systems reduce satisfaction [49].
Weak Personalisation	Failure to adapt to user needs feels impersonal [37].
Lack of User Control	Few options to adjust or override decisions [21].
Over-Automation	Excess automation alienates users [17].
Limited Error Recovery	Little support to correct AI mistakes [29].

Table 14. Inclusiveness Issues Between Users and AI Technologies

Challenge	Reasoning
Lack of Diverse Input	Excluding perspectives drives biased outcomes [53].
Exclusion of Key Users	Leaving out relevant users fosters alienation [74].
Cultural Insensitivity	Designs that ignore cultural context isolate groups [23].
Ignored User Feedback	Dismissing diverse feedback misses needs [35].
Top-Down Implementation	Imposed solutions generate resistance [53].
Inadequate Communication	Poor outreach to affected groups causes misunderstanding [21].
Biased Outputs	Skewed results exclude or disadvantage users [3].
Unaddressed Equity Concerns	Failure to act on equity issues appears unfair [63].

bias, security vulnerabilities, or compliance failures. In high-stakes environments, demonstrated performance enhances confidence in system robustness, linking experience closely with indicators such as reliability and safety [21]. Conversely, negative experiences can quickly erode trust, underscoring the need for user-centred design that prioritises intuitive interaction and satisfaction, as reflected in Table 13.

3.2.9 Inclusiveness. User inclusiveness is a critical factor in fostering trust in AI, as it ensures that development, deployment, and use actively incorporate diverse perspectives and experiences. Because AI affects a wide range of communities, design must reflect varied needs to avoid privileging some groups and reproducing existing inequalities through unrepresentative data [29]. Involving diverse users during development enhances transparency, fosters a sense of ownership, and demonstrates that the technology is responsive to input [61]. Inclusiveness also strengthens confidence by aligning AI with societal values such as fairness, justice, and accountability [73]. Moreover, when users feel involved in shaping AI, they are more likely to adopt and support its application, while ongoing collaboration enables systems to evolve with real-world needs [7, 48]. Ensuring inclusiveness across contexts, however, remains challenging, as reflected in Table 14.

3.2.10 Knowledge. Knowledge is a fundamental component of trust in AI, as it ensures that users understand both the capabilities and limitations of the technology. Informed users are better positioned to engage responsibly, interpret outcomes with confidence, and balance reliance on AI with human judgment [11]. Recognising where AI can complement rather than replace expertise strengthens trust in the system as a supportive tool, while knowledge also prevents underutilisation by enabling users to identify opportunities for innovation and efficiency. Conversely, overreliance without adequate understanding increases risk, highlighting the need for balanced perspectives [38]. Knowledge further overlaps with adaptability and accountability, since informed

Table 15. Knowledge Issues Between Users and AI Technologies

Challenge	Reasoning
Lack of Understanding	Limited grasp of how AI works breeds mistrust [38].
AI Misconceptions	Myths create unrealistic hopes or fears [75].
Unclear Benefits	Unseen value weakens support [47].
Concept Complexity	Concepts overwhelm non-experts [11].
Job Displacement Fears	Replacement anxieties fuel resistance [39].
Overestimation of Capabilities	Inflated expectations lead to disillusionment [13].
Regulatory Unfamiliarity	Poor knowledge of rules raises compliance concerns [32]

Table 16. Privacy Issues Between Users and AI Technologies

Challenge	Reasoning
Unauthorised Data Sharing	Third-party sharing without explicit consent [19].
Surveillance Concerns	Intrusive monitoring generates privacy fears [16].
Weak Anonymisation	Re-identification risk persists despite masking [68].
Long Data Retention	Keeping data without clear justification erodes trust [53].
Limited User Control	Users cannot manage or delete their data [19].
Opaque Consent	Complex consent processes undermine agency [77].

users are more likely to integrate AI effectively and to hold developers responsible for system outcomes. Ongoing education is therefore essential, though challenges remain in providing accessible learning at scale, as reflected in Table 15.

3.2.11 Privacy. Privacy is a critical factor in building trust in AI, as it ensures that sensitive information is protected and managed responsibly. Users must feel confident that their data is securely stored, safeguarded from unauthorised access, and handled in line with established standards, making privacy central to sustained trust [19]. Effective safeguards such as encryption, anonymisation, and strict access controls prevent misuse while aligning systems with regulations and industry standards [76]. Strong privacy practices not only reinforce confidence in security and ethical integrity but also reduce reputational risks and signal a commitment to responsible innovation [66]. Privacy further overlaps with transparency, accountability, and ethics, as clear communication about data use, consent, and deletion empowers users to exercise control and agency. Balancing privacy with transparency and accessibility remains challenging, particularly in sensitive domains, underscoring its role as a core trust indicator within the framework summarised in Table 16.

3.2.12 Regulatory. Regulatory alignment is a critical factor in establishing and sustaining trust in AI, as it ensures that technologies operate within recognised legal, ethical, and industry standards [33]. Compliance provides assurance that AI respects rights and responsibilities, functioning within defined boundaries that protect users and the wider public. Adherence to regulations signals a commitment to accountability, fairness, and transparency, while mechanisms such as auditability and openness allow decisions to be traced, reviewed, and explained, strengthening user confidence [20]. Regulatory alignment also supports safety and reliability by requiring testing, certification, and monitoring to reduce risks to individuals and communities [52]. Furthermore, compliance promotes long-term sustainability by ensuring adaptability to evolving legal and ethical expectations. While alignment reinforces ethics, transparency, and accountability, navigating rapidly changing

Table 17. Regulatory Issues Between Users and AI Technologies

Challenge	Reasoning
Lack of Clear Rules	Regulatory gaps create uncertainty [78].
Inconsistent Standards	Cross-region differences cause confusion [46].
Slow Regulatory Updates	Lagging rules leave oversight gaps [39].
Weak Enforcement	Limited enforcement drives non-compliance [79].
Bias/Fairness Regulation Gaps	Insufficient safeguards against biased outcomes [43].
Compliance Costs	Burdensome costs limit access for SMEs [49].

Table 18. Reliability Issues Between Users and AI Technologies

Challenge	Reasoning
Inconsistent Performance	Variable outcomes undermine trust [70].
Frequent Errors	High error rates reduce confidence [23].
Poor Scalability	Performance drops under scale or complexity [10].
Inadequate Testing	Weak validation causes real-world failures [66].
Insufficient Monitoring	Lack of ongoing oversight hides degradation [29].

regulatory environments without compromising innovation remains a significant challenge, as reflected in Table 17.

3.2.13 Reliability. Reliability is a key determinant of trust in AI, as it reflects consistent performance, accurate results, and the capacity to meet expectations over time. Dependable systems provide repeatable outcomes across diverse conditions, reducing errors and fostering confidence in decision-making and automation [80]. Reliability extends beyond technical precision to encompass fairness and ethical consistency, ensuring results remain objective across different user groups and contexts [79]. It also strengthens accountability, as predictable outcomes make it easier to evaluate performance, trace decisions, and verify compliance. Maintaining reliability, however, requires continual updates, recalibration, and quality assurance to prevent performance degradation [49]. In critical operations, reliability is especially important for continuity and minimal disruption, while failures or inconsistencies can erode trust rapidly. Accordingly, reliability remains a core trust indicator within the framework summarised in Table 18.

3.2.14 Robustness. Robustness is a critical component of trustworthy AI, as it ensures that systems can function reliably and effectively across diverse conditions and scenarios. AI technologies are deployed in dynamic environments where variability, unforeseen inputs, and complex challenges are common, making resilience essential for trust [54]. Robust systems maintain accuracy despite data fluctuations, changing user behaviour, or external conditions, giving users confidence that performance will remain stable without constant intervention [45]. They also handle anomalies by adapting responses, flagging irregularities, or reverting to safe operational modes, preventing errors that might erode trust [7]. Robustness overlaps with reliability, adaptability, and safety, while also supporting scalability by enabling AI to operate effectively across contexts without major modification [52]. Demonstrating robustness therefore reassures users of resilience and dependability, as outlined in Table 19.

3.2.15 Safety. Safety is a critical determinant of trust in AI, as it ensures that technologies operate reliably while protecting users from potential harm. Robust safety measures provide assurance that systems will function as intended, even in dynamic or unpredictable environments,

Table 19. Robustness Issues Between Users and AI Technologies

Challenge	Reasoning
System Failures	Crashes or failures disrupt operation [34].
Real-World Variability Sensitivity	Good in lab, weak in dynamic settings 52].
Poor Edge-Case Handling	Rare or unexpected inputs trigger errors [46].
Limited Scalability	Robustness degrades when scaling up [59].
Inadequate Error Handling	Weak recovery mechanisms prolong faults [73].
Insufficient Stress Testing	Lack of diverse testing leads to surprises [12].
Failure to Adapt	Inability to cope with evolving conditions reduces resilience [26].

Table 20. Safety Issues Between Users and AI Technologies

Challenge	Reasoning
Autonomous Decision-Making Risks	Actions without human oversight in critical contexts [79].
Algorithmic Errors	Faulty logic can cause harmful outcomes [49].
Cybersecurity Weaknesses	Exploits can create safety hazards [68].
Lack of Fail-Safes	Missing kill-switches or safe modes increase risk [66].
Insufficient Testing	Deployment without thorough safety testing [54].
Physical Safety Hazards	Malfunctioning robots/vehicles endanger users [41].

thereby reinforcing perceptions of resilience and reliability [20]. Demonstrating safe performance in high-stakes settings is central to sustaining confidence, as malfunctions can produce significant consequences when AI interacts with the physical world. Prioritising safety not only protects user well-being but also demonstrates alignment with ethical principles and compliance with regulatory standards [81]. In this way, safety acts as a bridge between technological innovation and social responsibility, ensuring that progress does not compromise protection.

As AI becomes more autonomous, the demand for advanced safety protocols intensifies. Systems must include safeguards that allow adaptation while avoiding unintended harm, supported by continuous monitoring, testing, and auditing to ensure improvements do not introduce new risks. Safety overlaps with reliability, robustness, and ethics, reinforcing expectations that AI will act predictably, transparently, and in accordance with human values [70]. Long-term safety further depends on iterative updates and independent validation to keep pace with evolving contexts. Balancing innovation with rigorous testing remains a persistent challenge, underscoring the importance of sustained oversight, as reflected in Table 20.

3.2.16 Security. Security is a fundamental requirement for trust in AI, as it ensures protection from cyberattacks, data breaches, and other malicious threats. Without robust safeguards, AI risks compromising both the data it processes, and the confidence users place in the technology [19]. Strong, multilayered frameworks are therefore essential to preserve the integrity of AI operations and the confidentiality of sensitive information. Effective security mechanisms reassure users that their data is protected from unauthorised access and misuse, a priority in domains where trust depends directly on safeguarding personal or proprietary information [82]. Security also underpins reliability, as defences against external threats help ensure AI continues to function as intended even under adverse conditions [60, 70].

Table 21. Security Issues Between Users and AI Technologies

Challenge	Reasoning
Data Breaches	Unauthorised access to sensitive data erodes trust [83].
Adversarial Attacks	Inputs crafted to mislead or subvert models [19].
Model Theft	Reverse-engineering or theft threatens intellectual property [65].
Weak Encryption	Insufficient encryption exposes data to misuse [45].
Malware/Ransomware	Compromise disrupts integrity and availability [3].
Insider Threats	Misuse of privileged access by internal actors [84].
System Integrity Risks	Undetected tampering corrupts outcomes [37].
Supply-Chain Vulnerabilities	Insecure third-party components introduce risk [19].

Table 22. Training Issues Between Users and AI Technologies

Challenge	Reasoning
Insufficient Training	Users struggle to interact with or assess AI [35].
Opaque Training Materials	Complex/unclear materials hinder learning [13].
Missing Ethical Content	Training that omits ethics risks harm [61].
Resource-Intensive Programs	High time/cost burdens reduce uptake [6].

Security further overlaps with privacy, since breaches undermine confidentiality, and with accountability, as developers are responsible for anticipating and mitigating vulnerabilities. Practical measures such as encryption, access controls, monitoring, and rapid incident response provide tangible assurance of protection [12]. However, maintaining effective AI security remains a continual challenge, as evolving threats, hidden vulnerabilities, and complex interdependencies can quickly erode confidence. Addressing these risks through comprehensive and adaptive strategies is therefore essential to sustaining trust, as reflected in Table 21.

3.2.17 Training. Training is a vital element in fostering trust in AI, as it equips both developers and users with the skills needed to manage complex technologies effectively. Without adequate preparation, risks of misuse, inefficiency, or unintended errors increase, undermining confidence in the system [36]. Comprehensive and continuous training ensures that AI is designed and applied to high standards of safety, ethics, and reliability [54]. Well-trained users are more likely to trust AI, as competence reduces misinterpretations and harmful outcomes while improving preparedness for unpredictable behaviours, particularly in high-stakes settings [85]. Training also supports knowledge acquisition, adaptability, and accountability by helping users adjust to evolving systems and understand the implications of their actions. Regular updates further align stakeholders with technological advances and best practices, though ensuring effective and accessible training at scale remains a challenge, as reflected in Table 22 [32].

3.2.18 Transparency. Transparency is a fundamental component of trust in AI, as it provides clarity about how technologies function, the data they rely on, and the processes by which decisions are made. This openness reduces uncertainty, enabling users to understand and verify operations, which fosters confidence in outcomes [68]. Crucially, transparency extends beyond the provision of technical details: information must be communicated in accessible and meaningful ways so that both technical and non-technical audiences can comprehend system processes [41]. When users understand the rationale behind AI-driven decisions, concerns about fairness,

Table 23. Transparency Issues Between Users and AI Technologies

Challenge	Reasoning
Algorithmic Opacity	Hidden logic blocks understanding of decisions [15].
Bias and Fairness Blind spots	Lack of transparency prevents bias assessment [75].
Opaque Decision Processes	Limited insight into how decisions are made reduces accountability [3].
Unstated Model Limits	Not communicating constraints inflates expectations [68].
Poor Lifecycle Disclosure	Weak reporting on monitoring and improvement [85].

accountability, and ethics can be addressed more effectively. In this sense, transparency overlaps with these trust indicators by allowing users to evaluate whether outcomes are impartial, reliable, and aligned with expectations [86].

Transparency also provides mechanisms for identifying and correcting errors, biases, or unethical practices, thereby reinforcing accountability by ensuring developers and operators can be held responsible [34]. In addition, transparent systems are more likely to comply with legal frameworks and ethical guidelines, strengthening both public confidence and regulatory trust. Openness can also offer a competitive advantage, as users favour technologies they can understand and monitor. However, achieving transparency remains challenging, as organisations must balance accessibility with intellectual property protection and the complexity of modern AI systems. These challenges highlight the ongoing need for clear and practical approaches to transparency, as reflected in Table 23.

4 Findings and Discussion

The increasing integration of AI across diverse sectors requires a comprehensive understanding of the key factors influencing user trust. Trust in AI is a complex, multidimensional construct that depends on the careful alignment of technical, ethical, and regulatory considerations [15]. It must be actively cultivated across the AI lifecycle, from design and implementation to ongoing management and oversight. Unlike traditional technologies, AI often operates in contexts involving automation, autonomous decision-making and advanced data processing, making trust establishment more challenging and context dependent [46].

4.1 Balancing Technical Performance and Ethical Responsibility

Trust in AI cannot be sustained through technical performance alone. Efficiency, accuracy, and reliability must be reinforced by ethical safeguards [82]. High-performing AI systems that lack fairness, transparency, or accountability risk eroding user confidence, particularly when errors occur or outcomes are perceived as unjust.

Fairness is a prominent concern. Users expect AI to operate without discrimination and to deliver equitable treatment across demographic, geographic, and socio-economic factors. Biased decision-making in recruitment or loan approvals, for instance, can undermine not only trust in a specific system but also broader societal confidence in AI technologies [87].

Transparency and accountability are equally vital. Users require clarity on how AI systems process data and generate outputs, including traceable decision pathways and identifiable responsibility for system behaviour [24]. Without clear accountability structures, organisations risk losing trust when errors arise. Security further underpins ethical AI, requiring assurance that personal and sensitive data are safeguarded from breaches and that systems remain resilient against malicious attacks [88].

Developers should therefore adopt a proactive approach, embedding ethical principles into AI design from the earliest stages. Fairness, transparency, and security must be treated as core design requirements rather than as afterthoughts or compliance obligations. These factors should be validated through continuous monitoring across the system's lifecycle [78].

4.2 Continuous Engagement and Trust Maintenance

Trust in AI is dynamic, evolving in response to user experiences and ongoing system performance. While technical and ethical design establish initial trust, maintaining it requires deliberate and continuous engagement between developers and users [89].

Users value AI systems that incorporate open communication, feedback collection, and transparent updates. These mechanisms reassure users that concerns are acknowledged and addressed. Without such engagement, even technically sound systems risk a gradual decline in trust, as users may interpret a lack of communication as disengagement or neglect [90].

Such engagement can take several practical forms:

- Regular updates on system changes, fixes, and improvements [52].
- User education to explain new features, algorithmic changes, or updated privacy safeguards [18].
- Structured feedback channels where users can raise concerns or suggest improvements [73].

This two-way interaction reinforces accountability, positioning trust as an active component of system governance rather than a static design feature. Importantly, engagement must be sustained over the long term, not concentrated only during deployment, to ensure that trust grows alongside the AI system's evolution [3].

4.3 Contextual Differences in Trust Requirements

Trust in AI is highly context dependent. Different sectors require distinct safeguards, communication strategies, and performance priorities. In high-stakes environments such as healthcare or autonomous vehicles, trust is closely tied to safety, reliability, and compliance with strict ethical standards [91]. In these settings, even a single malfunction can have serious consequences, making rigorous testing, redundancy systems, and clear accountability essential.

By contrast, in lower-stakes domains such as entertainment, retail recommendations, or customer service chatbots, users tend to prioritise usability, convenience, and personalisation [55]. Minor errors may be tolerated if systems are easy to use, adaptable to user preferences, and transparent about how recommendations are generated. For example, a diagnostic AI tool in healthcare must undergo clinical trials, provide detailed explainability, and incorporate fail-safe protocols. Conversely, a music recommendation algorithm can focus on tailoring content to preferences and offering users transparent controls [70].

This variation highlights the inadequacy of a one-size-fits-all approach. Developers must calibrate trust-building measures to the stakes and sensitivities of each application. High-risk environments demand stronger ethical and technical safeguards, while lower-risk applications can prioritise usability and adaptability [83].

4.4 Role of Regulatory Frameworks in Trust Development

Regulatory frameworks provide the structural foundation for reinforcing trust in AI. They establish baseline standards for safety, fairness, and privacy, while also creating mechanisms for accountability in cases of harm or misuse. Compliance serves not only as a legal obligation but also as a trust signal, demonstrating to users that a system has been subjected to independent scrutiny [89].

The rapid pace of AI innovation often outstrips the development and adaptation of regulatory frameworks, creating oversight gaps. This misalignment can weaken trust, particularly when users perceive that systems are being deployed more quickly than they can be reliably assessed for safety and fairness [33].

To address this challenge, structured dialogue between regulators, developers, and user communities is required. Regulatory mechanisms must evolve in parallel with technological advancements to remain relevant and effective. Proactive regulation, particularly in areas such as data privacy, algorithmic bias mitigation, and transparency requirements, can reduce risk and reassure users that their rights and interests are protected [67]. In this respect, regulation functions both as a safeguard and as a trust-building instrument, ensuring that AI systems are developed and deployed within accountable governance structures.

4.5 Implications for Researchers, Practitioners, and Policymakers

The findings have clear implications for multiple stakeholder groups. For researchers, the multidimensional nature of trust in AI requires continued investigation into the interplay of technical, ethical, and contextual factors [63]. Further empirical studies are needed to identify effective trust-building strategies across different sectors and to assess how user trust evolves over time in response to design modifications, regulatory interventions, and societal changes.

For practitioners, the results emphasise the importance of embedding trust-building measures at the earliest stages of design. Fairness, transparency, accountability, and security should be treated as primary design objectives rather than secondary considerations addressed after deployment. This requires interdisciplinary collaboration, user-centred design methodologies, and ongoing evaluation to ensure systems remain aligned with user expectation [23].

For policymakers, the findings highlight the need for regulatory approaches that can adapt to rapid technological change. Frameworks should balance innovation with protection, addressing risks such as bias, privacy breaches, and opacity, while enabling responsible experimentation. Policymakers also play a critical role in fostering collaboration between public, private, and academic sectors to ensure that trust in AI is sustained as systems become increasingly complex and pervasive [12].

4.6 Challenges

Trust has long been recognised as a central component of human relationships, and this importance extends into the interaction between humans and technology. In the context of AI, trust is essential for societal acceptance and sustained use, but it is uniquely difficult to secure. Unlike traditional technologies, AI often operates with attributes such as natural language processing, affective computing, and conversational capacities. These human-like qualities complicate the way people evaluate trustworthiness, because the boundary between human decision-making and machine decision-making becomes blurred. In this setting, values such as fairness, benevolence, and compassion have been increasingly recognised as necessary elements of trust, extending beyond technical proficiency to include broader social and ethical considerations [52, 67]

One of the most pressing challenges is the problem of interpretability. Many AI systems, particularly those built on deep learning, operate as opaque "black boxes." Their outputs may be accurate, but the pathways leading to those outputs are often hidden or highly complex, making them difficult for users to scrutinise. This lack of transparency undermines accountability and restricts meaningful oversight. Explainable AI has been developed in response to this issue, aiming at providing insights into system logic and outcomes [32, 85]. However, a tension persists: the most powerful and accurate models tend to be the least interpretable. Simplifying them for the sake of explanation can compromise performance, while leaving them opaque risks user

scepticism and mistrust [50]. Striking the right balance between interpretability and accuracy remains unresolved, particularly in high-stakes sectors where trust is paramount.

Beyond technical opacity, there are broader governance and ethical challenges. Issues of bias and fairness continue to undermine trust in AI, with biased training data or poorly designed algorithms producing discriminatory outcomes [79]. In domains such as recruitment, finance, healthcare, and construction, these outcomes can have significant material consequences, eroding both individual and societal confidence in AI systems. Governance structures often lag behind technological progress, leaving regulatory gaps that exacerbate mistrust. Even where frameworks exist, their implementation can be inconsistent, fragmented across jurisdictions, and slow to adapt to the pace of innovation [13]. This misalignment between technological development and governance capacity represents a substantial barrier to the cultivation of trust.

Another challenge relates to accessibility and inclusiveness. AI adoption is uneven, and systems are often designed without sufficient consideration of the needs of diverse user groups. This can exclude individuals with lower levels of digital literacy, disadvantaged communities, or industries such as construction, where end-users may not have regular access to advanced digital infrastructure [19]. When accessibility is compromised, AI risks reinforcing existing inequalities rather than reducing them. These structural challenges remind us that trust cannot be assumed simply by virtue of technical capability but must be actively constructed through systems that are transparent, fair, inclusive, and governed in ways that align with societal expectations [9].

4.7 Opportunities

While the challenges are substantial, there are equally important opportunities to address them. These opportunities rest on a twofold strategy: strengthening the technical foundations of AI systems and embedding ethical considerations throughout the design, deployment, and monitoring processes.

Improving the quality and robustness of data offers one of the clearest opportunities for building trust. Data bias, poor generalisability, and inadequate validation remain central risks to the reliability of AI systems [21]. Addressing these issues requires more rigorous data collection, the use of diverse datasets, and continuous testing across varied environments. Systems should also be designed to handle rare or extreme cases, enhancing adaptability and ensuring performance does not collapse under unexpected conditions [49, 53]. In industries such as construction, this could mean validating predictive tools across different project types and geographic contexts, thereby ensuring that outputs are not narrowly tailored to one environment but robust enough to perform reliably across many.

Strengthening accountability and compliance frameworks is another key opportunity. Clear governance structures that allocate responsibility for AI outcomes improve traceability and assure users that systems can be monitored and controlled. Organisations can reinforce this through auditing processes, documentation, and formal oversight mechanisms [77]. At a broader level, collaboration between industry and regulators is necessary to ensure that laws keep pace with technological change. Proactive regulation, if responsive and flexible, can both protect users and provide clarity for developers, reducing regulatory risk and fostering trust [38].

Opportunities also exist in communication and transparency. Designing systems with interfaces that provide clear explanations of how outputs are generated, their limitations, and their associated risks can empower users and reduce scepticism. Feedback loops between developers and users create further opportunities for trust-building, as they allow systems to adapt in line with user expectations rather than being imposed as static technologies [7]. For example, explainable AI models, if communicated effectively, can bridge the gap between technical complexity and human comprehension, enabling users to exercise oversight and challenge outputs when necessary [92].

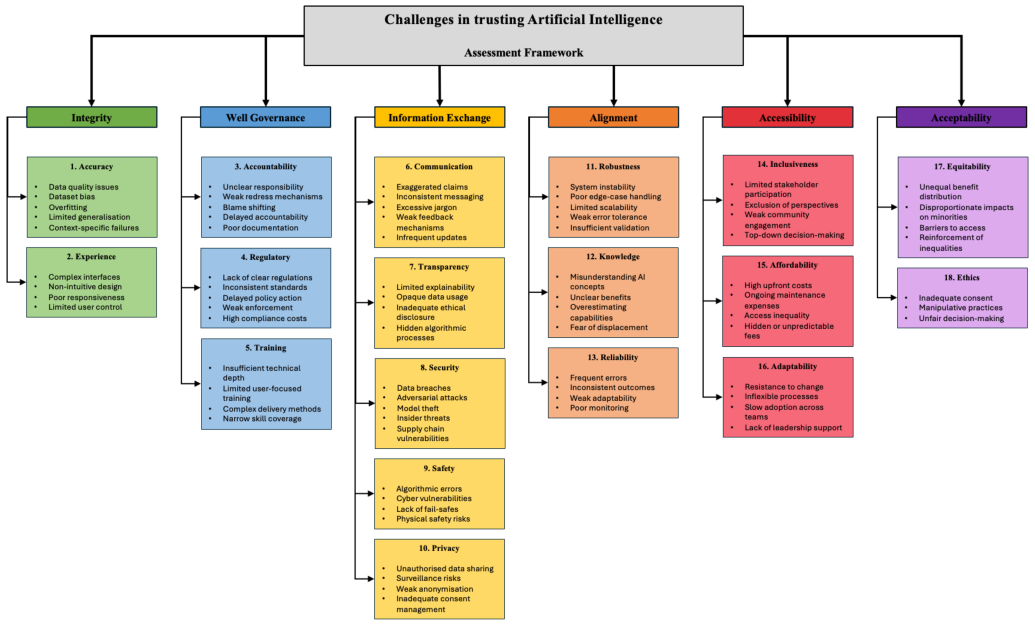

Fig. 4. Framework of indicators guiding user trust in AI technologies.

Capacity-building through education and training is also essential. Developers need a grounding not only in technical design but also in ethical principles such as fairness, inclusiveness, and safety. At the same time, users benefit from training that improves their understanding of AI capabilities and limitations. This dual education fosters more informed use, reducing unrealistic expectations and preventing overreliance on systems [67].

Finally, privacy and security represent areas where proactive investment can directly strengthen trust. Enhanced encryption, stronger consent management protocols, and protection against cyberattacks reassure users that their data is being managed responsibly. Transparent communication about these protections reinforces confidence that AI systems are both secure and accountable [67]. When combined with ethical design principles, these measures support not only compliance but also broader societal trust.

Taken together, these opportunities point to an integrated pathway forward. Technical advancements must be matched with governance, education, and communication initiatives, ensuring that AI systems are both functionally effective and socially acceptable. This integrated approach recognises that trust is not simply the absence of failure but the presence of reliability, fairness, and accountability across the entire lifecycle of AI systems [79].

4.8 Key Findings

This study identified 18 crucial indicators that significantly influence trust in AI systems (Figure 4). These indicators encompass both technical and ethical dimensions, providing a holistic understanding of how AI technologies are perceived and trusted by users. Among these, Table 24 highlights five indicators that have a particularly substantial impact on user confidence in AI technologies.

The findings demonstrate that for AI to be genuinely trusted, it must excel in both technical performance and ethical responsibility. Technical excellence provides assurance that systems operate reliably, efficiently, and safely, while ethical safeguards address wider social expectations

Table 24. Key Indicators that Influence Users Trust Towards AI Technologies

Indicator	Purpose
Transparency	Transparency is critical in AI systems, with users expecting clear explanations of how decisions are made, particularly in complex or high-stakes situations [28]. Understanding algorithms, risks, biases, and limitations allows users to verify processes, enhancing accountability [3].
Accountability	Accountability ensures responsibility for AI outcomes is clearly defined and traceable. Users need mechanisms to identify who is responsible when AI fails or makes errors [93]. Clear accountability structures reassure users that errors will be addressed appropriately, maintaining trust in systems that operate within ethical and legal frameworks [49].
Safety	Safety is paramount for AI systems, where users must trust that AI will operate reliably and without causing harm [79]. Incorporating fail-safes, emergency mechanisms, and rigorous testing ensures safe operation and fosters trust, particularly in high-risk applications [20].
Privacy	Privacy is essential for trust in AI, especially as systems handle sensitive personal data. Users need confidence that their data is secure from unauthorised access or misuse [29]. Strong privacy protections, such as encryption and anonymisation, not only ensure regulatory compliance but also reassure users that data is handled securely [13].
Reliability	Trust in AI depends on its ability to perform consistently and accurately over time. Users expect reliable outcomes across diverse contexts with minimal errors [49]. AI systems must operate continuously without failures to maintain user confidence [62].

regarding fairness, accountability, and respect for rights [94]. This dual emphasis recognises that trust in AI is not built on technical proficiency alone, but on a system's capacity to reflect and uphold societal values. By prioritising both dimensions, developers and policymakers can foster user confidence, encourage responsible adoption, and ensure long-term societal acceptance [19].

The analysis also reveals how these trust indicators are interconnected with the broader trust framework identified in this study. For example, transparency and accountability reinforce integrity and governance, ensuring decisions are explainable and responsibility is clearly defined [55]. Similarly, safety and reliability correspond with expectations of accessibility and acceptability, where systems must not only perform consistently but also operate in ways that are equitable and inclusive. Privacy strengthens trust in information exchange by safeguarding sensitive data and reinforcing the legitimacy of AI systems in the eyes of users [77]. This mapping underscores the multidimensional character of trust, confirming that no single factor is sufficient; rather, a combination of technical, ethical, and contextual measures must be integrated to sustain confidence [37].

While transparency, accountability, and privacy have received considerable academic attention, indicators such as inclusiveness, adaptability, and affordability are comparatively underexplored. These overlooked dimensions are critical for sectors such as construction, where digital literacy varies, costs must be controlled, and systems need to function reliably in complex and changing environments. Failure to address these factors risks widening inequalities and undermining adoption, particularly in industries with diverse workforces or limited resources. Recognising these gaps highlights important directions for future research and practice, ensuring that trust frameworks capture the full breadth of user expectation [95].

4.9 Research Contributions

This article investigates the role of trust indicators in AI systems and examines how they influence user trust and acceptance across multiple sectors. Drawing on a systematic literature review (n = 57), the study identifies 18 trust indicators organised into key categories, offering insights into how trust can be cultivated and sustained in contemporary AI applications. These indicators were analysed to assess their impact on the perceived trustworthiness of AI, with particular attention to transparency, accountability, reliability, and other critical aspects of user engagement.

The scope of this research is limited to trust indicators relevant to AI systems currently in operation, particularly those that directly affect system transparency and user interaction. While broader research into governance and ethics continues to expand, this study focuses on practical applications where trust is decisive for user acceptance.

The contributions of this article are as follows:

- A consolidated body of knowledge concerning trust indicators in AI systems and their influence on user trust and system adoption.
- An examination of the potential and existing applications of trust-building mechanisms in AI systems, highlighting the importance of transparency, accountability, and reliability.
- A review of the opportunities and challenges that AI developers face when integrating trust-building elements into AI technologies.
- Groundwork for future research, identifying areas where further investigation into underexplored indicators such as inclusiveness, adaptability, and affordability is needed to improve user confidence in AI systems.
- An analysis of how trust indicators function within operational environments, with emphasis on the role of transparency, ethical guidelines, and accountability mechanisms in strengthening user trust.

Future research should continue to investigate the adoption challenges associated with trust-building measures in AI systems and explore how these measures can be refined to foster greater user confidence, particularly in sectors where AI operates in complex or safety-critical contexts.

4.10 Research Limitations

The study has several limitations that should be acknowledged. (a) The scope of the research is constrained, focusing only on trust indicators identified through the literature review, and future studies could expand this scope to include additional dimensions of trust in AI or explore sector-specific challenges for a more comprehensive understanding. (b) Further literature reviews are needed to broaden the current findings and capture emerging opportunities and challenges in fostering trust, as AI continues to evolve rapidly. (c) AI trust issues remain dynamic, and the diversity of applications and sectors creates inconsistencies, since trust indicators may vary significantly depending on the use case, making it difficult to apply generalised frameworks across different contexts.

In addition, (d) this study is based primarily on secondary data, and primary research such as interviews or surveys with AI practitioners and users could provide valuable insights into real-world experiences and perspectives on the opportunities and challenges of trust-building. (e) The research also did not incorporate a SWOT analysis, which could offer a more strategic understanding of the strengths, weaknesses, opportunities, and threats associated with trust in AI. Given these limitations, future research should examine more practical aspects of trust-building in AI, including cost considerations, real-world implementation challenges, and user acceptance, with interviews and surveys providing a clearer picture of how trust-building mechanisms can be applied in practice to enhance AI trustworthiness.

5 Conclusion

This study conducted a systematic review of 18 trust indicators in AI systems, examining how they shape user trust and acceptance across multiple sectors. The review confirmed that trust in AI is multidimensional, requiring a balance between technical performance factors such as accuracy, reliability, and robustness and ethical principles including transparency, accountability, privacy, fairness, and security [64]. Technical capability is necessary but insufficient on its own, as ethical safeguards must be embedded into system design from the outset to ensure AI technologies are both effective and socially legitimate [96].

The findings demonstrate that trust in AI is highly context dependent. High-stakes environments such as healthcare, autonomous systems, and construction place greater emphasis on safety, accountability, and reliability, while lower-stakes domains often prioritise usability, convenience, and adaptability [16]. This variation shows that a one-size-fits-all approach to trust-building is inadequate and that sector-specific frameworks, accountability structures, and regulatory measures are required to ensure AI systems meet the unique needs of different industries [21].

Despite increasing recognition of the importance of trustworthy AI, challenges remain. Deep learning systems still operate as "black boxes," creating barriers to transparency and interpretability [70]. Users must be able to understand and, where appropriate, challenge AI outputs, particularly in safety-critical applications. Progress in explainable AI, supported by effective communication and engagement strategies, will be central to addressing these concerns [67]. Regulatory frameworks must also evolve in step with technological change to provide oversight and reinforce confidence in the responsible deployment of AI [40].

As AI becomes increasingly central to Industry 4.0, its potential to transform industries, improve processes, and mitigate risks is clear [24]. The cultivation of trust will remain a decisive factor in this transformation, shaping whether AI systems are embraced, contested, or rejected by society. The ability of developers, policymakers, and practitioners to address challenges of transparency, accountability, and inclusiveness while capitalising on opportunities for technical innovation will determine the trajectory of trustworthy AI. By embedding both technical excellence and ethical responsibility into AI systems, the path towards sustainable, widely accepted, and socially beneficial AI can be realised [97].

Acknowledgements

The authors thank the editor and anonymous referees for their invaluable comments on an earlier version of the manuscript.

References

[1] G. D'Amico, P. L'Abbate, W. Liao, T. Yigitcanlar, and G. Ioppolo. 2020. Understanding sensor cities: Insights from technology giant company driven smart urbanism practices. *Sensors* 20, 16 (2020), 4391.

[2] T. Yigitcanlar, K. Degirmenci, L. Butler, and K. Desouza. 2022. What are the key factors affecting smart city transformation readiness? *Evidence from Australian cities. Cities* 120, 1 (2022), 103434.

[3] S. Jain, M. Luthra, S. Sharma, and M. Fatima. 2020. Trustworthiness of artificial intelligence. In *Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems*. IEEE, 907–912.

[4] A. Kaplan, T. Kessler, J. Brill, and P. Hancock. 2023. Trust in artificial intelligence: Meta-analytic findings. *Human Factors* 65, 2 (2023), 337–359.

[5] T. Yigitcanlar, D. Agdas, and K. Degirmenci. 2023. Artificial intelligence in local governments: perceptions of city managers on prospects, constraints and choices. *AI and Society* 38, 3 (2023), 1135–1150.

[6] T. Yigitcanlar, K. Degirmenci, L. Butler, and K. Desouza. 2022. What are the key factors affecting smart city transformation readiness? *Evidence from Australian cities. Cities* 120, 1 (2022), 103434.

[7] J. Byabazaire, G. O'Hare, and D. Delaney. 2020. Data quality and trust: Review of challenges and opportunities for data sharing in IoT. *Electronics* 9, 12 (2020), 2083.

[8] R. Hoffman, S. Mueller, G. Klein, and J. Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023), 1096257.
[9] E. Glikson and A. Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.
[10] K. Reinhardt. 2023. Trust and trustworthiness in AI ethics. *AI and Ethics*, 3, 3 (2023), 735–744.
[11] A. Holzinger. 2021. The next frontier: AI we can really trust. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Cham: Springer International Publishing. 427–440.
[12] N. Omrani, G. Rivieccio, U. Fiore, F. Schiavone, and S. Agreda. 2022. To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change* 181 (2022), 121763.
[13] J. Zerilli, U. Bhatt, and A. Weller. 2022. How transparency modulates trust in artificial intelligence. *Patterns* 3, 4 (2022), 100455.
[14] W. Li, T. Yigitcanlar, W. Browne, and A. Nili. 2023a. The making of responsible innovation and technology: An overview and framework. *Smart Cities* 6, 4 (2023a), 1996–2034.
[15] S. Chowdhury, P. Budhwar, P. Dey, S. Joel-Edgar, and A. Abadie. 2022. AI-employee collaboration and business performance: Integrating knowledge-based view, socio-technical systems and organisational socialisation framework. *Journal of Business Research* 144, 1 (2022), 31–49.
[16] Y. Pan and L. Zhang. 2023. Integrating BIM and AI for smart construction management: Current status and future directions. *Archives of Computational Methods in Engineering* 30, 2 (2023), 1081–1110.
[17] D. Sargiotis. 2024. Ethical AI in information technology: Navigating bias, privacy, transparency, and accountability. *Adv Mach Lear Art Inte* 5, 3 (2024), 1–14.
[18] B. Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems* 10, 4 (2020), 1–31.
[19] M. Mylrea and N. Robinson. 2023. Artificial intelligence (AI) trust framework and maturity model: applying an entropy lens to improve security, privacy, and ethical AI. *Entropy* 25, 10 (2023), 1429.
[20] A. Kumar, T. Braud, S. Tarkoma, and P. Hui. 2020. Trustworthy AI in the age of pervasive computing and big data. In *Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 1–6.
[21] S. Na, S. Heo, S. Han, Y. Shin, and Y. Roh. 2022. Acceptance model of artificial intelligence (AI)-based technologies in construction firms: Applying the Technology Acceptance Model (TAM) in combination with the Technology–Organisation–Environment (TOE) framework. *Buildings* 12, 2 (2022), 90.
[22] O. Gillath, T. Ai, M. Branicky, S. Keshmiri, R. Davison, and R. Spaulding. 2021. Attachment and trust in artificial intelligence. *Computers in Human Behavior* 115, 1 (2021), 106607.
[23] M. Ryan. 2020. In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics* 26, 5 (2020), 2749–2767.
[24] M. Islam, G. Chen, and S. Jin. 2019. An overview of neural network. *American Journal of Neural Networks and Applications* 5, 1 (2019), 7–11.
[25] M. Agarwal and A. Saxena. 2019. An overview of natural language processing. *International Journal for Research in Applied Science and Engineering Technology* 7, 5 (2019), 2811–2813.
[26] K. Siau and W. Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.
[27] R. Marasinghe, T. Yigitcanlar, S. Mayere, T. Washington, and M. Limb. 2024. Computer vision applications for urban planning: A systematic review of opportunities and constraints. *Sustainable Cities and Society* 100, 1 (2024), 105047.
[28] R. Afshar, Y. Zhang, J. Vanschoren, and U. Kaymak. 2022. Automated reinforcement learning: An overview. arXiv:2201.05000. Retrieved from https://arxiv.org/abs/2201.05000
[29] S. Thiebes, S. Lins, and A. Sunyaev. 2021. Trustworthy artificial intelligence. *Electronic Markets* 31, 2 (2021), 447–464.
[30] J. Bareis. 2024. The trustification of AI. Disclosing the bridging pillars that tie trust and AI together. *Big Data and Society* 11, 2 (2024), 20539517241249430.
[31] T. Yigitcanlar, K. Degirmenci, and T. Inkinen. 2024. Drivers behind the public perception of artificial intelligence: insights from major Australian cities. *AI and Society* 39, 3 (2024), 833–853.
[32] O. Vereschak, G. Bailly, and B. Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.
[33] A. Bostrom, J. Demuth, C. Wirz, M. Cains, A. Schumacher, D. Madlambayan, A. Bansal, A. Bearth, R. Chase, K. Crosman, and I. Ebert-Uphoff. 2024. Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. *Risk Analysis* 44, 6 (2024), 1498–1513.
[34] S. Jangoan, G. Krishnamoorthy, M. Muthusubramanian, and K. Sharma. 2024. Demystifying explainable AI: Understanding, transparency, and trust. *International Journal for Multidisciplinary Research* 6, 2 (2024), 1–13.

[35] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou. 2023. Trustworthy AI: From principles to practices. *ACM Computing Surveys* 55, 9 (2023), 1–46.
[36] M. Lee and K. Rich. 2021. Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[37] M. Molina and S. Sundar. 2024. Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media and Society* 26, 6 (2024), 3638–3656.
[38] B. Knowles and J. Richards. 2021. The sanction of authority: Promoting public trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 262–271.
[39] H. Choung, P. David, and A. Ross. 2023. Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human–Computer Interaction* 39, 9 (2023), 1727–1739.
[40] P. Schmidt, F. Biessmann, and T. Teubner. 2020. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 29, 4 (2020), 260–278.
[41] C. Wickramasinghe, D. Marino, J. Grandio, and M. Manic. 2020. Trustworthy AI development guidelines for human system interaction. In *Proceedings of the 2020 13th International Conference on Human System Interaction*. IEEE, 130–136.
[42] T. Zhang, Y. Qin, and Q. Li. 2021. Trusted artificial intelligence: Technique requirements and best practices. In *Proceedings of the 2021 International Conference on Cyberworlds*. IEEE, 303–306.
[43] M. Braun, H. Bleher, and P. Hummel. 2021. A leap of faith: Is there a formula for "trustworthy" AI? *Hastings Center Report* 51, 3 (2021), 17–22.
[44] J. Kim, M. Giroux, and J. Lee. 2021. When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations. *Psychology and Marketing* 38, 7 (2021), 1140–1155.
[45] A. Schepman and P. Rodway. 2023. The general attitudes towards artificial intelligence scale (GAAIS): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human–Computer Interaction* 39, 13 (2023), 2724–2741.
[46] A. Ferrario and M. Loi. 2022. How explainability contributes to trust in AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1457–1466.
[47] K. Kuru. 2022. Trustfsdv: Framework for building and maintaining trust in self-driving vehicles. *IEEE Access* 10 (2022), 82814–82833.
[48] S. Lockey, N. Gillespie, D. Holm, and I. Someh. 2021. A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. *Proceeding of the 54th Hawaii International Conference (HICSS)*. Hawaii, USA, 5465.
[49] S. Dorton and S. Harper. 2022. A naturalistic investigation of trust, AI, and intelligence work. *Journal of Cognitive Engineering and Decision Making* 16, 4 (2022), 222–236.
[50] K. Crockett, M. Garratt, A. Latham, E. Colyer, and S. Goltz. 2020. Risk and trust perceptions of the public of artifical intelligence applications. In *Proceedings of the 2020 International Joint Conference on Neural Networks*. IEEE, 1–8.
[51] L. Butler, T. Yigitcanlar, and A. Paz. 2020. How can smart mobility innovations alleviate transportation disadvantage? Assembling a conceptual framework through a systematic review. *Applied Sciences* 10, 18 (2020), 6306.
[52] M. Ağca, S. Faye, and D. Khadraoui. 2022. A survey on trusted distributed artificial intelligence. *IEEE Access* 10 (2022), 55308–55337.
[53] A. Hasija and T. Esper. 2022. In artificial intelligence (AI) we trust: A qualitative investigation of AI technology acceptance. *Journal of Business Logistics* 43, 3 (2022), 388–412.
[54] R. Chatila, V. M. Fisher, F. Giannotti, K. Morik, S. Russell, and K. Yeung. 2021. Trustworthy AI. *Reflections on Artificial Intelligence for Humanity*. 13–39.
[55] Butler Tom, and Leona O'Brien. 2019. Artificial intelligence for regulatory compliance: Are we there yet? *Journal of Financial Compliance* 3, 1 (2019), 44–59.
[56] W. Li, T. Yigitcanlar, A. Nili, and W. Browne. 2023b. tech giants' responsible innovation and technology strategy: An international policy review. *Smart Cities* 6, 6 (2023b), 3454–3492.
[57] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, and T. Maharaj. 2020. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213. Retrieved from https://arxiv.org/abs/2004.07213
[58] Q. Liao and S. Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1257–1268.
[59] W. Li, T. Yigitcanlar, I. Erol, and A. Liu. 2021. Motivations, barriers and risks of smart home adoption: From systematic literature review to conceptual framework. *Energy Research and Social Science* 80 (2021), 102211.
[60] M. Mora-Cantallops, S. Sánchez-Alonso, E. García-Barriocanal, and M. Sicilia. 2021. Traceability for trustworthy AI: A review of models and tools. *Big Data and Cognitive Computing* 5, 2 (2021), 20.

[61] J. Balakrishnan and Y. Dwivedi. 2021. Role of cognitive absorption in building user trust and experience. *Psychology and Marketing* 38, 4 (2021), 643–668.
[62] H. Choung, P. David, and A. Ross. 2023a. Trust and ethics in AI. *AI and Society* 38, 2 (2023a), 733–745.
[63] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. Zelaya, and A. Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 272–283.
[64] G. Nadella, S. Meduri, H. Gonaygunta, and S. Podicheti. 2023. Understanding the role of social influence on consumer trust in adopting AI tools. *International Journal of Sustainable Development in Computing Science* 5, 2 (2023), 1–18.
[65] P. Shinde and S. Shah. 2018. A review of machine learning and deep learning applications. In *Proceedings of the 2018 4th International Conference on Computing Communication Control and Automation*. IEEE, 1–6.
[66] I. Troshani, S. Rao Hill, C. Sherman, and D. Arthur. 2021. Do we trust in AI? Role of anthropomorphism and intelligence. *Journal of Computer Information Systems* 61, 5 (2021), 481–491.
[67] A. Papenmeier, D. Kern, G. Englebienne, and C. Seifert. 2022. It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction* 29, 4 (2022), 1–33.
[68] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 624–635.
[69] M. Vössing, N. Kühl, M. Lind, and G. Satzger. 2022. Designing transparency for effective human-AI collaboration. *Information Systems Frontiers* 24, 3 (2022), 877–895.
[70] F. Chen, J. Zhou, A. Holzinger, K. Fleischmann, and S. Stumpf. 2023. Artificial Intelligence ethics and trust: From principles to practice. *IEEE Intelligent Systems* 38, 6 (2023), 5–8.
[71] S. Bejger and S. Elster. 2020. Artificial Intelligence in economic decision making: how to assure a trust? Ekonomia i prawo. *Economics and Law* 19, 3 (2020), 411–434.
[72] M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, and T. Janowski. 2020. Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly* 37, 3 (2020), 101493.
[73] A. Leschanowsky, S. Rech, B. Popp, and T. Bäckström. 2024. Evaluating privacy, security, and trust perceptions in conversational AI. arVix preprint ArXiv:2406.09037. Retrieved from https://arxiv.org/abs/2406.09037
[74] O. Bitkina, H. Jeong, B. Lee, J. Park, J. Park, and H. Kim. 2020. Perceived trust in artificial intelligence technologies: A preliminary study. *Human Factors and Ergonomics in Manufacturing & Service Industries* 30, 4 (2020), 282–290.
[75] A. McNamara and S. Sepasgozar. 2021. Intelligent contract adoption in the construction industry: Concept development. *Automation in Construction* 122 (2021), 103452.
[76] L. Robert Jr, G. Bansal, N. Melville, and T. Stafford. 2020. Introduction to the special issue on AI fairness, trust, and ethics. *AIS Transactions on Human-Computer Interaction* 12, 4 (2020), 172–178.
[77] R. Shrestha, K. Kafle, and C. Kanan. 2022. An investigation of critical issues in bias mitigation techniques. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1943–1954.
[78] Smuha. 2019. The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International* 20, 4 (2019), 97–106.
[79] R. Yang and S. Wibowo. 2022. User trust in artificial intelligence: A comprehensive conceptual framework. *Electronic Markets* 32, 4 (2022), 2053–2077.
[80] V. Chamola, V. Hassija, A. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar. 2023. A review of trustworthy and explainable artificial intelligence (xai). *IEEe Access* 11 (2023), 78994–79015.
[81] A. Andreotta, N. Kirkham, and M. Rizzi. 2022. AI, big data, and the future of consent. *Ai and Society* 37, 4 (2022), 1715–1728.
[82] T. Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines* 30, 1 (2020), 99–120.
[83] F. Rossi. 2018. Building trust in artificial intelligence. *Journal of International Affairs* 72, 1 (2018), 127–134.
[84] L. Dhirani, N. Mukhtiar, B. Chowdhry, and T. Newe. 2023. Ethical dilemmas and privacy issues in emerging technologies: A review. *Sensors* 23, 3 (2023), 1151.
[85] Y. Zhang, Q. Liao, and R. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.
[86] M. Rotta, D. Sell, R. dos Santos Pacheco, and T. Yigitcanlar. 2019. Digital commons and citizen coproduction in smart cities: Assessment of Brazilian municipal e-government platforms. *Energies* 12, 14 (2019), 2813.
[87] O. Akinrinola, C. Okoye, O. Ofodile, and C. Ugochukwu. 2024. Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. *GSC Advanced Research and Reviews* 18, 3 (2024), 050–058.
[88] T. Yigitcanlar, K. Desouza, L. Butler, and F. Roozkhosh. 2020. Contributions and risks of artificial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature. *Energies* 13, 6 (2020), 1473.

[89] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux. 2020. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics* 26, 6 (2020), 3333–3361.
[90] D. Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.
[91] R. Alvarado. 2023. What kind of trust does AI deserve, if any? *AI and Ethics* 3, 4 (2023), 1169–1183.
[92] D. Kaur, S. Uslu, K. Rittichier, and A. Durresi. 2022. Trustworthy artificial intelligence: A review. *ACM Computing Surveys* 55, 2 (2022), 1–38.
[93] D. Varona and J. Suárez. 2022. Discrimination, bias, fairness, and trustworthy AI. *Applied Sciences* 12, 12 (2022), 5826.
[94] S. Wiesmüller. 2023. Conceptualisation of the relational governance of artificial intelligence. In *Proceedings of the Relational Governance of Artificial Intelligence: Forms and Interactions*. Cham: Springer. 91–163.
[95] K. De Fine Licht and J. de Fine Licht. 2020. Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI and Society* 35, 4 (2020), 917–926.
[96] M. Ashoori and J. Weisz. 2019. In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. arXiv: 1912.02675. Retrieved from https://arxiv.org/abs/1912.02675
[97] N. Shadbolt, K. O'Hara, D. De Roure, W. Hall, N. Shadbolt, K. O'Hara, D. De Roure, and W. Hall. 2019. Privacy, trust and ethical issues. In *Proceedings of the Theory and Practice of Social Machines*. 149–200, Cham: Springer.

Appendix A: Salient Characteristics of the Reviewed Literature

Title	Authors	Year	Region	Journal	Accuracy	Experience	Privacy	Accountability	Training	Regulatory	Communication	Transparency	Security	Robustness	Knowledge	Reliability	Inclusiveness	Affordability	Adaptability	Equitability	Safety	Ethics
					Integrity			Well-governance			Information Exchange			Alignment			Accessibility			Acceptability		
Artificial Intelligence Ethics and Trust: From Principles to Practice	Chen, F., Zhou, J., Holzinger, A., Fleischmann, K.R. and Stumpf, S.	2023	Europe	IEEE Intelligent Systems	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓	✓
Trustworthy Artificial Intelligence: A Review	Kaur, D., Uslu, S., Rittichier, K.J. and Durresi, A.	2022	Europe	ACM Computing Surveys	✗	✓	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗
Trustworthy AI: From Principles to Practices	Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B.	2023	Europe	ACM Computing Surveys	✓	✓	✗	✓	✓	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗	✓
Trustworthy AI	Chatila, R., Evers, V.	2021	Europe	Reflections on artificial intelligence for humanity	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✓	✓
A Leap of Faith: Is There a Formula for "Trustworthy" AI?	Braun, M., Bleher, H., Hummel, P.	2021	Europe	Hastings Center Report	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗
Trust and trustworthiness in AI ethics	Reinhardt, K.	2023	Europe	AI and Ethics	✓	✗	✗	✓	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗	✓
Trustworthiness of Artificial Intelligence	Jain, S., Krishna, T., Verma, M., Gupta, A.	2020	Europe	2020 6th International Conference on Advanced Computing and Communication Systems	✗	✗	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗	✗	✓	✓	✗

(Continued)

Continued

Title	Authors	Year	Region	Journal	Accuracy	Experience	Privacy	Accountability	Training	Regulatory	Communication	Transparency	Security	Robustness	Knowledge	Reliability	Inclusiveness	Affordability	Adaptability	Equitability	Safety	Ethics
					Integrity			Well-governance			Information Exchange			Alignment			Accessibility			Acceptability		
Trustworthy AI in the Age of Pervasive Computing and Big Data	Kumar, A., Srinivas, B.	2020	Europe	2021 6th International Conference on Advanced Computing and Communication Systems	✓	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗
A Review of Trustworthy and Explainable Artificial Intelligence (XAI)	Chamola, V., Hassija, V., Gupta, V., Guizani, M.	2023	Europe	IEEE Access	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗	✓
Trustworthy AI Development Guidelines for Human System Interaction	Wickramasinghe, C., Ahmed, M., Weerasinghe, S.	2020	Europe	2020 6th International Conference on Advanced Computing and Communication Systems	✓	✗	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✗	✓	✓	✗	✗
Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems	Shneiderman, B.	2020	Europe	ACM Transactions on Interactive Intelligent Systems (TiiS)	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✓	✗	✓	✓	✗	✓	✓	✓
Data governance: Organizing data for trustworthy Artificial Intelligence	Janssen, M., Brous, P., Estévez, E., Barbosa, L. S., Janowski, T.	2020	Europe	Government information quarterly	✗	✓	✗	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗
Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims	Brundage, M., Avin, S., Wang, J., Krueger, G., Hadfield, G., and Dafoe, A.	2020	Europe	Computers and Society	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Trustworthy artificial intelligence	Thiebes, S., Lins, S., and Sunyaev, A.	2021	Europe	Electronics Markets	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗	✓	✗
The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust	Schepman, A., Rodway, P., and Read, S.	2023	Europe	International Journal of Human–Computer Interaction	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓	✓

(Continued)

Continued

Title	Authors	Year	Region	Journal	Accuracy	Experience	Privacy	Accountability	Training	Regulatory	Communication	Transparency	Security	Robustness	Knowledge	Reliability	Inclusiveness	Affordability	Adaptability	Equitability	Safety	Ethics
					Integrity			Well-governance			Information Exchange			Alignment			Accessibility			Acceptability		
Trust in Artificial Intelligence: Meta-Analytic Findings	Kaplan, A., Bond, S. D., and Pearson, G. D.	2023	Europe	Human Factors	✗	✗	✓	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓
Acceptance Model of Artificial Intelligence (AI)-Based Technologies in Construction Firms: Applying the Technology Acceptance Model (TAM) in Combination with the Technology–Organisation–Environment (TOE) Framework	Na, S., Lee, J. H., Baek, J. S., and Kim, H.	2022	North America	Buildings	✗	✓	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗	✓	✗	✓
How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies	Vereschak, O., Liao, Q. V., Wang, D., and Riedl, M.	2021	North America	Proceedings of the ACM on Human-Computer Interaction	✗	✗	✓	✗	✓	✗	✓	✓	✗	✗	✓	✗	✓	✗	✗	✓	✓	✓
Role of cognitive absorption in building user trust and experience	Balakrishnan, J., and Sundar, S.	2021	North America	Psychology and Marketing	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗	✗
Understanding the Role of Social Influence on Consumer Trust in Adopting AI Tools	Nadella, G., Pandey, R., and Singh, V.	2023	North America	International Journal of Sustainable Development in Computing Science	✓	✗	✓	✗	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✓	✗	✓	✗
Integrating BIM and AI for Smart Construction Management: Current Status and Future Directions	Pan, Y., Zhang, L., and Ji, S.	2023	Asia	Archives of Computational Methods in Engineering	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Intelligent contract adoption in the construction industry: Concept development	Mcnamara, A., Sepasgozar, S., and Karimi, A.	2021	Asia	Automation in construction	✗	✗	✓	✗	✓	✗	✓	✓	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗

(Continued)

Continued

Title	Authors	Year	Region	Journal	Accuracy	Experience	Privacy	Accountability	Training	Regulatory	Communication	Transparency	Security	Robustness	Knowledge	Reliability	Inclusiveness	Affordability	Adaptability	Equitability	Safety	Ethics
					Integrity			Well-governance			Information Exchange			Alignment			Accessibility			Acceptability		
The Next Frontier: AI We Can Really Trust	Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B.	2021	North America	In Joint European conference on machine learning and knowledge discovery in databases	✗	✗	✓	✗	✓	✗	✓	✓	✓	✗	✓	✗	✓	✗	✗	✓	✓	✓
Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation	Molina, M. D., Sundar, S. S., and Le, T.	2024	North America	New Media and Society	✗	✓	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗
When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations	Kim, J., and Sundar, S. S.	2021	North America	Psychology and Marketing	✓	✗	✗	✓	✗	✗	✓	✓	✗	✓	✗	✓	✗	✗	✗	✓	✓	✓
Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance.	Hoffman, R.R., Mueller, S.T., Klein, G. and Litman, J.	2023	North America	Frontiers in Computer Science	✓	✗	✓	✓	✗	✓	✗	✗	✓	✓	✗	✓	✓	✗	✗	✗	✓	✗
Artificial Intelligence (AI) trust framework and maturity model: applying an entropy lens to improve security, privacy, and ethical AI	Mylrea, M., and Gourisetti, S. N. G.	2023	Asia	Entropy	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗	✓
Trust in AI and its role in the acceptance of AI technologies	Choung, H., David, P. and Ross, A.	2023	Europe	International Journal of Human–Computer Interaction	✗	✗	✓	✓	✓	✗	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗
Trust and ethics in AI	Choung, H., David, P. and Ross, A.Choung, H., David, P. and Ross, A.	2023	Europe	Ai and Society	✓	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓	✗	✓
The relationship between user trust, model accuracy and explanations in AI.	Papenmeier, A., and Peters, J.	2022	North America	ACM Transactions on Computer-Human Interaction	✗	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✓	✗

(Continued)

Continued

Title	Authors	Year	Region	Journal	Accuracy	Experience	Privacy	Accountability	Training	Regulatory	Communication	Transparency	Security	Robustness	Knowledge	Reliability	Inclusiveness	Affordability	Adaptability	Equitability	Safety	Ethics
					Integrity			Well-governance			Information Exchange			Alignment			Accessibility			Acceptability		
In artificial intelligence (AI) we trust: A qualitative investigation of AI technology acceptance	Hasija, A. and Esper, T.L	2022	North America	Journal of Business Logistics	✓	✗	✗	✓	✗	✓	✓	✓	✗	✗	✗	✓	✓	✗	✗	✓	✗	✓
AI, big data, and the future of consent	Andreotta, A.J., Kirkham, N. and Rizzi, M.	2022	North America	Ai and Society	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓	✗
How explainability contributes to trust in AI	Ferrario, A. and Loi, M.	2022	North America	Proceedings of the 2022 ACM Conference on Fairness, Accountability and Transparency	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗
Designing transparency for effective human-AI collaboration	Vössing, M., Kühl, N., Lind, M. and Satzger, G.	2022	Oceania	Information Systems Frontiers	✗	✗	✓	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓
Designing for responsible trust in AI systems: A communication perspective	Liao, Q. V., Gruen, D. M., and Miller, S.	2022	Oceania	Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency	✓	✓	✗	✓	✗	✗	✓	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
How transparency modulates trust in artificial intelligence	Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C.	2022	North America	Patterns	✗	✗	✓	✗	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗
A survey on trusted distributed artificial intelligence	Ağca, M. A., and Tsiatsis, V.	2022	Europe	IEEE Access	✗	✗	✓	✓	✗	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✓	✓	✓
To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts	Omrani, N., Vermeir, J., and Vinck, B.	2022	Other Regions	Technological Forecasting and Social Change	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓
A naturalistic investigation of trust, AI, and intelligence work	Dorton, S. L., and Pirolli, P.	2022	North America	Journal of Cognitive Engineering and Decision Making	✓	✓	✗	✓	✗	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗
Discrimination, bias, fairness, and trustworthy AI	Varona, D., and Bickmore, T.	2022	North America	Applied Sciences	✗	✗	✓	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✓	✓	✓

(Continued)

Continued

Title	Authors	Year	Region	Journal	Accuracy	Experience	Privacy	Accountability	Training	Regulatory	Communication	Transparency	Security	Robustness	Knowledge	Reliability	Inclusiveness	Affordability	Adaptability	Equitability	Safety	Ethics
					Integrity			Well-governance			Information Exchange			Alignment			Accessibility			Acceptability		
Traceability for trustworthy AI: a review of models and tools	Mora-Cantallops, M., and Sicilia, M. A.	2021	Europe	Big Data and Cognitive Computing	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions	Lockey, S., Gillespie, N., and Dietz, G.	2021	Europe	Hawaii International Conference on System Sciences	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Trusted artificial intelligence: technique requirements and best practices	Zhang, T., and Dafoe, A.	2021	Europe	2021 International Conference on Cyberworlds	✗	✗	✓	✓	✗	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗
The sanction of authority: Promoting public trust in AI	Knowles, B. and Richards, J.T.	2021	Europe	Proceedings of the 2021 ACM conference on fairness, accountability and transparency	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗	✓
Attachment and trust in artificial intelligence	Gillath, O., Ai, T., Branicky, M.S., Keshmiri, S., Davison, R.B. and Spaulding, R.	2021	Europe	Computers in Human Behavior	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust	Lee, M. K., and Karahalios, K.	2021	Asia	Proceedings of the 2021 CHI conference on human factors in computing systems	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	✓	✗
Do we trust in AI? Role of anthropomorphism and intelligence	Troshani, I., Jazi, M. D., and Sanderson, M.	2021	Asia	Journal of Computer Information Systems	✗	✓	✓	✗	✓	✗	✓	✓	✓	✗	✓	✗	✗	✓	✓	✗	✓	✗
Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI	Jacovi, A., Marasovic, A., and Miller, T.	2021	Europe	Proceedings of the 2021 ACM conference on fairness, accountability, and transparency	✗	✓	✗	✗	✓	✗	✓	✓	✗	✓	✗	✗	✗	✓	✗	✓	✗	✓
Transparency and trust in artificial intelligence systems	Schmidt, P., and Biessmann, F.	2020	Europe	Decision Systems	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓	✓	✓	✗

(Continued)

Continued

Title	Authors	Year	Region	Journal	Accuracy	Experience	Privacy	Accountability	Training	Regulatory	Communication	Transparency	Security	Robustness	Knowledge	Reliability	Inclusiveness	Affordability	Adaptability	Equitability	Safety	Ethics
					Integrity			Well-governance			Information Exchange			Alignment			Accessibility			Acceptability		
Human trust in artificial intelligence: Review of empirical research.	Glikson, E., Cheshin, A., and Avidov-Ungar, O.	2020	North America	Academy of Management Annals	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓
Data quality and trust: Review of challenges and opportunities for data sharing in iot.	Byabazaire, J., Matovu, F., and Namayanja, J.	2020	Oceania	Electronics	✗	✓	✓	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗
The relationship between trust in AI and trustworthy machine learning technologies	Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., and van Moorsel, A.	2020	Europe	Proceedings of the 2020 conference on fairness, accountability and transparency	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✓	✗	✓
In AI we trust: ethics, artificial intelligence, and reliability	Ryan, M., and Stahl, B. C.	2020	Oceania	Science and Engineering Ethics	✓	✗	✗	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✗	✓
Risk and trust perceptions of the public of artificial intelligence applications	Crockett, K., and Latham, A.	2020	Other Regions	2020 International Joint Conference on Neural Networks	✓	✓	✓	✗	✓	✗	✓	✓	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗
Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making	Zhang, Y., Yao, X., and Li, P.	2020	Other Regions	Proceedings of the 2020 conference on fairness, accountability, and transparency	✓	✓	✗	✓	✗	✓	✓	✓	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗
Artificial Intelligence in economic decision making: how to assure a trust?	Bejger, S., and Koenigstein, N.	2020	Other Regions	Ekonomia i prawo. Economics and Law,	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
Introduction to the special issue on AI fairness, trust, and ethics	Robert Jr, L. P., Pierce, C. E., and Marlow, N.	2020	Other Regions		✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✓	✓	✓

Received 12 December 2024; revised 17 November 2025; accepted 8 January 2026