Human-AI Trust in the Workplace: A Survey

www.surveyx.cn

Abstract

This survey paper examines the dynamic evolution of trust between humans and AI systems, emphasizing its impact on employees' wellbeing, the methodologies for measuring trust levels, intervention strategies to enhance trust, and the integration of workplace technology to support these processes. The paper highlights the foundational definitions and frameworks of human-AI trust, exploring its cognitive and emotional dimensions and the factors influencing its evolution over time. Key findings reveal that trust is a multifaceted construct, essential for AI acceptance and effective utilization in workplace environments, particularly in safety-critical areas such as healthcare. The survey reviews traditional and innovative trust measurement techniques, including adaptive trust calibration and explainable AI methodologies, underscoring their importance in fostering user confidence. Intervention strategies are discussed, focusing on training programs that incorporate personality traits, transparent communication, and human-in-the-loop approaches to enhance trust. The role of AI representations, large language models, and AI decision support systems in shaping trust dimensions is analyzed, highlighting the need for ethical considerations and transparent communication. The paper concludes by addressing the challenges and future directions in the field, advocating for refined trust evaluation methods and the development of personalized explanation interfaces to support user understanding and trust. By synthesizing these insights, the survey provides a comprehensive understanding of the ongoing development of human-AI trust and its implications for workplace technology integration.

1 Introduction

1.1 Importance of Human-AI Trust

Human-AI trust is essential in contemporary workplaces, significantly impacting user satisfaction and performance in interactions with AI systems [1]. Establishing trust in AI is critical for its acceptance and effective use across various domains [2]. As AI technologies become integral to organizational processes, a structured understanding of trust is vital for their successful implementation [3]. This is particularly relevant in fields such as science, health, and humanity, where AI integration can dramatically alter workplace dynamics [4].

Trust plays a crucial role in interactions with technology, especially when dealing with systems that do not emulate human thought processes [5]. The transformative influence of AI on decision-making highlights the necessity of understanding trust dynamics in technology [6]. This understanding is particularly critical in complex environments where AI is expected to support human decision-making [7]. Moreover, the reliability of AI predictions is paramount, especially in high-stakes applications where erroneous decisions can lead to severe consequences [8].

The dynamics of trust in digital human-AI collaboration further underscore the importance of comprehending how AI impacts trust within teams [9]. As AI becomes ubiquitous, understanding user trust in AI-enhanced technologies is crucial, especially in safety-critical fields like medicine, where reactions to AI recommendations can significantly affect decision-making [10]. The complexity

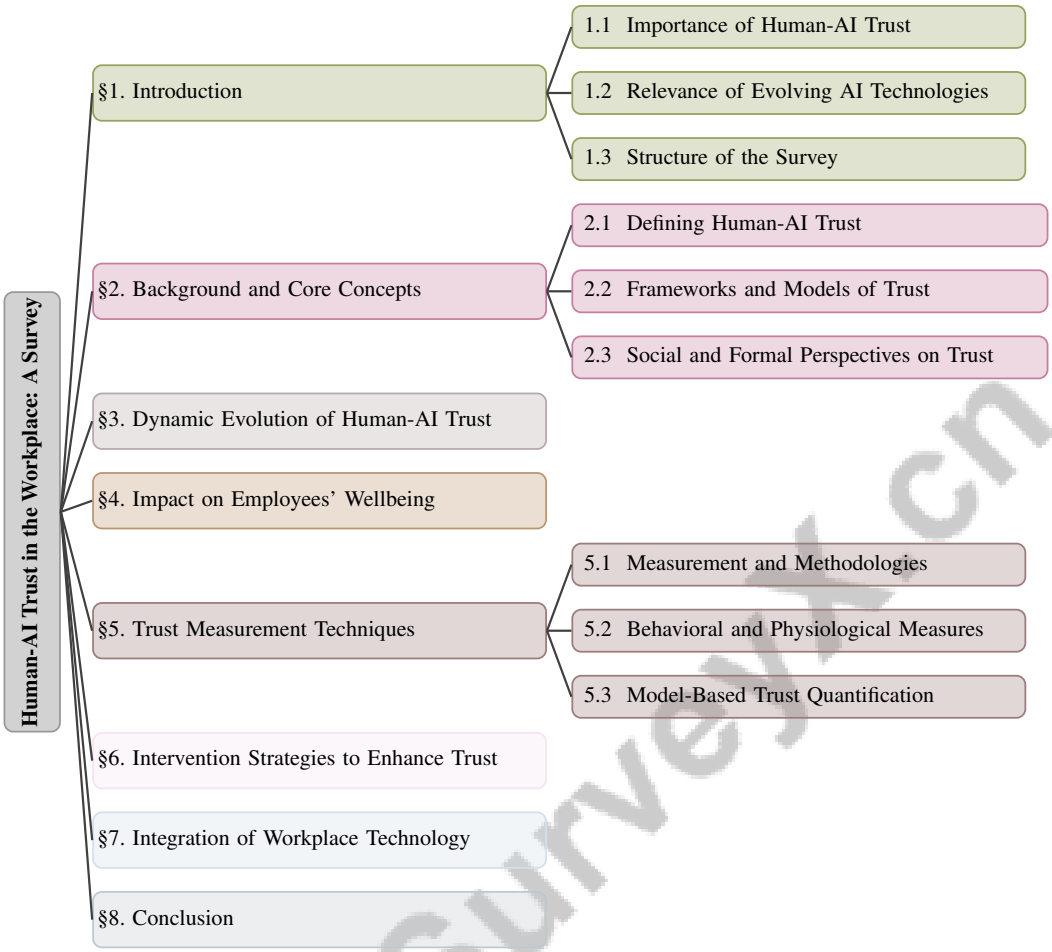


Figure 1: chapter structure

and opacity of AI systems often breed distrust, particularly in medicine, where clarity in decision-making processes is essential [11].

Furthermore, developing trustworthy AI systems increasingly draws from psychological insights into trust formation, exploring interpersonal, human-automation, and human-AI trust dynamics [12]. Addressing the fragmented understanding of trust in human-AI interactions is key to establishing a standardized framework for studying AI trust [13]. As workplace technologies evolve, their effects on job motivation and employee engagement further highlight the necessity of fostering a trusting relationship between humans and AI [14]. Cultivating human-AI trust is thus vital for maximizing the benefits of AI technologies in the workplace.

1.2 Relevance of Evolving AI Technologies

The rapid evolution of AI technologies is reshaping workplace dynamics and the trust relationships between humans and AI systems. As AI becomes increasingly integrated into sectors like healthcare, finance, and autonomous systems, its influence extends beyond technological advancements to affect human interactions and trust [2]. This integration necessitates a nuanced understanding of AI representation and machine intelligence levels, which are crucial for shaping trust in AI systems [4].

The lack of comprehensive models for understanding trust in human-AI interactions complicates decision-making, particularly in high-stakes areas such as healthcare and finance [5]. The dynamic nature of human-AI teams emphasizes the importance of understanding trust within these collaborative environments, where AI's role is increasingly significant [6]. Research on AI's impact on team collaboration highlights the need for a sophisticated comprehension of trust dynamics between human and AI collaborators [11].

As AI technologies permeate daily life, fostering a deeper understanding of trust in AI is essential for technology acceptance and effective utilization [15]. The fragmented research landscape regarding the impact of evolving workplace technologies on job motivation further underscores the necessity to cultivate trust [16]. Addressing the challenges of establishing trust in various human-AI interaction contexts is crucial, particularly given the black-box nature of AI systems, which limits trust and complicates the adoption of AI recommendations in critical decision-making scenarios. Thus, the ongoing evolution of AI technologies calls for a comprehensive approach to understanding and building trust in workplace settings.

1.3 Structure of the Survey

This survey is organized to provide an in-depth exploration of human-AI trust within workplace settings, structured around several key themes. The introduction emphasizes the significance of human-AI trust and the relevance of evolving AI technologies, setting the stage for subsequent discussions. The survey then examines foundational definitions and various frameworks and models of trust, categorizing research based on AI representation and machine intelligence, while highlighting cognitive and emotional trust dimensions [17].

Next, the dynamic evolution of human-AI trust is analyzed, focusing on its progression over time and the factors influencing this evolution. The impact of trust on employee well-being, including job satisfaction and mental health, is scrutinized. The survey also reviews trust measurement techniques, encompassing traditional methodologies and innovative approaches like those within the Technology Acceptance Model (TAM) framework, which stress the importance of trust in technology acceptance [18].

The discussion on intervention strategies to enhance trust in AI-human interactions highlights the need for comprehensive training programs, transparent communication practices, and human-in-the-loop approaches. These strategies aim to bridge the knowledge gap regarding consumer perceptions of AI, particularly in news media and industrial applications, and are essential for building confidence in AI technologies as they become increasingly integrated into decision-making processes across various sectors [13, 18, 16, 9, 19]. The role of workplace technology in supporting trust development is analyzed, focusing on AI representations and the influence of large language models and AI decision support systems. The survey concludes by summarizing key findings and proposing future research directions, addressing ongoing challenges in the field of human-AI trust.The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Defining Human-AI Trust

Human-AI trust is a complex and essential element for the effective integration of AI in organizational settings. It is defined by user confidence in AI systems, significantly influenced by the quality of system explanations, which affect user comprehension and satisfaction [1]. Trust is crucial for viewing AI as reliable partners in decision-making, enhancing collaborative efficiency in workplaces [3]. Ferrario et al.'s incremental trust model, which categorizes trust into simple, reflective, and paradigmatic forms, highlights the layered development of trust in both human-human and human-AI interactions [3].

The creation and maintenance of trust in AI are vital amidst ethical concerns and privacy risks, necessitating responsible AI deployment [12]. In dynamic task environments, trust in human-AI teams evolves over time and is shaped by task nature and roles [7]. Trust dynamics in digital human-AI collaborations are influenced by varying roles and tasks [9].

Trust influences user perceptions and acceptance of AI technologies, underscoring its significance for workplace integration [18]. Trust's context-dependent nature complicates establishing appropriate trust levels, as user perceptions vary with applications and environments [13]. The impact of workplace technologies on employee motivation highlights the need for cultivating trust in AI systems, as these technologies transform work environments and affect engagement [14].

Understanding how individuals use advice from AI versus human peers is crucial, influenced by beliefs about the performance of these sources [10]. The introduction of AI teammates can alter

trust perceptions and behaviors, differing from those associated with human teammates, revealing unique dynamics of human-AI trust [20]. Modeling trust in AI is complex due to the intricate nature of human trust [5]. Trust in AI is a cognitive mechanism enabling users to anticipate AI behavior [15]. Viewing AI as a more reliable decision-maker compared to humans, who may be seen as biased, adds complexity to trust dynamics [6].

The complexity and transparency of AI decision-making processes significantly affect human-AI trust [11]. The trust score, a metric quantifying agreement between a classifier's prediction and a modified nearest-neighbor classifier, enhances prediction reliability understanding [8]. Addressing trust and distrust in AI is crucial for technology acceptance across various domains [2]. The lack of comprehensive models complicates decision-making, particularly in high-stakes areas like healthcare and finance [5]. Defining and fostering human-AI trust is pivotal for maximizing AI's potential benefits in workplace settings.

2.2 Frameworks and Models of Trust

Understanding trust dynamics in human-AI interactions requires exploring frameworks and models that define and assess trust's complexities. Trust is inherently multidimensional and context-dependent, evolving over time [13]. This understanding is crucial for developing robust frameworks that accommodate trust's dynamic nature across diverse scenarios.

The activation-integration model delineates advice utilization stages, offering insights into trust dynamics in human-AI interactions [10]. This complements the taxonomy of trustworthiness metrics, categorizing trust into technical dimensions like safety, accuracy, and robustness, and non-technical dimensions such as ethics and legality [2]. These taxonomies systematically assess AI system trust by considering performance and ethical implications.

Shen et al.'s framework emphasizes behavior certificates for assessing AI trust, focusing on interpretability and reliability [4]. This aligns with research adopting dual perspectives through social and formal lenses, integrating concepts like trusting beliefs and disposition to trust.

Jacovi et al.'s formalization of trust highlights vulnerability and the ability to anticipate AI decisions, critical for understanding trust dynamics [15]. This approach is enriched by frameworks categorizing trust as influenced by system characteristics and user attributes, particularly in automated driving contexts [16].

The trust score measures classifier reliability based on the spatial distribution of training examples, integrating statistical learning with knowledge representations to enhance AI robustness and explainability. This highlights transparency and user-centric evaluation methods' role in fostering trust, particularly in sectors like manufacturing and news media. Research indicates openness in AI processes significantly influences consumer perceptions and confidence, necessitating systematic methodologies—such as controlled experiments and surveys—to assess transparency levels' effects on user trust. As AI technologies proliferate, understanding trust's multidimensional nature, encompassing human-like and functionality-related trust, is essential for enhancing user acceptance and ensuring ethical AI development [18, 19].

2.3 Social and Formal Perspectives on Trust

Examining trust from social and formal perspectives is essential for understanding its multifaceted nature in human-AI interactions. Social perspectives focus on relational and emotional aspects, highlighting how interpersonal dynamics and user experiences shape trust in AI systems. Factors like faith in technology, familiarity, emotional response, capability-based trust, and affective trust are crucial for assessing user engagement with AI [21]. These dimensions emphasize users' prior experiences and emotional connections in forming trust, influencing acceptance and reliance on AI.

Conversely, formal perspectives emphasize structural and procedural components underpinning trust dynamics. These perspectives use quantifiable measures and models to evaluate AI systems' reliability and predictability. Key elements include assessing contractual trust, anticipating AI decision outcomes, and using behavior certificates aggregating evidence to predict future model behavior. This approach contrasts with subjective trust interpretations, establishing a framework for understanding trust-building and evaluation in human-AI interactions, addressing potential AI misuse or disuse [15, 19, 4, 6]. This evaluates AI's technical competence, including accuracy, safety, and

robustness, critical for ensuring consistent and trustworthy performance. Integrating formal metrics with social factors provides a comprehensive framework for evaluating trust, emphasizing AI systems' need to be technically sound and socially attuned to user needs.

The convergence of social and formal perspectives is relevant in evolving workplace technologies, where trust shapes employee motivation and engagement. Despite AI's increasing integration in organizational settings, substantial gaps remain in understanding how technologies influence motivation and contextual factors mediating this relationship [14]. Addressing these gaps requires a holistic approach considering AI's technical capabilities and social dynamics of human interactions, fostering a nuanced understanding of trust in AI systems. This dual perspective is crucial for designing AI technologies that are efficient and reliable and resonate with users on emotional and relational levels, enhancing acceptance and integration into workplace environments.

In recent years, the exploration of human-AI trust has gained significant attention within the field of artificial intelligence. Understanding the nuances of this trust is crucial for developing systems that foster positive user experiences and enhance collaboration between humans and AI. As illustrated in Figure 2, the hierarchical structure of key concepts in the dynamic evolution of human-AI trust is depicted, highlighting the primary categories of trust evolution. This figure elucidates the various factors influencing trust, as well as the intricate relationship between trust dynamics and psychological safety. Each primary category is further delineated into subcategories, emphasizing essential cognitive and emotional factors, system performance, transparency, user experience, and the pivotal role of psychological safety in the successful integration of AI technologies. This comprehensive framework not only aids in understanding the complexities of trust but also serves as a foundation for future research and development in the field.

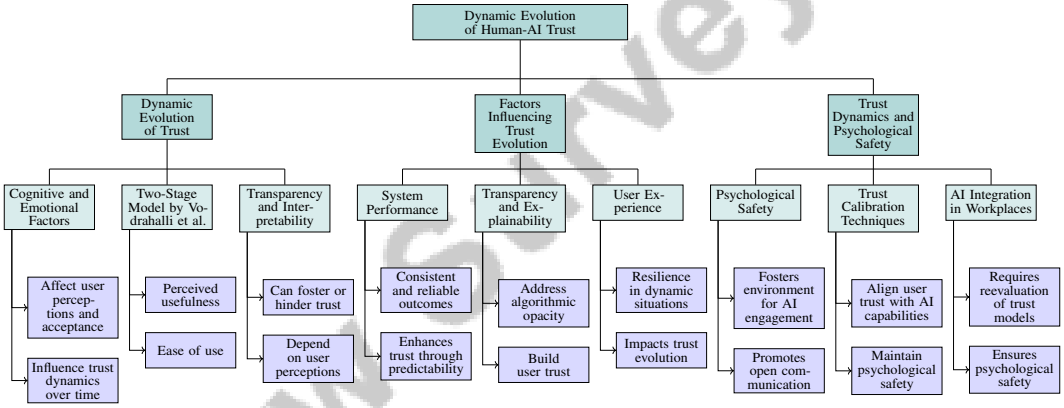


Figure 2: This figure illustrates the hierarchical structure of key concepts in the dynamic evolution of human-AI trust, highlighting the primary categories of trust evolution, factors influencing trust, and the relationship between trust dynamics and psychological safety. Each category is further broken down into subcategories, emphasizing cognitive and emotional factors, system performance, transparency, user experience, and the role of psychological safety in AI integration.

3 Dynamic Evolution of Human-AI Trust

3.1 Dynamic Evolution of Trust

Trust in AI systems is dynamic, shaped by cognitive and emotional factors that affect user perceptions and acceptance [18]. It evolves over time through interactions, necessitating an understanding of its fluctuations across different contexts [13]. In long-term engagements, trust can shift based on user experiences and system performance. The two-stage model by Vodrahalli et al. details how individuals decide to trust AI advice, emphasizing perceived usefulness and ease of use as critical for AI acceptance [10, 18]. The opaqueness of AI systems complicates trust dynamics, as transparency and interpretability can either foster or hinder trust depending on user perceptions [4]. Trust tends to increase when AI is perceived as more impartial and reliable than human counterparts, driven by shared values [6, 5]. Trust scores provide a quantitative measure of prediction reliability, adapting to complex decision-making contexts [8]. In workplace environments, aligning user trust with the

capabilities and limitations of AI systems is crucial for maintaining psychological safety and fostering collaborative relationships. As AI technologies advance, managing trust's dynamic nature is vital for their successful integration and acceptance.

3.2 Factors Influencing Trust Evolution

Trust evolution in AI systems is influenced by several factors, with system performance being paramount. Consistent and reliable outcomes foster trust, as users are more likely to rely on predictable systems [7]. AI systems that clearly explain their decisions enhance user understanding and trust, addressing concerns about algorithmic opacity. Transparency is critical, as the black-box nature of AI and potential biases can lead to distrust and hinder adoption [2]. Mechanisms like trust calibration cues and explainable AI improve transparency, building user trust. Trust scores, offering quantitative reliability measures, exemplify efforts to enhance transparency [8]. User experience also impacts trust evolution; systems demonstrating resilience in dynamic situations are more trusted [11]. Prior beliefs about AI and human performance, along with value similarity between users and AI, significantly influence trust and advice utilization. The unique interdependencies and communication needs in human-AI collaborations necessitate reevaluating traditional trust models, as these interactions often diverge from established norms [7]. Addressing these factors is essential for managing trust evolution and ensuring AI systems are seen as trustworthy partners.

3.3 Trust Dynamics and Psychological Safety

The relationship between trust dynamics and psychological safety is crucial for AI integration in workplaces. Psychological safety, the belief that the team is safe for interpersonal risk-taking, fosters an environment where employees can engage with AI systems and express concerns without fear of retribution [9]. This environment is essential for effective human-AI collaboration, promoting open communication and information sharing necessary for building trust. Trust dynamics affect psychological safety by influencing perceptions of AI reliability and predictability. As trust evolves, it can either enhance or undermine psychological safety, depending on whether AI systems meet user expectations and demonstrate transparency [15]. Providing clear, understandable explanations for AI decisions is key to maintaining psychological safety, enabling users to feel secure in their interactions [11]. Integrating AI in workplaces requires reevaluating traditional trust models, as human-AI interdependencies demand new approaches to managing trust and ensuring psychological safety [7]. Developing trust calibration techniques that align user trust with AI capabilities is crucial for maintaining psychological safety and fostering trusting relationships [6].

4 Impact on Employees' Wellbeing

4.1 Trust in Team Dynamics

Establishing trust in AI systems is crucial for effective team collaboration, particularly in organizations increasingly adopting AI technologies. This trust depends on addressing user concerns regarding privacy, accountability, and transparency, which are vital for user acceptance and collaboration [12]. Aligning AI behavior with user expectations is essential for positioning AI as a reliable team partner.

Trust in AI systems varies across sectors such as healthcare, finance, and autonomous vehicles, where the stages of trust development significantly influence team dynamics [13, 2]. The source of advice, whether from AI or human agents, also impacts decision-making and collaborative processes within teams [10]. Transparency is key to fostering trust dynamics, especially in environments with human-AI-human interactions [4]. Trust calibration cues and clear explanations for AI decisions enhance user understanding and facilitate collaboration [18]. Advances in explainable AI further support user trust and comprehension, positively affecting team dynamics [11].

Factors such as perceived rapport, enjoyment, peer influence, facilitating conditions, and self-efficacy contribute positively to trust in AI teammates, enhancing overall team dynamics [22]. However, perceived trustworthiness and affective interpersonal trust are often lower for AI compared to human counterparts [20], underscoring the need for ongoing efforts to cultivate and sustain trust in AI systems.

Trust in AI significantly shapes team interactions and engagement, altering collaboration structures. Key factors include perceived trustworthiness, emotional rapport, and socio-technical aspects of human-AI configurations, which are vital for fostering effective collaboration. Trust in AI affects individual attitudes and perceptions, influencing overall team trust and psychological safety, critical for the success of AI-enhanced teamwork [20, 22, 18, 9, 19]. By addressing these trust-related factors, organizations can enhance collaboration and maximize the benefits of AI integration in team settings.

4.2 Social Consequences and Public Apprehension

AI integration brings significant social consequences and public apprehensions, particularly concerning trust dynamics. Distrust in AI systems can disrupt public trust cycles, hindering broader acceptance and integration [23]. This skepticism is intensified by concerns over AI's role in surveillance and control, emphasizing the need for ethical considerations often neglected in existing frameworks [9].

In news production, AI can enhance operational efficiency, but ethical considerations must be prioritized to maintain public trust [19]. Deploying AI in sensitive areas requires understanding ethical implications and developing frameworks that address the multi-faceted nature of trust, ensuring AI systems are perceived as trustworthy.

Public apprehension towards AI is exacerbated by a lack of transparency and accountability in decision-making processes, leading to skepticism and resistance. Improving transparency and explainability of AI systems is essential for fostering a trusting relationship between AI technologies and society. Transparency significantly influences consumer trust, especially in sectors like healthcare and news media, where AI algorithms often create a 'black box' scenario. Implementing explainable AI methodologies and establishing robust frameworks for evaluating AI behavior can ensure AI systems are reliable and aligned with ethical and regulatory standards, promoting informed public engagement and confidence in AI applications [4, 11, 24, 15, 19]. Addressing these social impacts and public concerns is vital for fostering a more trustful and ethically responsible integration of AI technologies.

5 Trust Measurement Techniques

Category	Feature	Method
Model-Based Trust Quantification	Interactive Engagement	CS[25]

Table 1: This table presents a summary of methodologies employed in model-based trust quantification within human-AI interactions. It highlights the integration of interactive engagement techniques and computational methods that facilitate the measurement of trust dynamics. These approaches are instrumental in understanding the psychological and technological factors that influence user trust in AI systems.

Exploring trust measurement techniques unveils a range of methodologies for assessing trust in human-AI interactions. Table 1 provides an overview of the methodologies used in model-based trust quantification, emphasizing the role of interactive engagement in measuring trust within human-AI interactions. Additionally, Table 2 offers a detailed comparison of various methodologies for measuring trust in human-AI interactions, emphasizing their unique features and application contexts. This section delves into traditional and innovative approaches, capturing the complexities of user trust in AI systems, and highlights their significance in understanding user experiences and expectations.

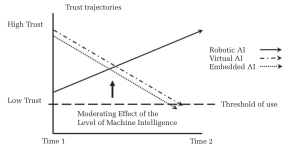
5.1 Measurement and Methodologies

Trust measurement in human-AI interactions employs both traditional and innovative methodologies. Traditional approaches often involve quantitative measures like surveys and questionnaires to gauge user perceptions and satisfaction with AI systems [10]. These tools offer structured insights into trust-related constructs, enhancing understanding of user experiences with AI technologies.

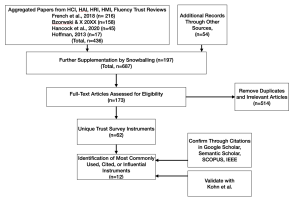
Recent developments have introduced dynamic methodologies that accommodate AI's evolving capabilities and the fluid nature of trust [2]. Adaptive Trust Calibration (ATC) exemplifies this innovation by adjusting trust levels based on contextual factors, aligning user expectations with AI

performance [15]. The integration of Explainable Artificial Intelligence (XAI) methodologies further enhances trust measurement by providing understandable explanations for AI decisions, addressing transparency and interpretability concerns [11, 4]. Scoring systems for behavior certificates evaluate correctness, relevance, and understandability, offering insights into both traditional and innovative trust measurement methodologies.

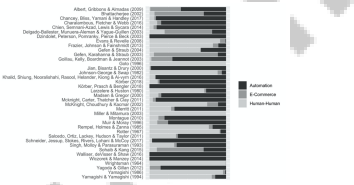
Innovative models derive trust scores by combining direct observations, recommendations, and social similarities, reflecting the social dynamics influencing user perceptions of AI [5]. Manipulationist approaches assess trust by measuring levels before and after altering AI trustworthiness, providing nuanced insights into trust dynamics [15]. Despite these advancements, current research faces limitations, including a lack of longitudinal studies and methodological diversity [2]. Addressing these gaps is crucial for developing robust methodologies that accurately capture the multifaceted nature of trust in human-AI interactions.



(a) Trust Trajectories: The Moderating Effect of Machine Intelligence[17]



(b) The image depicts a flowchart illustrating the process of identifying unique trust survey instruments through a systematic review.[21]



(c) Research on Automation, E-Commerce, and Human-Human Interaction[25]

Figure 3: Examples of Measurement and Methodologies

As illustrated in Figure 3, these visual examples provide insights into the diverse methodologies and approaches utilized in contemporary trust measurement research. The first image delineates the evolution of trust levels over time concerning different AI modalities, highlighting trust's dynamic nature. The second image systematically depicts the process of identifying unique trust survey instruments through a systematic review, aggregating insights from various fields. The third image presents a comparative analysis of research frequencies on automation, e-commerce, and human-human interaction, illustrating the breadth of methodologies employed in trust measurement [17, 21, 25].

5.2 Behavioral and Physiological Measures

Behavioral and physiological measures offer innovative approaches for understanding and quantifying trust in human-AI interactions, capturing real-time data that reflect users' implicit trust responses. Behavioral metrics, such as response time, task performance, and interaction patterns, provide insights into user engagement and reliance on AI technologies [10]. These metrics are valuable for assessing evolving trust dynamics during interactions [15].

Physiological measures, including heart rate variability, galvanic skin response, and eye-tracking, deepen the understanding of users' emotional and cognitive states during AI interactions [4]. These measures capture subtle physiological changes, providing a comprehensive view of trust as it relates to stress, anxiety, and cognitive load. Integrating physiological data with behavioral observations enables a multi-dimensional assessment of trust [11].

These measures are particularly relevant in high-stakes environments, such as healthcare and autonomous systems, where trust is critical for effective decision-making and collaboration [5]. By leveraging behavioral and physiological data, researchers can develop accurate trust models that account for the interplay between cognitive and emotional factors. Despite their potential, the application of these measures in trust research faces challenges, including sophisticated data analysis techniques and ethical considerations surrounding user monitoring. Addressing these challenges is essential for advancing the field and developing trustworthy AI systems [13, 2, 19, 11].

5.3 Model-Based Trust Quantification

Model-based approaches for quantifying trust in human-AI interactions utilize structured frameworks that systematically evaluate and measure trust dynamics, addressing the interplay of psychological, sociocultural, and technological factors influencing user perceptions. These approaches leverage empirical data and advanced methodologies, such as structural equation modeling, to assess trustworthiness, which is increasingly essential in contexts where AI systems are integrated into decision-making processes [13, 25, 6, 4]. By establishing standardized procedures for measuring trust, these frameworks elucidate the conditions under which users can confidently rely on AI.

An innovative method employs dynamic conversational techniques to measure trust through interactive prompts, facilitating personal expression and reflection [25]. This captures nuanced trust perceptions by engaging users in dialogue about their experiences and expectations of AI systems. Model-based trust quantification is further enhanced by machine learning algorithms that analyze user interactions and feedback, predicting trust levels through assessments of agreement between classifier outputs and alternative models [8, 4]. This adaptability is crucial as AI systems evolve, ensuring trust assessments remain relevant and accurate.

Trust scores derived from model-based approaches offer quantitative measures easily communicated and interpreted. These scores serve as standardized benchmarks for evaluating trust levels across different applications, enabling organizations to compare trustworthiness effectively and drive continuous improvement initiatives in human-AI collaboration [18, 9, 8, 6]. By quantifying trust consistently, organizations can better understand AI's impact on user trust and make informed decisions about system design and deployment.

Model-based trust quantification emphasizes evaluating AI systems based on behavioral evidence and reliability rather than solely on interpretability. This method leverages behavior certificates to predict future model performance, addressing fundamental trustworthiness issues in AI technologies. By focusing on mechanisms that foster trust, such as perceived impartiality and functionality, this approach clarifies trust's role in AI acceptance and offers a nuanced framework for improving AI's societal integration and guiding ethical development [18, 3, 4, 6]. Combining dynamic conversational methods with computational models and trust scores provides a robust framework for evaluating and enhancing trust in AI systems.

Feature	Measurement and Methodologies	Behavioral and Physiological Measures	Model-Based Trust Quantification
Measurement Approach	Surveys And Questionnaires	Behavioral And Physiological	Structured Frameworks
Data Type	Quantitative	Real-time Data	Empirical Data
Application Context	General AI Interactions	High-stakes Environments	Decision-making Processes

Table 2: Table ef presents a comparative analysis of three distinct trust measurement methodologies in human-AI interactions: surveys and questionnaires, behavioral and physiological measures, and model-based trust quantification. It highlights the diverse data types, measurement approaches, and application contexts relevant to each methodology, providing insights into their respective strengths and applicability in different settings.

6 Intervention Strategies to Enhance Trust

6.1 Training and Personality Integration

Integrating personality traits into AI systems, especially large language models (LLMs), significantly boosts user trust by enhancing AI response consistency and adaptability, as demonstrated by the PersLLM method [26]. This integration not only improves perceived reliability but also aligns AI behavior with user expectations, creating a personalized experience [12]. Designing AI teammates to emphasize rapport and enjoyment over anthropomorphism is crucial for increasing user trust and satisfaction [22]. Training users to effectively collaborate with AI teammates further enhances familiarity and positive experiences, reinforcing trust [20]. Incorporating trust calibration cues (TCCs) within training programs helps users understand AI capabilities and limitations, facilitating accurate trust calibration [27, 16]. A human-in-the-loop approach is vital for improving understanding and trust in AI systems, ensuring that human oversight enhances transparency and reliability [11]. Active trust management frameworks that adjust based on real-time feedback are essential for maintaining trust in dynamic human-AI interactions [7].

6.2 Transparent Communication and Explainability

Transparent communication and explainability are fundamental to fostering trust in AI systems by enabling users to comprehend and accept AI technologies. The ExClaim framework underscores the importance of providing rationales for AI predictions, particularly for non-expert users, enhancing trust through clear explanations [28]. This aligns with the broader emphasis on explainable AI (XAI), essential for building trust by elucidating AI systems' processes and decisions. Transparency is particularly crucial in psycho-counseling, where disclosing AI identity is vital for ethical standards and trust [29]. Ongoing research should address the ethical implications of AI transparency and reliability, as these factors are crucial for sustaining trust in human-AI interactions [6]. A theoretical framework synthesizing existing literature supports the model of transparent communication, detailing how trust is established and maintained in human-AI interactions [3]. This framework highlights the necessity for AI systems to communicate their processes and limitations transparently, allowing users to achieve a pragmatic understanding of AI operations [1]. However, the effectiveness of trust scores in enhancing transparency may diminish in high-dimensional spaces, highlighting the need for ongoing refinement of trust measurement techniques [8].

6.3 Human-in-the-Loop and Feedback Mechanisms

Human-in-the-loop (HITL) approaches and feedback mechanisms are pivotal for enhancing trust in AI systems by integrating human oversight into AI decision-making processes. These methodologies facilitate interactive engagement, allowing users to provide feedback that refines system performance and aligns AI behavior with user expectations [21]. Feedback mechanisms are crucial for trust enhancement, enabling continuous improvement of AI systems based on user input. They support adaptive learning processes, allowing AI systems to adjust operations in response to feedback, ensuring responsiveness to user needs [13]. HITL approaches deepen the understanding of trust dynamics by exploring both positive and negative aspects of AI integration in teams, helping AI systems navigate complex social environments and mitigate potential negative impacts, such as trust erosion due to perceived threats or biases [9]. Future research should prioritize refining trust measurement instruments, exploring trust dynamics over time, and addressing current study limitations. By developing sophisticated trust metrics and adaptable frameworks, researchers can better navigate the intricate dynamics of trust in human-AI interactions, enhancing the design of reliable AI systems centered around user needs [4, 13, 18, 6, 9].

7 Integration of Workplace Technology

7.1 AI Representations and Trust Dimensions

AI representations significantly influence trust dimensions within workplace technology, shaping user perceptions and interactions. Design elements such as anthropomorphism and perceived intelligence are crucial in fostering trust and acceptance [22]. The PersLLM approach integrates personality traits into AI, aligning behavior with user expectations to enhance relatability and engagement [26]. Transparent AI representations providing understandable explanations are vital for building cognitive trust, enabling users to assess reliability [28]. Ethical considerations are equally important, addressing user concerns regarding AI implications and promoting informed user relationships [29, 6]. In team settings, AI's perceived rapport and enjoyment can enhance trust and collaboration, highlighting the need for socially attuned, technically proficient AI systems [22]. Challenges in trust calibration arise in high-dimensional spaces, where geometric relationships among data points may lose informativeness [8].

7.2 Role of LLMs and AI Decision Support Systems

Large Language Models (LLMs) and AI Decision Support Systems are increasingly influential in shaping workplace trust dynamics. LLMs, especially those developed through the PersLLM method, enhance trust by facilitating consistent, adaptable interactions that meet user expectations [26]. This personalization improves communication and understanding between AI systems and users. AI Decision Support Systems, utilizing LLM capabilities, enhance decision-making processes by providing accurate, timely information, empowering informed decisions and increasing trust [18]. The integration of LLMs into these systems enables effective processing and analysis of large

datasets, essential for decision-making in complex environments. Transparency and explainability are critical, as users trust systems offering clear rationales for decisions, as emphasized by the ExClaim framework [28, 1]. Ethical considerations in LLM and AI Decision Support Systems design and deployment are vital for sustaining trust, addressing user concerns about AI's role and impact [29, 6]. Ensuring AI systems are technically competent and socially attuned is essential for successful workplace integration and acceptance.

8 Conclusion

8.1 Challenges and Future Directions

Human-AI trust remains a multifaceted challenge, particularly in understanding the intricate dynamics of trust and the seamless integration of AI in diverse environments. One pressing issue is the development of AI systems capable of learning adaptively from errors and incorporating multifaceted trust signals to bolster reliability and user acceptance. The demand for standardized measures that accurately capture the quality of AI explanations and their impact on user trust is acute, especially in critical sectors like healthcare and criminal justice.

Future research must focus on refining methodologies for trust evaluation, emphasizing the interaction between trust and user characteristics. Enhancing trust assessments through context awareness and knowledge graph embeddings can increase accuracy and adaptability. Furthermore, advancing AI assurance techniques to boost model trustworthiness and achieving explainability in AI systems are pivotal areas for development. Expanding datasets to cover a wider range of scenarios, including psycho-counseling, will enrich the evaluation of AI responses and their ethical dimensions.

The implications of emerging trends in AI transparency for consumer engagement and societal impacts warrant further exploration, as these are crucial for establishing and maintaining trust. Investigating adaptive trust calibration methods across various contexts and identifying social factors that influence trust are essential for constructing robust trust models. Additionally, enhancing the quality of AI explanations and developing personalized explanation interfaces to improve user comprehension and trust are critical future research directions. By addressing these challenges and focusing on these areas, the field can advance towards more trustworthy and effectively integrated AI systems across multiple domains. Future research should also explore the applicability of these findings in other human-AI collaboration contexts and examine user behavior in adopting AI technologies.

References

[1] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257, 2023.

[2] Saleh Afroogh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. Trust in ai: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11(1):1–30, 2024.

[3] Andrea Ferrario, Michele Loi, and Eleonora Viganò. In ai we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology*, 33(3):523–539, 2020.

[4] Max W. Shen. Trust in ai: Interpretability is not necessary or sufficient, while black-box interaction is necessary and sufficient, 2022.

[5] Siddharth Mehrotra. Modelling trust in human-ai interaction. In *AAMAS*, pages 1826–1828, 2021.

[6] Michael Gerlich. Exploring motivators for trust in the dichotomy of human—ai trust dynamics. *Social Sciences*, 13(5):251, 2024.

[7] Melanie J McGrath, Andreas Duenser, Justine Lacey, and Cecile Paris. Collaborative human-ai trust (chai-t): A process framework for active management of trust in human-ai collaboration. *arXiv preprint arXiv:2404.01615*, 2024.

[8] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. *Advances in neural information processing systems*, 31, 2018.

[9] Essi Janhunen, Tuuli Toivikko, Kirsimarja Blomqvist, and Dominik Siemon. Trust in digital human-ai team collaboration: A systematic review. 2024.

[10] Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 763–777, 2022.

[11] Andreas Holzinger. The next frontier: Ai we can really trust. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 427–440. Springer, 2021.

[12] Yugang Li, Baizhou Wu, Yuqi Huang, and Shenghua Luan. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-ai trust. *Frontiers in Psychology*, 15:1382693, 2024.

[13] Fatemeh Alizadeh, Oleksandra Vereschak, Dominik Pins, Gunnar Stevens, Gilles Bailly, and Baptiste Caramiaux. Building appropriate trust in human-ai interactions. In *20th European Conference on Computer-Supported Cooperative Work (ECSCW 2022)*, volume 6, 2022.

[14] New work: New motivation? a comp.

[15] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 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*, pages 624–635, 2021.

[16] Moritz Körber, Eva Baseler, and Klaus Bengler. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied ergonomics*, 66:18–31, 2018.

[17] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020.

[18] Hyesun Choung, Prabu David, and Arun Ross. Trust in ai and its role in the acceptance of ai technologies. *International Journal of Human–Computer Interaction*, 39(9):1727–1739, 2023.

[19] Justyna Żywiołek. Building trust in ai-human partnerships: Exploring preferences and influences in the manufacturing industry. *Management Systems in Production Engineering*, 32(2), 2024.

[20] Eleni Georganta and Anna-Sophie Ulfert. My colleague is an ai! trust differences between ai and human teammates. *Team Performance Management: An International Journal*, 30(1/2):23–37, 2024.

[21] Yosef S. Razin and Karen M. Feigh. Converging measures and an emergent model: A meta-analysis of human-automation trust questionnaires, 2023.

[22] Keke Hou, Tingting Hou, and Lili Cai. Exploring trust in human–ai collaboration in the context of multiplayer online games. *Systems*, 11(5):217, 2023.

[23] Bran Knowles, Jason D'Cruz, John T. Richards, and Kush R. Varshney. Humble machines: Attending to the underappreciated costs of misplaced distrust, 2022.

[24] Christian Meske and Enrico Bunde. Transparency and trust in human-ai-interaction: The role of model-agnostic explanations in computer vision-based decision support. In *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, pages 54–69. Springer, 2020.

[25] Mengyao Li, Areen Alsaid, Sofia I. Noejovich, Ernest V. Cross, and John D. Lee. Towards a conversational measure of trust, 2020.

[26] Zheni Zeng, Jiayi Chen, Huimin Chen, Yukun Yan, Yuxuan Chen, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. Persllm: A personified training approach for large language models, 2024.

[27] Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-ai collaboration. *Plos one*, 15(2):e0229132, 2020.

[28] Sai Gurrapu, Lifu Huang, and Feras A. Batarseh. Exclaim: Explainable neural claim verification using rationalization, 2023.

[29] Lizhi Ma, Tong Zhao, Huachuan Qiu, and Zhenzhong Lan. No general code of ethics for all: Ethical considerations in human-bot psycho-counseling, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.