Bidirectional Human-Machine Trust in Emerging Human-Machine Relationships

Authors: Yubin Xie, Zhou Ronggang, Zhou Ronggang

Date: 2025-02-07T00:00:00+00:00

Abstract

With the rapid development of artificial intelligence technology, the frequency of human-machine interaction continues to increase, and interaction patterns are becoming increasingly complex. Traditional human-machine trust models have primarily focused on unidirectional trust, namely human trust in machines. However, as intelligent systems gradually acquire autonomy and decision-making capabilities, the bidirectionality of human-machine trust has emerged as a core research issue. Building upon a review of recent theoretical models of human-machine trust, this study proposes a theoretical structural model of bidirectional human-machine trust based on dispositional trust, perceived trust, and behavioral trust, particularly emphasizing the critical role of "perceived trust" as an interactive channel for mutual trust between humans and machines. Furthermore, this paper systematically reviews the latest advances in measurement and computational modeling methods for human-machine trust, focusing on methods for measuring machine trust in humans and their practical implications, and proposes future research directions, aiming to provide new perspectives and a guiding framework for theoretical development and technological application in the field of human-machine collaboration.

Full Text

The Bidirectional Trust in the Context of New Human-Machine Relationships

XIE Yubin[1,2], **ZHOU Ronggang**[1,3,4]
[1] School of Economics and Management, Beihang University, Beijing 100191, China
[2] Department of Systems Engineering, City University of Hong Kong, Hong Kong 999077, China
[3] Key Laboratory of Data Intelligence and Management, Beihang University, Beijing 100191, China

[4] Laboratory for Low-carbon Intelligent Governance, Beihang University, Beijing 100191, China

Abstract

With the rapid advancement of artificial intelligence technology, the frequency and complexity of human-machine interactions have increased dramatically. Traditional human-machine trust models have primarily focused on unidirectional trust—namely, human trust in machines. However, as intelligent systems gradually acquire autonomy and decision-making capabilities, the bidirectional nature of trust has emerged as a central research topic. This study reviews recent theoretical models of human-machine trust and proposes a theoretical framework for bidirectional trust based on dispositional trust, perceived trust, and behavioral trust, with particular emphasis on the critical role of perceived trust as an interactive channel for mutual trust. Furthermore, this paper systematically examines the latest developments in trust measurement and computational modeling methods, focusing specifically on approaches for measuring machine trust in humans and their practical implications. Finally, we identify future research directions to provide new perspectives and a guiding framework for theoretical development and technological applications in human-machine collaboration.

Keywords: artificial intelligence, human-machine mutual trust, trust, trust measurement, human-machine teams

1. Introduction

Effective teamwork depends on trust [**?**, **?**], and mutual trust between humans and intelligent machines is similarly regarded as foundational for successful collaboration [**?**, **?**]. Research has shown that inappropriate or insufficient trust negatively impacts team performance [**?**, **?**]. For instance, misplaced trust can deteriorate human-machine relationships and harm team efficiency [**?**]. As AI and intelligent technologies become widely deployed in production and daily life, researchers increasingly recognize that AI' s role in human-machine collaboration is strengthening, and the relationship is evolving from traditional "assistant-subordinate" dynamics toward "equal partnership" and even "symbiotic integration" [**?**, **?**]. This transformation has shifted human-machine relationships from unidirectional to bidirectional [**?**, **?**]. Given this bidirectional nature, cultivating mutual trust between humans and AI is essential. Over the past decade, extensive research has examined human trust in AI and machines [**?**, **?**, **?**, **?**], while discussions of AI/machine trust in humans remain relatively limited.

AI trust in humans is grounded in the shared mental model hypothesis [**?**, **?**]. The Explainable Artificial Intelligence (XAI) program launched by DARPA in 2015 proposed a psychological model of AI explanation, emphasizing the importance of consistency between AI and human minds [**?**]. Murphy [**?**] similarly

noted that investigating how robots intelligently infer human beliefs, desires, and goals to take appropriate actions is key to shaping explainable AI. Robots must establish mental models to communicate and cooperate effectively with humans, thereby earning human trust. Consequently, robots should not only understand human trust patterns but also convey trust in ways that align with human psychological perception, which is considered crucial for achieving psychological alignment between AI and humans. This demonstrates that trust, as a psychological variable, plays a vital role in human-AI alignment. Qi Yue et al. [**?**] were among the first to address the transformation of human-machine trust relationships in the AI era, emphasizing the importance of AI trust in humans. Drawing on interpersonal trust models and human-AI trust frameworks, they proposed that AI can assess its trust level in humans by actively perceiving its own and users' states, thereby determining control allocation. This study provided the first systematic definition of AI trust in humans, laying a theoretical foundation for exploring human-machine trust mechanisms. However, the discussion of human perception of AI trust and the transmission and measurement of bidirectional trust remains incomplete.

Exploring machine trust in humans, particularly human perception of machine trust, is crucial for human-machine interaction and technological applications. Although we cannot yet determine whether AI truly "trusts" humans, AI can assess individual trustworthiness and adjust its behavior accordingly, thereby transmitting trust signals. As technology advances, machines have surpassed novice employees in certain tasks [**?**]. In the new paradigm of equal human-machine partnership, critical interaction design questions arise: When machines identify human capability deficiencies, should they proactively offer advice or intervene, and how deeply should they engage? For example, automotive active safety systems often forcibly override human operations when triggered. While this significantly improves efficiency and safety in some scenarios, one of the three laws of robotics emphasizes that "robots must obey human commands," preserving humans' ultimate decision-making authority at ethical and legal levels, including control over AI system "on/off" switches and authorization [**?**, **?**]. Human trust in AI determines usage intentions, which is typically closely related to system performance [**?**]. Conversely, AI trust in humans determines whether to delegate tasks to humans, depending on whether human capabilities and performance meet system expectations. In human-machine interaction, humans' subjective perception of AI trust influences whether they exercise ultimate control such as "on/off" switches. Similar to human teams, when one party does not feel trusted by the other, uncertainty and insecurity increase, leading to defensive behavior and reduced interaction [**?**]. Likewise, when humans do not perceive trust from AI, they often question the system' s reliability and cooperative intentions, reducing dependence and acceptance. Thus, the absence of bidirectional trust not only weakens cooperative relationships but may also hinder technology adoption. Human perception of AI trust is not only a core issue in technical design but also a key element for promoting long-term AI development.

Based on this understanding, our literature review first focuses on how machine trust behaviors convey the perception of "being trusted" to humans, arguing that this perception constitutes the interactive channel for bidirectional trust. We propose a theoretical model of human-machine bidirectional trust based on human disposition, perception, and behavior. Second, trust measurement is a core issue in current human-machine trust research. Although existing literature has addressed machine trust to some extent, a clear framework for its creation and measurement has not yet emerged, lacking systematic review and synthesis. Therefore, our second focus examines measurement methods for human-machine mutual trust, particularly exploring approaches for machine trust in humans across dispositional, perceived, behavioral, and integrated computational dimensions. These aspects collectively construct our complete research framework on human-machine bidirectional trust. To systematically address these questions, we first collected research papers related to human-machine trust; specific collection and screening criteria are detailed in Table 1 . This paper subsequently discusses: (1) the evolution and current challenges of human-machine trust; (2) bidirectional trust models and their theoretical foundations; (3) measurement and modeling methods for human-machine bidirectional trust; (4) relevant application case analyses; and (5) future research directions. Based on this, we propose a comprehensive theoretical and applied model of human-machine mutual trust that integrates human trust psychological structures, human-machine relationships, interaction behaviors, and bidirectional trust transmission mechanisms, and we look ahead to future research.

2.1 The Evolution of Human-Machine Trust

Trust is a multidimensional concept widely applied in psychology, sociology, economics, and human-computer interaction [**?**, **?**, **?**]. Initially, human-machine trust focused primarily on human trust in machines: Muir [**?**] examined the development and erosion of human trust in machines within automated systems, while Lee and See [**?**] defined trust in automation as an attitude toward collaborating with an agent under uncertainty, sparking interest in how affect, attitudes, and situational factors influence trust [**?**, **?**, **?**, **?**]. This phase witnessed empirical validation by numerous scholars [**?**]. Subsequent research further concentrated on applications in autonomous vehicles [**?**, **?**, **?**], AI agents [**?**, **?**], and spacecraft [**?**], as well as trust calibration and repair in dynamic environments [**?**, **?**].

As machines have gained autonomous learning and interaction capabilities, trust has shifted from unidirectional dependence to bidirectional mutual trust [**?**]. De Visser et al. [**?**] first emphasized the importance of machine trust in humans, while Azevedo-Sa et al. [**?**] proposed the concept of "artificial trust" and constructed a bidirectional trust model encompassing both human trust in machines and machine trust in humans. Jorge et al. [**?**, **?**, **?**] further expanded the connotation of AI trust, noting that it includes not only assessments of human capability and willingness but also factors such as perceived costs and benefits,

while defining the structure and measurement methods for machine trust in humans. Qi Yue et al. [?] identified key factors influencing AI trust in humans and constructed a human-AI mutual trust model. Overall, human-machine trust is evolving from static unidirectional to dynamic bidirectional models, with increasing importance placed on individual psychological and cognitive differences as machine intelligence advances. However, systematic exploration of mutual trust relationships and bidirectional trust transmission mechanisms remains needed.

2.2 New Human-Machine Relationships in the AI Era

Traditional instrumentalism views machines as subordinate tools, but as AI autonomy and complexity increase [?, ?], human-machine relationships are transforming from "tool-based" to "partner-based" and even "symbiotic" [?]. New human-machine relationships manifest in three ways: First, from assistance to collaboration—whereas traditional relationships involved AI assisting humans, AI now possesses autonomous perception, decision-making, and learning capabilities, enabling participation in complex task division and collaboration. Second, human-machine team structures—Sycara and Lewis [?] proposed three roles for machines in teams: supporting individual tasks, acting as equal members, or assisting the entire team. In new relationships, AI is also viewed as a proxy substitute for human teams, serving as a "synthetic human" [?]. Third, bidirectional interaction and trust—in these new relationships, core challenges for AI integration include user acceptance and adoption, and perceiving AI as teammates rather than tools [?]. Trust is considered the core psychological variable affecting these issues [?, ?, ?, ?]. The defining feature of this new relationship is bidirectional interaction, requiring researchers to further focus on the definition, measurement, and dynamic changes of bidirectional trust.

2.3 Challenges of Human-Machine Trust in New Relationships

Building trust in new human-machine relationships faces multiple challenges, particularly in AI-dependent collaborative scenarios where bidirectional trust definition and measurement are critical. Although existing research has explored the core connotation of human-machine trust from interpersonal perspectives [?, ?, ?, ?], four research gaps remain:

First, the interaction and transmission mechanisms of bidirectional trust lack psychological grounding. Current research primarily focuses on definitions and structural models, with insufficient exploration of the psychological mechanisms underlying bidirectional trust, especially the transmission processes in emotional, informational, and decision-making exchanges. For example, human perception of AI trust and their trust needs. This paper integrates interpersonal and human-machine trust models to propose a bidirectional trust model based on human perception of "being trusted," with in-depth analysis of psychological transmission mechanisms.

Second, there is a lack of understanding regarding individual characteristic differences. In human trust in machines, differences in emotion, attitude, and personality significantly affect trust outcomes. Individual trust propensity and algorithm aversion may become potential barriers to trust establishment. This paper incorporates these traits into the analysis, focusing on their roles in trust interaction and transmission.

Third, measurement methods have primarily focused on questionnaires, psychological perception, and behavioral observation, with mature measures for human trust in machines but underdeveloped dimensions and methods for measuring machine trust in humans. This paper integrates existing methods to construct a measurement system applicable to bidirectional trust.

Fourth, research on influencing factors and effects of bidirectional trust is insufficient. Current exploration of bidirectional trust's antecedents and consequences, particularly at experimental and empirical levels, remains limited. This paper analyzes key variables and mechanisms of bidirectional trust to address these gaps and improve theoretical and practical frameworks.

3. A Bidirectional Trust Model Based on Trust Perception

Following the developmental trajectory of trust theory models and systematically reviewing different trust development stages, this study proposes a bidirectional trust model based on trust perception, integrating interpersonal team trust theory and human-machine mutual trust frameworks. This model aims to clarify the development process and influencing mechanisms of human-machine bidirectional trust, providing theoretical support and practical guidance for building more efficient and reliable human-machine collaboration systems.

3.1 Framework Elements of Human-Machine Mutual Trust

Early trust research by Mayer et al. [**?**] constructed a three-factor model encompassing competence, benevolence, and integrity based on interpersonal trust structures. Subsequently, Lee and See [**?**] proposed three foundations for trust in automation: performance, process, and purpose. Earle and Siegrist [**?**] distinguished between relational trust and calculus trust, emphasizing relationship importance. Merritt and Ilgen [**?**] introduced dispositional trust, proposing trust as a continuum between dispositional and history-based trust. Hoff and Bashir [**?**] later constructed a three-layer human-machine trust model comprising dispositional, situational, and learned trust. Gao et al. [**?**] proposed a dynamic trust framework for autonomous vehicles, identifying four layers: dispositional, initial, real-time, and post-hoc trust. Recently, as AI autonomy and complexity have increased, researchers have begun examining machine trust in humans. Jorge et al. [**?**] adapted Mayer et al.' s framework, dividing human credibility into three dimensions: competence (task success), benevolence (willingness to selflessly help other agents), and integrity (demonstrating truthful, honest, and ethical behavior). Their theoretical foundation focused on building evaluation systems

for trustworthy humans. Qi Yue et al. [**?**] similarly drew on interpersonal trust models to propose a dynamic human-AI mutual trust model comprising initial, perception, and behavioral stages, emphasizing the importance of the perception stage and subdividing it into system state perception and user state perception.

Existing research typically divides trust into two layers: dispositional trust as an inherent trait independent of specific situations with high stability, and history-based trust generated through interaction processes, situationally influenced and dynamically adjusted. Researchers generally agree that trust is a dynamic developmental process encompassing initial, perceived, and post-hoc trust. However, while existing models often treat human-machine trust as a credibility concept focused on building evaluation systems for trustworthy humans and AI, individual behavior is also influenced by the perception of "being trusted" [**?**]. Current human-machine trust frameworks lack systematic theoretical models for interactive perception, particularly regarding mutual trust. Trust as a psychological characteristic lacks clear models for affective transmission between dispositional and behavioral trust. For example, the critical dimension of human perception of AI trust behavior remains underexplored. In human teams, research has demonstrated that trust operates only when humans can perceive others' trust behaviors [**?**, **?**], yet studies on human perception of AI behavioral trust remain scarce.

From this perspective, this study proposes a dynamically evolving three-stage human-machine mutual trust model divided into dispositional trust, perceived trust, and behavioral trust (Figure 1 [Figure 1: see original paper]). The model emphasizes perceived trust as the critical bridge between dispositional and behavioral trust, highlighting its transmission role between AI, agents, and humans. Dispositional trust represents the initial stage, originating from individual inherent traits independent of specific situations, laying the foundation for subsequent trust development. Perceived trust gradually forms during interaction, reflecting dynamic perception of the other' s behavior, attitude, and trust, serving as the core for affective trust transmission and dynamic adjustment. Behavioral trust represents the ultimate manifestation of trust, expressed through concrete reliance, cooperation, and actions—post-hoc trust based on behavioral feedback that reflects the final outcome of trust relationships. This model highlights the dynamic evolution of trust from dispositional to behavioral stages and reveals transmission mechanisms in bidirectional interaction, offering a new theoretical perspective and practical basis for building effective human-machine collaboration.

The model' s advantages include: (1) Dynamic evolution characteristics—comprehensively demonstrating trust' s developmental process from dispositional to perceived to behavioral trust, accommodating the complexity and variability of trust relationships in human-machine interaction; (2) Bidirectional trust transmission—focusing systematically on bidirectional interaction between humans and intelligent agents, emphasizing perceived trust' s bridging role and its significance in affective transmission and dynamic adjustment, providing unique

guidance for optimizing human-machine interaction; (3) Expanded perspective on dispositional trust—introducing an algorithmic trust perspective to explore algorithmic initial trust sources and individual algorithm aversion tendencies, offering new theoretical foundations for algorithmic trust research; and (4) In-depth behavioral trust analysis—highlighting machine behavior' s impact on human-machine trust, such as the negative impact on "perceived trust"when machines reject human requests, revealing emotional and behavioral consequences of trust misalignment. Overall, this model centers on dynamic, bidirectional, and affective transmission characteristics, comprehensively demonstrating the complex mechanisms of human-machine trust.

3.2 Dispositional Trust

In human-machine trust research, dispositional trust refers to users' initial trust level exhibited before actual interaction with technology or systems (such as AI or automated equipment), based on inherent trust propensity and cognitive traits. It does not depend on specific interaction contexts or experiences but is determined by personality characteristics, general attitudes, cultural backgrounds, and prior experiences. Research by Merritt and colleagues [**?**, **?**, **?**, **?**] indicates that dispositional trust is primarily influenced by four key factors—culture, age, gender, and personality—exhibiting universality and stability as the foundation of initial trust. In our trust model, dispositional trust serves as an important starting point for building human-machine trust, laying the groundwork for subsequent perceived and behavioral trust and driving the establishment and development of the entire trust relationship.

Algorithms are the core driver of AI development, and their importance in intelligent machines has become increasingly prominent as technology transitions from automation to AI. Previous human-machine trust research focused on user trust in automated equipment or AI products, with less attention to attitudes toward underlying algorithms. In the AI era, algorithms built on big data and machine learning demonstrate powerful capabilities but also entail unpredictability and unexplainability, increasing user comprehension difficulty and potentially shaking trust, particularly regarding decision-making opacity and fairness issues [**?**]. In academia, the behavior or behavioral tendency of users refusing to accept or use algorithmic recommendations and services is termed algorithm aversion [**?**, **?**, **?**]. Algorithm aversion has manifested across multiple domains (e.g., healthcare, e-commerce, autonomous driving, law) and correlates with individual factors such as gender, age, and Big Five personality traits [**?**]. Research shows that algorithm aversion negatively impacts human-machine trust [**?**, **?**] and significantly affects user adoption of algorithmic recommendations [**?**] and organizational relationships in teams [**?**, **?**]. Despite confirmed impacts, algorithm aversion has not been adequately considered in existing trust models.

Given algorithms' importance in the AI era and their role in trust processes, this study incorporates algorithm aversion into the human-machine bidirectional

trust theoretical model as a component of dispositional trust. We hypothesize that algorithm aversion tendency negatively correlates with trust propensity —that is, individuals with algorithm aversion exhibit lower initial trust levels. Meanwhile, algorithm aversion' s influence mechanisms on perceived and behavioral trust may mirror but oppose those of trust propensity. By incorporating algorithm aversion, we aim to more comprehensively understand human-machine trust formation mechanisms and reveal trust evolution processes. However, the specific mechanisms through which algorithm aversion affects perceived and behavioral trust and its long-term impacts require further empirical research, which will provide valuable theoretical support for algorithm design and human-machine interaction optimization.

Regarding intelligent machines' dispositional trust in humans, Qi Yue et al. [**?**] proposed that intelligent machines' trust propensity in humans is determined by trust experience and disposition during the initial AI trust stage, primarily derived from system designers' trust assumptions about users. Our model expands this view, suggesting that intelligent machines' trust propensity is influenced not only by designers but may also gradually exhibit human-like trust patterns as intelligence levels increase. Specifically, intelligent machines' initial trust may reflect designers' tendencies, but through increased interaction, machines adjust trust propensity levels based on user behavior and feedback. For example, machines may revise trust in users based on behavioral patterns and cooperative willingness. Johnson and Obradovich' s [**?**] experiments found that intelligent systems (such as ChatGPT) exhibit trust propensity toward human populations, indicating that intelligent machines not only inherit designers' trust settings but may also autonomously adjust trust. We argue that intelligent machines' dispositional trust in humans, similar to human trust, is primarily shaped by designers' trust logic, machine learning experience, and human historical behavior. Intelligent machines gradually optimize trust propensity by learning user behaviors and qualities, offering a new research perspective on the dynamic evolution of human-machine bidirectional trust.

3.3 Perceived Trust

Perceived trust represents an important manifestation of bidirectional trust and a critical interactive channel for human-machine mutual trust. In interpersonal team research, perceived trust is considered a key factor for team cooperation and effectiveness. Perceived trust refers to the subjective evaluation by the trusted party of the trust conveyed by the trustor. It is influenced not only by the trust level exhibited by the trustor but also by the trusted party' s acceptance and response to that trust. In human teams, trust transmission occurs primarily through mutual perception of trust [**?**]. The feeling of being trusted is a widely mentioned form of trust in human teams [**?**, **?**]. Baer et al. [**?**] assessed employees' confidence and their perception of supervisors' willingness to accept their vulnerabilities as indicators of feeling trusted, finding that employees' sense of being trusted correlates closely with supervisor support and

acceptance. Gillespie [?] measured employees' perception of supervisor trust by asking whether supervisors were willing to rely on them at work and share personal views and sensitive information. Lau and Lam [?] found that supervisors' trust in employees, when perceived by employees, had stronger effects on performance and attitudes than employees' trust in supervisors. Individuals also respond to each other based on perceived trust [?]. These studies demonstrate that trust perception possesses sound psychometric properties in human organizations and serves as an important predictor of individual behavior and cooperative willingness in team environments [?, ?, ?].

Drawing on interpersonal trust transmission theory, we treat perceived trust as an important transmission channel in our human-machine bidirectional trust model. Regardless of whether AI possesses emotional trust, it makes judgments about humans based on specific algorithms and takes actions that transmit trust-like signals. Perceived trust includes two main dimensions: (1) Perception of the other' s state and behavior by humans/intelligent machines. For example, human users perceive AI reliability and effectiveness through its performance, response speed, and decision transparency. Similarly, intelligent machines perceive human reliability and consistency by observing human behavior (e.g., decision patterns, response speed, cooperative intent). (2) Perception of the other' s trust by humans/intelligent machines. For instance, human users perceive whether machines trust them through decision-making methods and interactive responses. Likewise, intelligent machines may perceive whether humans trust them through feedback on their behavior. This perception directly affects trust formation, as individuals adjust trust levels based on perceived behaviors. Perceived trust plays a critical role in trust formation and maintenance, serving as an important factor in human-machine collaboration and trust quality enhancement, particularly providing new pathways for dynamic trust construction in complex interactive environments.

3.4 Behavioral Trust

Behavioral trust refers to actual reliance or cooperative behaviors made by individuals in specific situations based on integrated processing of dispositional and perceived trust, representing the action manifestation of trust relationships and the core dimension of all trust models. Both Gao et al. [?] and Qi Yue et al. [?] emphasize the importance of behavioral trust. Gao et al. combine it with real-time trust, arguing that situational and system characteristics affect system performance, thereby forming real-time trust. Qi Yue et al. define it as whether the trusted party executes decisions and how execution outcomes affect the system. Behavioral trust manifests in three aspects: (1) Dependence behavior—humans deciding whether to delegate tasks based on trust in machines. For example, in autonomous driving systems, users decide whether to transfer driving control to the vehicle based on system trust and their willingness to use autonomous vehicles [?]. (2) Cooperative behavior—humans choosing to collaborate with intelligent machines. For instance, in healthcare, finance, and

customer service, users participate in cooperation and share sensitive information based on trust in intelligent systems [**?**]. (3) Advice adoption behavior—users accepting machine recommendations. Intelligent machines' trust behaviors exhibit human-like patterns, such as users adopting product recommendations or diagnostic suggestions provided by systems [**?**].

Regarding intelligent machines' trust behaviors toward humans, the new human-machine relationship emphasizes that intelligent machines are not merely tools but can act as agents with certain decision-making capabilities and intelligence levels. This role transformation enables intelligent machines to gradually exhibit human-like trust behavior patterns. Xie et al. [**?**] demonstrated that when machines reject human suggestions or instructions based on judgment, users feel untrusted by the machine, thereby reducing their willingness to use and attitudes toward the machine. This paper argues that intelligent machines' trust behaviors toward humans primarily include dependence, cooperation, and advice adoption behaviors, showing certain symmetry with human trust behaviors toward intelligent machines. However, how intelligent machines transmit trust signals to humans through behavior remains a direction worthy of further research and exploration.

In summary, this paper divides human-machine mutual trust into three stages—dispositional, perceived, and behavioral trust—demonstrating the dynamic evolutionary characteristics of trust in interaction and proposing appropriate trust measurement and computational methods based on each stage' s features, providing systematic framework support for theoretical and practical development in the human-machine trust domain.

4. Measurement and Computational Modeling Methods for Human-Machine Mutual Trust

The purpose of proposing theoretical models is to develop targeted measurement and computational modeling methods for human-machine mutual trust based on different stages' measurement characteristics, thereby enabling dynamic monitoring and calibration of mutual trust. Extensive research has explored methods for measuring human trust in intelligent machines [**?**, **?**, **?**, **?**], with trust measurement primarily including three approaches: subjective scale reporting, physiological measurement, and behavioral measurement. This section summarizes existing measurement methods and, drawing on interpersonal trust measurement experience, proposes a measurement framework and methods applicable to human-machine bidirectional trust. The research focuses on: developing stage-specific measurement tools for dispositional, perceived, and behavioral trust; exploring multidimensional, multilevel measurement methods that integrate subjective reports, physiological signals, and behavioral data to construct dynamic monitoring and calibration systems; and designing trust modeling tools adapted to human-machine interaction characteristics by drawing on interpersonal trust quantification methods. Ultimately, this study aims to provide a systematic, operational theoretical and methodological framework for measur-

ing and modeling human-machine bidirectional trust, laying the foundation for dynamic evaluation and intelligent adjustment.

4.1 Measurement of Dispositional Trust

According to the theoretical model framework in Chapter 3, human dispositional trust in machines includes two dimensions: trust propensity and algorithm aversion. Trust propensity is commonly measured using Merritt et al.' s [**?**] Propensity to Trust Machines Scale, which contains six items and is widely used in trust propensity research on machines, technology, and AI [**?**, **?**], providing a reliable foundation for quantifying trust propensity. Algorithm aversion is typically measured through questionnaires assessing participants' choice tendencies in specific contexts. For example, Reich et al. [**?**] directly measured algorithm aversion by asking, "Who would you trust more to predict personality traits? (0=human, 100=algorithm)." Additionally, Shariff et al. [**?**] assessed algorithm aversion by comparing people' s expectation differences between algorithms and human operators in identical scenarios. With the prevalence of algorithm aversion, scholars have called for more systematic quantification research, including developing algorithm aversion propensity scales and constructing algorithm aversion indices [**?**]. These quantitative tools support the transition to trust measurement and modeling, and we integrate these research advances as core components of the dispositional trust layer in our theoretical model.

In contrast, research on machine dispositional trust in humans remains preliminary. Existing studies such as Johnson and Obradovich [**?**] have attempted to measure intelligent machines' trust propensity in humans by observing ChatGPT' s behavior in trust games. The trust game is a classic method for measuring trust and cooperation, revealing whether participants choose to trust others and whether they are trustworthy through fund allocation behavior [**?**]. In trust games, the trustor decides whether to trust the other party, and the trustee decides whether to reciprocate that trust. The trustor can choose to delegate partial or full funds to the trustee, which are multiplied by a fixed factor (e.g., 3x) by the experimenter. The amount transferred by the trustor reflects trust level, while the amount returned by the trustee reflects trustworthiness. Based on this framework, Johnson and Obradovich measured ChatGPT' s trust propensity in humans by having it act as the trustor interacting with humans. They found that under appropriate incentive conditions, ChatGPT exhibited trust propensity toward humans overall. While this method measures trust propensity through behavioral data, its limitation lies in its inability to explain the causes, motivations, and influencing factors of intelligent machine trust. Currently, no subjective measurement tools analogous to human trust propensity scales exist.

Based on existing research and measurement methods, we propose the following recommendations: reference existing trust propensity scale frameworks to design scales applicable to intelligent machines' trust propensity in humans; leverage large language model assessment capabilities by asking LLMs (such

as ChatGPT) to complete scale items; and further develop new measurement methods combining behavioral data with model outputs. Through these exploratory methods, we can further improve human-machine bidirectional trust measurement and modeling, providing more scientific and effective support for dynamic trust monitoring and calibration.

4.2 Measurement of Perceived Trust

Perceived trust measurement is an important research direction in human-machine interaction, particularly for exploring the interactive aspects of human-machine bidirectional trust. Subjective scales, as a widely used trust measurement tool, have been validated across multiple domains. Alsaid et al. [**?**] reviewed trust measurement instruments, noting that Jian' s [**?**] 12-item trust scale is widely used for automated system trust assessment. Additionally, Merritt et al. [**?**] developed a multidimensional trust scale that refines trust into dimensions such as reliability and applicability. Hoffman et al. [**?**] developed a dynamic trust scale that captures trust changes over time or situations. Beyond general scales, researchers have developed self-report scales for specific application scenarios involving autonomous driving [**?**, **?**], computers [**?**], robots [**?**], and automated systems [**?**]. These scales share common features: (1) post-task evaluation—focusing on perception of machine current behavior rather than overall attitudes; and (2) scenario dependency—scale items closely related to machine behavior.

In machine trust research, machines lack human emotions, making trust mechanisms unrealistic. However, in organizational management, trust measurement can be indirectly assessed through subordinates' perception of being trusted [**?**]. This approach provides inspiration for measuring machine trust in humans. This study recommends measuring intelligent machines' trust transmission to humans through language and behavior from the perspective of human perception of being trusted by machines. Specific steps include: collecting two types of scales—one measuring trust between human organizational members and another assessing human trust in AI and machines; after collecting the original item pool, screening items based on trust measurement perspective (perceived behavioral trust) and removing items measuring dispositional trust and other overall attitudes; then modifying scales by converting active statements to passive statements or swapping subjects and objects to measure participants' perceived trust. For example, an item could be designed: "In this scenario, I believe the intelligent machine trusts my decisions." Following this approach, Xie et al. [**?**] developed and validated a questionnaire for measuring human perception of being trusted by AI, assessing trust perception when autonomous vehicles and AI accept or reject human suggestions. This research shows that positive perceived trust significantly promotes human trust in AI and technology acceptance willingness. When users perceive that AI trusts their judgment, their trust in AI and technology acceptance willingness increase more easily. Xie et al.' s research not only improves quantification methods for perceived

trust but also provides a framework for optimizing human-machine interaction design, advancing trust measurement and technology acceptance research.

4.3 Measurement of Behavioral Trust

In human-machine trust research, behavior-based measurement methods assess trust levels by observing user-system interaction behaviors, reflecting users' trust in system capability, reliability, and intent. Common behavioral trust indicators include: system adoption rate (e.g., frequency of continued system use across multiple interactions) [**?**]; human-machine team performance (e.g., task completion efficiency or outcome quality) [**?**]; decision time (e.g., user hesitation or decisiveness during decision-making) [**?**]; system intervention frequency (e.g., how often users adjust or intervene in system behavior during tasks) [**?**]; and system monitoring and input (e.g., user monitoring intensity and operational behavior during tasks) [**?**]. Another approach measures trust by comprehensively evaluating a series of behavioral performances during human-machine interaction, such as nonverbal behaviors including body language (e.g., face touching, arm crossing, leaning back) [**?**], facial expressions and eye gaze (e.g., using expression analysis or eye tracking to determine user trust levels), and spatial distance and voice (e.g., changes in interaction distance or vocal tone) [**?**]. In autonomous driving, comprehensive behavioral indicators can fully assess driver trust in autonomous vehicles. Typical behavioral trust indicators include: takeover behavior (whether users actively assume control during tasks) [**?**], speed and throttle control [**?**], and gaze patterns (frequency and duration of attention to system interfaces) [**?**], all used to measure driver trust levels during simulated or actual driving. These indicators reflect drivers' dependence levels and psychological responses when facing autonomous systems. Overall, researchers commonly employ methods evaluating human responses to machine behaviors across different contexts to measure human behavioral trust in machines.

Drawing on these methods, machine trust in humans can be assessed by analyzing machine reactions to human behavior in different collaborative scenarios. For example, observing how machines adjust behavior based on human performance when collaborating with large language models (such as ChatGPT) can evaluate machine trust levels. Additionally, machines can be programmed with specific evaluation rules to define and measure trust in humans. For instance, in AI resume screening systems, trust can be assessed based on candidate performance; in personal credit rating, AI trust in users can be quantified by analyzing historical repayment behavior; in driving scoring systems, machine trust can be evaluated based on driver safety or risky behaviors. Jorge et al. [**?**] structurally measured AI trust by designing a virtual supermarket item-finding task game, using items found per unit time as a competence indicator, participants' cooperation willingness with AI as a benevolence indicator, and participants' lying frequency to AI as an integrity indicator, combining these three metrics as AI' s evaluation criteria for human trust.

4.4 Measurement of Integrated Dynamic Bidirectional Trust

The aforementioned human-machine trust measurement methods primarily focus on single dimensions of subjective psychology or behavior, typically involving only one direction of trust between human and machine parties. In practice, measuring bidirectional trust usually requires applying these tools separately to both parties and merging results to reflect bidirectional trust. While this approach partially addresses bidirectional trust measurement, results lack unity and require cumbersome tools. To address these limitations, researchers have proposed integrated dynamic bidirectional trust measurement methods that combine multimodal information such as subjective data (e.g., questionnaire feedback), behavioral data (e.g., interaction logs), and physiological data (e.g., heart rate, skin conductance) to enable dynamic tracking and assessment of trust relationships. This multimodal approach not only reveals trust changes during interaction but also more comprehensively reflects the complexity and interactivity of human-machine trust. Table 2 summarizes several typical integrated computational measurement models for human-machine trust. Through analyzing these models, we aim to propose a method for predicting bidirectional trust dynamic changes based on unified datasets. This method combines existing trust measurement technologies, integrates subjective, behavioral, and physiological data, and further explores how these data can effectively predict and evaluate the evolution of human-machine bidirectional trust during interaction.

In human-to-machine trust computational models, researchers typically combine psychological, physiological, and behavioral data, using machine learning algorithms to quantify and predict human trust in machines. These models' logic is based on relationships between trust factor representations observed during real-time human-machine interaction and corresponding objective measurement indicators and temporal dynamics [**?**, **?**, **?**]. Driver trust monitoring in autonomous driving is an important research direction, typically collecting real-time monitoring data including hand movements (e.g., steering wheel operation frequency and force), eye movements (e.g., gaze duration and trajectory), non-driving activities (e.g., driver distraction), system usage (e.g., autonomous driving function activation frequency), and physiological signals (e.g., EEG, skin conductance, and heart rate changes) [**?**, **?**, **?**]. Corresponding trust judgment indicators exhibit real-time and dynamic characteristics, commonly including takeover activities, frequency-domain and time-domain features of critical events, and gaze area switching [**?**, **?**, **?**, **?**]. Yu et al. [**?**] proposed a model for predicting driver trust in autonomous vehicles by combining objective and subjective indicators, evaluating trust levels by comparing driver hand position and movement. Avetisyan et al. [**?**] incorporated driver personality, initial trust, and dynamic trust into a human-vehicle trust monitoring model, revealing vehicle failure impacts on trust. Yi et al. [**?**] established a real-time identification model for driver trust in autonomous vehicles by fusing physiological signals (skin conductance, ECG) with takeover behavior using label-smoothed Convolutional Neural Networks

(CNN) and Long Short-Term Memory (LSTM) networks, dynamically capturing driver trust states and providing new technical means for precise trust assessment. Zhang et al. [**?**] proposed a new assessment method for evaluating subtle changes in driver trust in autonomous vehicles through EEG signals.

These methods share common characteristics: they integrate multimodal data from psychology, physiology, and especially behavior for dynamic monitoring, using machine learning algorithms to capture and predict human trust in machines. Notably, while psychological data are measured through targeted scales with certain specificity, physiological and behavioral data typically involve comprehensive monitoring of individual behavior, with trust measurement relying on general data collection and definition to derive trust prediction methods. Based on this approach, we can separately define general data in human-machine interaction processes to enable a single dataset to measure bidirectional trust simultaneously. This method systematically models the interaction between both parties' trust, allowing a single dataset to capture both individual trust in machines and machine trust in individuals. Physiological data possess bidirectionality—for example, in driving, gaze information and facial expressions can infer driver distrust in vehicles (e.g., gaze switching from secondary to primary driving tasks may reflect vehicle distrust) [**?**, **?**] while also revealing driver fatigue, cognitive load, and attention states [**?**]. Behavioral data are similarly bidirectional: driver behavior reflects both human trust in machines and can define machine trust in humans. For instance, Tesla' s driver safety scoring system evaluates driver safe driving capability through five manual driving behaviors (hard braking, sharp turns, dangerous following) and takeover time during autonomous driving usage. These indicators can be viewed as manifestations of machine trust in drivers. Takeover reaction time has also been defined by researchers as an indicator of human trust in autonomous systems—longer takeover times may indicate higher trust, while shorter times may indicate distrust [**?**, **?**].

Based on these methods, researchers can structurally define theoretical models of machine trust in humans and design unified physiological, behavioral, and psychological monitoring systems to dynamically estimate human-machine bidirectional trust levels. This comprehensive approach not only precisely captures trust dynamics in human-machine interaction but also provides robust support for system optimization and adaptive adjustment. In the future, this bidirectional trust modeling approach will further enhance intelligent system reliability and human-machine collaboration efficiency, providing a solid foundation for intelligent system innovation and development.

5.1 Application Case Studies

Human-machine mutual trust research holds significant importance across multiple key domains, substantially improving collaboration efficiency and contributing positively to accident prevention, safety analysis, and proactive intervention. Through mutual trust, machines can accurately understand and support human decision-making, reducing conflicts and "human-machine fighting" phenomena

[?]. For example, in collaborative driving scenarios between autonomous vehicles and human drivers, when the autonomous system possesses high driving capability while the human driver is limited in ability or suboptimal in state, the machine must accurately assess driver capability to decide whether to intervene or assume control [?, ?]. In this process, the autonomous system' s evaluation of driver capability reflects system trust in humans. Furthermore, when systems provide reminders, proactive interventions, or feedback on driving capability and safety scores, they directly trigger driver trust perceptions. This perception not only affects driver trust in the system but also influences technology acceptance and advice adoption willingness. Therefore, designing reasonable system reminder mechanisms based on human-machine mutual trust, particularly human trust perception, can balance the discomfort potentially caused by proactive intervention against the benefits of safety systems. Proper human-machine trust design helps optimize driver-autonomous system interaction experiences, enhance safety [?], increase user acceptance and dependence, and support more efficient collaborative driving.

In aviation, as technology reliability continues improving, pilots' response capabilities in special situations become critical for flight safety. Surveys show that over 75% of civil aviation accidents stem from human factors, with 41% related to improper handling of unexpected events [?]. When pilots experience acute stress reactions, they may exhibit physiological, psychological, and behavioral responses such as increased hormone secretion, rapid breathing, accelerated heartbeat, emotional tension, and cognitive dissonance, potentially losing cognitive skills for aircraft state awareness and leading to catastrophic consequences [?, ?]. In such situations, civil aviation autopilot systems (APS) not only assist pilots in executing flight tasks over extended periods but also provide critical safety support when pilots are in poor condition, compensating for potential errors or judgment mistakes. Through human-machine bidirectional trust theoretical models and measurement methods, real-time monitoring of pilot capability states can be achieved, analyzing error behaviors and generating comprehensive scores (viewable as system' s initial trust in pilots). This scoring mechanism not only helps systems intervene timely when necessary but also provides foundations for subsequent training. Meanwhile, function coordination and allocation based on bidirectional trust models help avoid "human-machine fighting" phenomena, thereby reducing aviation accident rates [?, ?]. This bidirectional trust mechanism provides important support for flight safety assurance and human-machine collaboration optimization.

These cases from different domains demonstrate the profound impact of machine trust in humans. Machine trust not only supports human behavior but also dynamically changes human trust levels, attitudes, and usage intentions toward machines. This further emphasizes the critical role of building reliable human-machine mutual trust systems in optimizing collaboration efficiency and enhancing safety.

5.2 Future Research Directions

Based on the above analysis of human-machine bidirectional trust theoretical frameworks and measurement methods, we identify three research areas worthy of further attention in the AI era:

(1) Development of measurement tools for machine trust in humans. In human-machine interaction, measurement tools for machine trust in humans are crucial, particularly when machines are applied to specific tasks and scenarios. Existing tools require validation and refinement through questionnaires and behavioral experiments. The applicability and adaptation of tools across different scenarios and task environments should also be addressed. For example, intelligent assistants, autonomous vehicles, and industrial robots have different trust requirements and evaluation standards. Future research should strive to establish more refined and precise measurement tools to enhance human-machine collaboration effectiveness.

(2) Human acceptance of machine trust and feedback interventions. Human technology acceptance and usage intentions are important topics in human-machine team research [**?**]. Particularly when machines score humans, output trust levels, or provide suggestions and interventions on behavior, the acceptability of these feedback and their acceptance degree directly affect human-machine interaction quality and effectiveness. Currently, systematic research on acceptance willingness and attitudes toward machine trust is lacking, especially regarding how different feedback types (e.g., positive vs. negative) affect human acceptance and how interaction design can be optimized to improve acceptance.

(3) Mechanisms of machine trust impact on human-machine collaboration performance and psychology, attitudes, and behaviors. Machine trust not only affects collaboration outcomes but also subtly influences users' psychological states, attitudes, and behaviors. Although existing research has addressed machine trust in humans, most remains at the theoretical model stage, lacking systematic empirical studies. Research should deeply explore machine trust's impact on human psychological states (e.g., self-confidence, autonomy, or sense of control) and whether machine trust triggers long-term attitude changes toward machines. Additionally, alignment between machine trust and human trust requires investigation, particularly whether machines exhibit "over-trust" or "under-trust" phenomena and how these affect collaboration effectiveness—critical directions for future research.

References

Chen, L. (2020). Labor order under "digital control" —A study of labor control of food delivery riders. *Sociological Studies*, 35(06), 113–135+244.

Gao, Z., Li, W., Liang, J., Pan, H., Xu, W., & Shen, M. (2021). Human-machine trust in autonomous vehicles. *Advances in Psychological Science*, 29(12), 2121–2132.

Luo, Y., Zhu, G., Qian, W., Wu, Y., Huang, J., & Yang, Z. (2023). Algorithm aversion in the AI era: Research framework and future prospects. *Management World*, 23(10), 205–227.

Qi, Y., Chen, J., Qin, S., & Du, F. (2024). Human-AI trust in the era of general artificial intelligence. *Advances in Psychological Science*, 32(12), 1–13.

Xu, W., Gao, Z., & Ge, L. (2024). New paradigm orientations and priorities for human factors science research in the intelligent era. *Acta Psychologica Sinica*, 56(3), 363–382.

Xu, W., & Ge, L. (2020). Engineering psychology in the intelligent era. *Advances in Psychological Science*, 28(9), 1409–1425.

Agreste, S., De Meo, P., Ferrara, E., Piccolo, S., & Provetti, A. (2015). Trust networks: Topology, dynamics, and measurements. *IEEE Internet Computing*, 19(6), 26–35.

Alhaji, B., Büttner, S., Sanjay Kumar, S., & Prilla, M. (2024). Trust dynamics in human interaction with an industrial robot. *Behaviour & Information Technology*, https://doi.org/10.1080/0144929X.2024.2316284

Allen, R., & Choudhury, P. (2022). Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organization Science*, 33(1), 149–169.

Alsaid, A., Li, M., Chiou, E. K., & Lee, J. D. (2023). Measuring trust: A text analysis approach to compare, contrast, and select trust questionnaires. *Frontiers in Psychology*, 14, 1192020.

Ambady, N., & Weisbuch, M. (2010). Nonverbal behavior. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 464–497). Hoboken, NJ: John Wiley & Sons.

Avetisyan, L., Ayoub, J., Yang, X. J., & Zhou, F. (2024). Building contextualized trust profiles in conditionally automated driving. *IEEE Transactions on Human-Machine Systems*, 54(6), 658–667.

Azevedo-Sa, H., Yang, X. J., Robert, L. P., & Tilbury, D. M. (2021). A unified bidirectional model for natural and artificial trust in human–robot collaboration. *IEEE Robotics and Automation Letters*, 6(3), 5913–5920.

Babashahi, L., Barbosa, C. E., Lima, Y., Lyra, A., Salazar, H., Argôlo, M., ··· & Souza, J. M. D. (2024). AI in the workplace: A systematic review of skill transformation in the industry. *Administrative Sciences*, 14(6), 127.

Baer, M. D., Dhensa-Kahlon, R. K., Colquitt, J. A., Rodell, J. B., Outlaw, R., & Long, D. M. (2015). Uneasy lies the head that bears the trust: The effects of feeling trusted on emotional exhaustion. *Academy of Management Journal*, 58(6), 1637–1657.

Baer, M. D., Frank, E. L., Matta, F. K., Luciano, M. M., & Wellman, N. (2021). Undertrusted, overtrusted, or just right? The fairness of (in)congruence between trust wanted and trust received. *Academy of Management Journal*, 64(1), 180–206.

Basu, C., & Singhal, M. (2016, March). Trust dynamics in human autonomous vehicle interaction: A review of trust models. In *2016 AAAI Spring Symposium Series - Technical Report* (pp. 85–91). Palo Alto, CA: AAAI Press.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1), 122–142.

Bonneviot, F., Coeugnet, S., & Brangier, E. (2021). Pedestrians-automated vehicles interaction: Toward a specific trust model. In N. L. Black, W. P. Neumann, & I. Noy (Eds.), *Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021)* (Vol. 221, pp. 568–574). Springer, Cham.

Caldwell, S., Sweetser, P., O' donnell, N., Knight, M. J., Aitchison, M., Gedeon, T., ···& Conroy, D. (2022). An agile new research framework for hybrid human-AI teaming: Trust, transparency, and transferability. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(3), 1–36.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5), 809–825.

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors*, 59(3), 333–345.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, 31(10), 692–702.

Chung, H., Holder, T., Shah, J., & Yang, X. J. (2024). Developing a team classification scheme for human-agent teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 68(1), 1394–1399.

Cuzzolin, F., Morelli, A., Cirstea, B., & Sahakian, B. J. (2020). Knowing me, knowing you: Theory of mind in AI. *Psychological Medicine*, 50(7), 1057–1061.

de Visser, E. J., & Pak, R., Shaw, T. H. (2018). From 'automation'to 'autonomy' : The importance of trust repair in human-machine interaction. *Ergonomics*, 61(10), 1409–1427.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3), 1155–1170.

Ding, Y., & Liang, Z. (2018). Structural optimization and measurement of Chinese employees' perception of being trusted. In W. Striełkowski, J. M. Black, S. A. Butterfield, C.-C. Chang, J. Cheng, F. P. Dumanig, R. Al-Mabuk, M. Urban, & S. Webb (Eds.), *Proceedings of the 2018 2nd International Conference on Management, Education and Social Science (ICMESS 2018)* (pp. 1392–1395). Atlantis Press.

Dong, Y., Hu, Z., Uchimura, K., & Murayama, N. (2010). Driver inattention monitoring system for intelligent vehicles: A review. *IEEE Transactions on Intelligent Transportation Systems*, 12(2), 596–614.

Earle, T. C., & Siegrist, M. (2006). Morality information, performance information, and the distinction between trust and confidence. *Journal of Applied Social Psychology*, 36(2), 383–416.

Ebnali, M., Hulme, K., Ebnali-Heidari, A., & Mazloumi, A. (2019). How does training effect users' attitudes and skills needed for highly automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 66, 184–195.

Fang, Z., Wang, J., Liang, J., Yan, Y., Pi, D., Zhang, H., & Yin, G. (2023). Authority allocation strategy for shared steering control considering human-machine mutual trust level. *IEEE Transactions on Intelligent Vehicles*, 9(1), 2002–2015.

Feng, F., Bao, S., Sayer, J., & LeBlanc, D. (2016). Spectral power analysis of drivers' gas pedal control during steady-state car-following on freeways. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 60(1), 729–733.

Fernández, A., Usamentiaga, R., Carús, J. L., & Casado, R. (2016). Driver distraction using visual-based sensors and algorithms. *Sensors*, 16(11), 1805.

Garcia, D., Kreutzer, C., Badillo-Urquiola, K., & Mouloua, M. (2015). Measuring trust of autonomous vehicles: A development and validation study. In C. Stephanidis (Ed.), *HCI International 2015-Posters' Extended Abstracts* (Vol. 529, pp. 610–615). Springer, Cham.

Gebru, B., Zeleke, L., Blankson, D., Nabil, M., Nateghi, S., Homaifar, A., & Tunstel, E. (2022). A review on human–machine trust evaluation: Human-centric and machine-centric perspectives. *IEEE Transactions on Human-Machine Systems*, 52(5), 952–962.

Georganta, E., & Ulfert, A. S. (2024). Would you trust an AI team member? Team trust in human–AI teams. *Journal of Occupational and Organizational Psychology*, 97, 1212–1241.

Gillespie, N. (2012). Measuring trust in organizational contexts: An overview of survey-based measures. In F. Lyon, G. Möllering, & M. Saunders (Eds.), *Handbook of research methods on trust* (pp. 175–188). Edward Elgar Publishing.

Gunning, D., Vorm, E., Wang, Y., & Turek, M. (2021). DARPA' s explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2, e61. https://doi.org/10.1002/ail2.61

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5), 517–527.

He, X., Nie, X., Zhou, R., Yang, J., & Wu, R. (2023). The risk-taking behavioural intentions of pilots in adverse weather conditions: An application of the theory of planned behaviour. *Ergonomics*, 66(8), 1043–1056.

Hieronymi, P. (2008). The reasons of trust. *Australasian Journal of Philosophy*, 86(2), 21–236.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434.

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*, 28(1), 84–88.

Hu, C., Huang, S., Zhou, Y., Ge, S., Yi, B., Zhang, X., & Wu, X. (2024). Dynamic and quantitative trust modeling and real-time estimation in human-machine co-driving process. *Transportation Research Part F: Traffic Psychology and Behaviour*, 106, 306–327.

Inga, J., Ruess, M., Robens, J. H., Nelius, T., Rothfuß, S., Kille, S., ···& Kiesel, A. (2023). Human-machine symbiosis: A multivariate perspective for physically coupled human-machine systems. *International Journal of Human-Computer Studies*, 170, 102926.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, 61(4), 577–586.

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71.

Johnson, T., & Obradovich, N. (2022). Measuring an artificial intelligence agent' s trust in humans using machine incentives. *arXiv preprint arXiv:2212.13371*. https://doi.org/10.48550/arXiv.2212.13371

Jorge, C. C., Jonker, C. M., & Tielman, M. L. (2024). How should an AI trust its human teammates? Exploring possible cues of artificial trust. *ACM Transactions on Interactive Intelligent Systems*, 14(1), 1–26.

Jorge, C. C., Tielman, M. L., & Jonker, C. M. (2022a). Artificial trust as a tool in human-AI teams. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 1155–1157). IEEE. https://doi.org/10.1109/HRI53351.2022.9889652

Jorge, C. C., Tielman, M. L., & Jonker, C. M. (2022b). Assessing artificial trust in human-agent teams: A conceptual model. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents* (Article 1-3). Association for Computing Machinery. https://doi.org/10.1145/3514197.3549696

Kamaraj, A. V., Lee, J., Domeyer, J. E., Liu, S. Y., & Lee, J. D. (2024). Comparing subjective similarity of automated driving styles to objective distance-based similarity. *Human Factors*, 66(5), 1545–1563.

Kamaraj, A. V., Lee, J., Parker, J. I., Domeyer, J. E., Liu, S. Y., & Lee, J. D. (2023). Bimodal trust: High and low trust in vehicle automation influence response to automation errors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67(1), 1144–1149. https://doi.org/10.1177/21695067231196244

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

Kaur, D., Uslu, S., Durresi, A., Mohler, G., & Carter, J. G. (2020). Trust-based human-machine collaboration mechanism for predicting crimes. In L. Barolli, F. Amato, F. Moscato, T. Enokido, & M. Takizawa (Eds.), *Advanced information networking and applications. AINA 2020* (Vol. 1151). Springer, Cham.

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2017). Calibrating trust to increase the use of automated systems in a vehicle. In N. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in human aspects of transportation* (Vol. 484). Springer, Cham.

Kintz, J. R., Banerjee, N. T., Zhang, J. Y., Anderson, A. P., & Clark, T. K. (2023). Estimation of subjectively reported trust, mental workload, and situation awareness using unobtrusive measures. *Human Factors*, 65(6), 1076–1093.

Kobayashi, G., Quilici-Gonzalez, M. E., Broens, M. C., & Quilici-Gonzalez, J. A. (2016). The ethical impact of the internet of things in social relationships: Technological mediation and mutual trust. *IEEE Consumer Electronics Magazine*, 5(3), 85–89.

Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y. C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, 12, 604977. https://doi.org/10.3389/fpsyg.2021.604977

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, 50(1), 569–598.

Lau, D. C., & Lam, L. W. (2008). Effects of trusting and being trusted on team citizenship behaviours in chain stores. *Asian Journal of Social Psychology*, 11(2), 141–149.

Lau, D. C., Lam, L. W., & Wen, S. S. (2014). Examining the effects of feeling trusted by supervisors in the workplace: A self-evaluative perspective. *Journal*

of Organizational Behavior, 35(1), 112–127.

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

Lee, J. D., Liu, S. Y., Domeyer, J., & DinparastDjadid, A. (2021). Assessing drivers' trust of automated vehicle driving styles with a two-part mixed model of intervention tendency and magnitude. *Human Factors*, 63(2), 246–260.

Lee, J. J., Knox, B., & Breazeal, C. (2013). Modeling the dynamics of nonverbal behavior on interpersonal trust for human-robot interactions. In *Trust and autonomous systems: Papers from the 2013 AAAI Spring Symposium* (pp. 46-47). AAAI.

Li, M., & Lee, J. D. (2022). Modeling goal alignment in human-AI teaming: A dynamic game theory approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 66(1), 1538–1542. https://doi.org/10.1177/1071181322661047

Li, M., Erickson, I. M., Cross, E. V., & Lee, J. D. (2024a). It' s not only what you say, but also how you say it: Machine learning approach to estimate trust from conversation. *Human Factors*, 66(6), 1724–1741.

Li, M., Kamaraj, A. V., & Lee, J. D. (2024b). Modeling trust dimensions and dynamics in human-agent conversation: A trajectory epistemic network analysis approach. *International Journal of Human–Computer Interaction*, 40(14), 3571-3582.

Lu, Z., Happee, R., Cabrall, C. D., Kyriakidis, M., & De Winter, J. C. (2016). Human factors of transitions in automated driving: A general framework and literature survey. *Transportation Research Part F: Traffic Psychology and Behaviour*, 43, 183–198.

Lyons, J. B., Wynne, K. T., Mahoney, S., & Roebke, M. A. (2019). Trust and human-machine teaming: A qualitative study. In W. Lawless, R. Mittu, D. Sofge, I. S. Moskowitz, & S. Russell (Eds.), *Artificial intelligence for the Internet of everything* (pp. 101–116). Academic Press.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of the 11th Australasian Conference on Information Systems* (Vol. 53, pp. 6–8).

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175, 121390.

Mathavara, K., & Ramachandran, G. (2022). Role of human factors in preventing aviation accidents: An insight. In *Aeronautics-New Advances* (pp. 1–26). IntechOpen. https://doi.org/10.5772/intechopen.106899

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

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human Factors*, 60(2), 262–273.

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, 53(4), 356–370.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, 50(2), 194–210.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don' t know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, 55(3), 520–534.

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors*, 57(1), 34–47.

Michelaraki, E., Katrakazas, C., Kaiser, S., Brijs, T., & Yannis, G. (2023). Real-time monitoring of driver distraction: State-of-the-art and future insights. *Accident Analysis & Prevention*, 192, 107241.

Möhlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R. W. (2021). Algorithmic management of work on online labor platforms: When matching meets control. *MIS Quarterly*, 45(4).

Montag, C., Kraus, J., Baumann, M., & Rozgonjuk, D. (2023). The propensity to trust in (automated) technology mediates links between technology self-efficacy acceptance artificial intelligence. *Computers in Human Behavior Reports*, 11, 100315.

Mueller, F. F., Lopes, P., Strohmeier, P., Ju, W., Seim, C., Weigel, M., ···& Maes, P. (2020). Next steps for human-computer integration. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI' 20)* (pp. 1–15). Association for Computing Machinery.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5-6), 527–539.

Murphy, R. R. (2024). What will robots think of us? *Science Robotics*, 9(86), eadn6096. https://doi.org/10.1126/scirobotics.adn6096

Murphy, R., & Woods, D. D. (2009). Beyond Asimov: The three laws of responsible robotics. *IEEE Intelligent Systems*, 24(4), 14–20.

Nies, H. (2009). Key elements in effective partnership working. In J. Glasby & H. Dickinson (Eds.), *International perspectives on health and social care: Partnership working in action* (pp. 56–67). Wiley-Blackwell.

Olson, D. M., & Xu, Y. (2021). Building Trust Over Time in Human-Agent Relationships. In *Proceedings of the 9th International Conference on Human-*

Agent Interaction (pp. 193–201). Association for Computing Machinery. https://doi.org/10.1145/3472307.3484178

Pakdamanian, E., Sheng, S., Baee, S., Heo, S., Kraus, S., & Feng, L. (2021). Deeptake: Prediction of driver takeover behavior using multimodal data. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 103, pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445563

Parnell, K. J., Wynne, R. A., Griffin, T. G., Plant, K. L., & Stanton, N. A. (2021). Generating design requirements for flight deck applications: Applying the perceptual cycle model to engine failures on take-off. *International Journal of Human–Computer Interaction*, 37(7), 611–629.

Payre, W., Cestac, J., Dang, N. T., Vienne, F., & Delhomme, P. (2017). Impact of training and in-vehicle task performance on manual control recovery in an automated car. *Transportation Research Part F: Traffic Psychology and Behaviour*, 46, 216–227.

Pitardi, V., & Marriott, H. R. (2021). Alexa, she' s not human but···Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing*, 38(4), 626–642.

Prahl, A., Leung, R. K. H., & Chua, A. N. S. (2022). Fight for flight: The narratives of human versus machine following two aviation tragedies. *Human-Machine Communication*, 4, 27–42.

Qu, Y., Hu, H., Liu, J., Zhang, Z., Li, Y., & Ge, X. (2023). Driver state monitoring technology for conditionally automated vehicles: Review and future prospects. *IEEE Transactions on Instrumentation and Measurement*, 72, Article 3000920, 1–20.

Regli, C., & Annighoefer, B. (2022). An anthropomorphic approach to establish an additional layer of trustworthiness of an AI pilot. In *Software Engineering 2022 Workshops* (pp. 160–180). Gesellschaft für Informatik e.V. https://doi.org/10.18420/se2022-ws-17

Reich, T., Kaju, A., & Maglio, S. J. (2023). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, 33(2), 285–302.

Robertson, I. W. T. (2021). *The development and initial validation of the trust in self-driving vehicles scale (tsdv)* (Doctoral dissertation). Rice University.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35(4), 651–665.

Sadrfaridpour, B., Saeidi, H., Burke, J., Madathil, K., & Wang, Y. (2016). Modeling and control of trust in human-robot collaborative manufacturing. In R. Mittu, D. Sofge, A. Wagner, & W. Lawless (Eds.), *Robust intelligence and trust in autonomous systems* (pp. 115–141). Springer.

Salamon, S. D., & Robinson, S. L. (2008). Trust that binds: The impact of collective felt trust on organizational performance. *Journal of Applied Psychology*, 93(3), 593.

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T. L., Chen, J. Y., & Hancock, P. A. (2014). A meta-analysis of factors influencing the development of trust in automation, Implications for Human-robot Interaction. Aberdeen: Army Research Laboratory.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, 58(3), 377–400.

Seet, M., Harvy, J., Bose, R., Dragomir, A., Bezerianos, A., & Thakor, N. (2020). Differential impact of autonomous vehicle malfunctions on human trust. *IEEE Transactions on Intelligent Transportation Systems*, 23(1), 548–557.

Shariff, A., Bonnefon, J. F., & Rahwan, I. (2021). How safe is safe enough? Psychological mechanisms underlying extreme safety demands for self-driving cars. *Transportation Research Part C: Emerging Technologies*, 126, 103069.

Shi, Z., O' Connell, A., Li, Z., Liu, S., Ayissi, J., Hoffman, G., ···& Matarić, M. J. (2024). Build your own robot friend: An open-source learning module for accessible and engaging AI education. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI 24/IAAI 24/EAAI 24)* (Article 2636, pp. 1–9). AAAI Press. https://doi.org/10.1609/aaai.v38i21.30359

Simons, T., Leroy, H., & Nishii, L. (2022). Revisiting behavioral integrity: Progress and new directions after 20 years. *Annual Review of Organizational Psychology and Organizational Behavior*, 9(1), 365–389.

Simpson, J. A. (2007). Psychological foundations of trust. *Current Directions in Psychological Science*, 16(5), 264–268.

Strauch, C., Mühl, K., Patro, K., Grabmaier, C., Reithinger, S., Baumann, M., & Huckauf, A. (2019). Real autonomous driving from a passenger' s perspective: Two experimental investigations using gaze behaviour and trust ratings in field and simulator. *Transportation Research Part F: Traffic Psychology and Behavior*, 66, 191–207.

Sycara, K., & Lewis, M. (2004). Integrating intelligent agents into human teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 46, 413–417.

Techer, F., Ojeda, L., Barat, D., Marteau, J. Y., Rampillon, F., Feron, S., & Dogan, E. (2019). Anger and highly automated driving in urban areas: The role of time pressure. *Transportation Research Part F: Traffic Psychology and Behaviour*, 64, 353–360.

Uggirala, A., Gramopadhye, A. K., Melloy, B. J., & Toler, J. E. (2004). Measurement of trust in complex and dynamic systems using a quantitative approach. *International Journal of Industrial Ergonomics*, 34(3), 217–227.

Ulfert, A. S., Georganta, E., Centeio Jorge, C., Mehrotra, S., & Tielman, M. (2024). Shaping a multidisciplinary understanding of team trust in human-AI teams: A theoretical framework. *European Journal of Work and Organizational Psychology*, 33(2), 158–171.

Walliser, J. C., de Visser, E. J., Wiese, E., & Shaw, T. H. (2019). Team structure and team building improve human–machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making*, 13(4), 258–278.

Walmsley, S., & Gilbey, A. (2017). Debiasing visual pilots' weather-related decision making. *Applied Ergonomics*, 65, 200–208.

Wang, J., & Moulden, A. (2021). AI Trust Score: A user-centered approach to building, designing, and measuring the success of intelligent workplace features. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 1–7). Association for Computing Machinery. https://doi.org/10.1145/3411763.3443452

Wang, Y., Wang, X., Tang, J., Zuo, W., & Cai, G. (2015). Modeling status theory in trust prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 29, 9460. https://doi.org/10.1609/aaai.v29i1.9460

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52, 113–117.

Wiggins, M. W., Azar, D., Hawken, J., Loveday, T., & Newman, D. (2014). Cue-utilisation typologies and pilots' pre-flight and in-flight weather decision-making. *Safety Science*, 65, 118–124.

Wong, J. H., Chiou, E. K., Gutzwiller, R. S., Cook, M. B., & Fallon, C. K. (2024). Human-artificial intelligence teaming for the U.S. Navy: Developing a holistic research roadmap. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 68(1), 380-385. https://doi.org/10.1177/10711813241260352

Xie, Y., Liu, Y., Zhou, R., Zhi, X., & Chan, A. H. (2024). Wait or Pass? Promoting intersection' s cooperation via identifying vehicle' s social behavior. *Accident Analysis & Prevention*, 206, 107724.

Xie, Y., Zhou, R., Chan, A. H. S., Jin, M., & Qu, M. (2023). Motivation to interaction media: The impact of automation trust and self-determination theory on intention to use the new interaction technology in autonomous vehicles. *Frontiers in Psychology*, 14, 1078438.

Xie, Y., Zhou, R., Chan, A.H.S. (In Press). Do You Trust Me? Measuring People's Perception of Being Trusted by AI in a Human–Agent Team. *International Journal of Human-Computer Interaction*.

Yang, C., Zhu, Y., & Chen, Y. (2021). A review of human–machine cooperation in the robotics domain. *IEEE Transactions on Human-Machine Systems*, 52(1), 12-25.

Yi, B., Cao, H., Song, X., Wang, J., Zhao, S., Guo, W., & Cao, D. (2024). How can the trust-change direction be measured and identified during takeover transitions in conditionally automated driving? Using physiological responses and takeover-related factors. *Human Factors*, 66(4), 1276–1301.

Yu, B., Bao, S., Zhang, Y., Sullivan, J., & Flannagan, M. (2021). Measurement and prediction of driver trust in automated vehicle technologies: An application of hand position transition probability matrix. *Transportation Research Part C: Emerging Technologies*, 124, 102957.

Yu, K., Berkovsky, S., Conway, D., Taib, R., Zhou, J., & Chen, F. (2018). Do I trust a machine? Differences in user trust based on system performance. In J. Zhou & F. Chen (Eds.), *Human and machine learning* (pp. 161–172). Springer.

Yuan, L., Gao, X., Zheng, Z., Edmonds, M., Wu, Y. N., Rossano, F., ···& Zhu, S. C. (2022). In situ bidirectional human-robot value alignment. *Science Robotics*, 7(68), eabm4183. https://doi.org/10.1126/scirobotics.abm4183

Zhang, T., Tao, D., Qu, X., Zhang, X., Lin, R., & Zhang, W. (2019). The roles of initial trust and perceived risk in public' s acceptance of automated vehicles. *Transportation Research Part C: Emerging Technologies*, 98, 207–220.

Zhang, T., Yang, J., Chen, M., Li, Z., Zang, J., & Qu, X. (2024). EEG-based assessment of driver trust in automated vehicles. *Expert Systems with Applications*, 246, 123196.

Zhou, L., Paul, S., Demirkan, H., Yuan, L., Spohrer, J., Zhou, M., & Basu, J. (2021). Intelligence augmentation: Towards building human-machine symbiotic relationship. *AIS Transactions on Human-Computer Interaction*, 13(2), 243-264.

Note: Figure translations are in progress. See original paper for figures.