frontiers | Frontiers in Psychology

TYPE Review
PUBLISHED 17 April 2024
DOI 10.3389/fpsyg.2024.1382693

Check for updates

OPEN ACCESS

EDITED BY
Hao Chen,
Nankai University, China

REVIEWED BY
Kaisheng Lai,
Jinan University, China
Lin Qiu,
Nanyang Technological University, Singapore

*CORRESPONDENCE
Shenghua Luan
✉ luansh@psych.ac.cn

RECEIVED 16 February 2024
ACCEPTED 04 April 2024
PUBLISHED 17 April 2024

CITATION
Li Y, Wu B, Huang Y and Luan S (2024) Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. *Front. Psychol.* 15:1382693. doi: 10.3389/fpsyg.2024.1382693

Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust

Yugang Li[1,2], Baizhou Wu[1,2], Yuqi Huang[1,2] and Shenghua Luan[1,2]*

[1]CAS Key Laboratory for Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China, [2]Department of Psychology, University of the Chinese Academy of Sciences, Beijing, China

The rapid advancement of artificial intelligence (AI) has impacted society in many aspects. Alongside this progress, concerns such as privacy violation, discriminatory bias, and safety risks have also surfaced, highlighting the need for the development of ethical, responsible, and socially beneficial AI. In response, the concept of trustworthy AI has gained prominence, and several guidelines for developing trustworthy AI have been proposed. Against this background, we demonstrate the significance of psychological research in identifying factors that contribute to the formation of trust in AI. Specifically, we review research findings on interpersonal, human-automation, and human-AI trust from the perspective of a three-dimension framework (i.e., the trustor, the trustee, and their interactive context). The framework synthesizes common factors related to trust formation and maintenance across different trust types. These factors point out the foundational requirements for building trustworthy AI and provide pivotal guidance for its development that also involves communication, education, and training for users. We conclude by discussing how the insights in trust research can help enhance AI's trustworthiness and foster its adoption and application.

KEYWORDS

trustworthy AI, interpersonal trust, human-automation trust, human-AI trust, AI ethics, warmth, competence, trust measurement

1 Introduction

Artificial intelligence (AI) is the driving force behind industry 4.0 and has profoundly affected manufacturing, business, work, and our daily life (Magd et al., 2022). The invention of generative AI technologies, such as ChatGPT, marks a particularly significant leap in AI's competence. While bringing significant changes to society, the development of AI has also sparked various concerns, including privacy invasion, hidden biases and discrimination, security risks, and ethical issues (Yang and Wibowo, 2022). One response to these concerns is the emergence of and emphasis on *trustworthy AI* that aims to strike a good balance between technological advancement and societal and ethical considerations (Li et al., 2023).

Trustworthy AI, defined as AI that is lawful, ethical, and robust [High-Level Expert Group on Artificial Intelligence (AI HLEG), 2019], represents a critical focus on responsible technology deployment. To develop trustworthy AI, multiple countries and international organizations have issued guidelines. For instance, the European Union issued the "Ethics Guidelines for Trustworthy AI" in April 2019 (High-Level Expert Group on Artificial Intelligence (AI HLEG), 2019); China published the "Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence" in June 2019 (National Governance Committee for the New Generation Artificial Intelligence, 2019); and on October 30, 2023, President Biden of the United States signed an executive order on the "safe, secure, and trustworthy development and use of artificial intelligence" (The White House, 2023). These guidelines lay out requirements for the development of AI that ensure safety and protect privacy, enhance transparency and accountability, and avoid discrimination.

Trust and trustworthiness are key psychological constructs that have been extensively explored in research on interpersonal, human-automation, and human-AI trust, providing many insights on how a person or an agent can become trustworthy. The aforementioned guidelines primarily outline requirements for developers and providers of AI, but do not pay sufficient attention to how end-users may develop trust in AI systems. Research on trust specifies users' expectations of AI, thus aiding the comprehension of their concerns and needs. It also assists in identifying which attributes of AI systems are crucial for establishing trust and improving their design. Furthermore, because trust has a significant impact on the adoption of AI, trust research may also help enhance the public's acceptance and adoption of AI technology.

In this paper, we apply trust theories to the context of trustworthy AI, aiming to shed lights on how to create reliable and trustworthy AI systems. In doing so, this paper makes several notable contributions to the field of AI trust research. First, we systematically review research on interpersonal, human-automation, and human-AI trust by viewing the perception of AI from the perspective of social cognition (Frischknecht, 2021). It serves to validate and build upon theoretical frameworks in previous literature reviews and meta-analyses from a broader, more coherent, and more inclusive angle. Second, based on a three-dimension framework of trust that encompasses trustor, trustee, and their interactive context, we compile and summarize a large number of factors that may influence trust in AI by reviewing a wide range of empirical studies. Third, by identifying and consolidating these influencing factors, our paper offers guidance on enhancing AI trustworthiness in its applications, bridging theoretical concepts and propositions with practical applications.

Overall, we aim to build a comprehensive framework for understanding and developing trustworthy AI that is grounded in end-users' perspectives. Here, we focus primarily on the formation of trust, and do not distinguish between specific applications in automation or AI but refer to them collectively as automation or AI (Yang and Wibowo, 2022). The following sections are organized according to the three types of trust, ending with a discussion on the implications of trust research on enhancing trustworthy AI.

2 Interpersonal trust

Rotter (1967) first defined interpersonal trust as the trustor's generalized expectancy for the reliability of another person's words or promises, whether verbal or written. This generalized expectancy, commonly known as trust propensity, is considered a personality characteristic that significantly influences actual behavior (Evans and Revelle, 2008). Trust typically arises in contexts characterized by risk and uncertainty. Mayer et al. (1995) framed interpersonal trust as a dyadic relationship between a trustor, the individual who extends trust, and a trustee, the entity being trusted, and treated trust as the willingness of the trustors to make themselves vulnerable despite knowing that the trustees' actions could significantly impact them and irrespective of the trustors' ability to monitor or control those actions. This section outlines critical factors that shape interpersonal trust.

2.1 Characteristics of interpersonal trust

Trust is a term frequently encountered in daily life, characterized by a multitude of definitions and generally regarded as a multidimensional concept. McAllister (1995) identified two dimensions of interpersonal trust: cognitive and affective, while Dirks and Ferrin (2002) expanded this to include vulnerability and overall trust. Jones and Shah (2016) further categorized trust into three dimensions: trusting actions, trusting intentions, and trusting beliefs. Different theories capture distinct characteristics of trust but exhibit several key commonalities.

First, trust is dyadic, involving a trustor and a trustee, each with certain characteristics. From a dyadic perspective, interpersonal trust can be categorized into three types: reciprocal trust, highlighting the dynamic interactions between parties; mutual trust, reflecting a shared and consistent level of trust; and asymmetric trust, indicating imbalances in trust levels within interpersonal relationships (Korsgaard et al., 2015). The propensity to trust of the trustors exhibits significant individual differences, influenced by gender (Dittrich, 2015; Thielmann et al., 2020), age (Bailey et al., 2015; Bailey and Leon, 2019), and personality traits (Ito, 2022). Trust propensity influences trust at an early stage, and the assessment of the trustworthiness of trustees (trust beliefs) may ultimately determine the level of trust (McKnight et al., 1998).

Second, interpersonal trust can be influenced by interactive contexts, such as social networks and culture (Baer et al., 2018; Westjohn et al., 2022). The trustworthiness of strangers is frequently evaluated through institutional cues, including their profession, cultural background, and reputation (Dietz, 2011). Third, trust occurs within uncertain and risky contexts, and it is closely linked to risk-taking behavior (Mayer et al., 1995). Fourth, trust is usually dynamic. Trust propensity embodies a belief in reciprocity or an initial trust, ultimately triggering a behavioral primitive (Berg et al., 1995). Trustors will determine whether to reinforce, decrease, or restore trust based on the outcomes of their interactions with trustees. Individuals form expectations or anticipations about the future behavior of the trusted entity based on various factors, such as past experience and social influences. Thus, the trust dynamic is a process of social learning that often evolves gradually and changes with interactive experiences (Mayer et al., 1995).

The above analysis shows that factors influencing interpersonal trust can be examined from three dimensions: the trustor, the trustee, and their interactive context. Interpersonal interactions correlate with changes in variables related to these three dimensions, ultimately leading to variations in the levels of trust and actual behavior.

2.2 Measurement of interpersonal trust

Trust, especially trust propensity, can be quantified using psychometric scales. These scales evaluate an individual's specific trust or disposition toward trusting others via a set of questions (Frazier et al., 2013). Example items include, "I am willing to let my partner make decisions for me" (Rempel et al., 1985), and "I usually trust people until they give me a reason not to trust them" (McKnight et al., 2002). Meanwhile, economic games, such as trust game, dictator game, public goods game, and social dilemmas, provide a direct and potentially more accurate means to assess trust by observing individual actions in well-defined contexts (Thielmann et al., 2020). This approach is beneficial for deducing levels of trust from real decisions and minimizing the impact of social desirability bias (Chen et al., 2023). When integrated, behavioral observations from economic games and self-reported beliefs yield a more comprehensive perspective on trust by combining the strengths of observed actions and declared beliefs.

The aforementioned methods represent traditional approaches to assessing interpersonal trust. Alternative methods for measuring interpersonal trust also exist. To avoid redundancy, these methods are reviewed in the "Measurement of Trust in Automation" section.

3 Trust in automation

Trust extends beyond human interactions. Its importance is notable in interactions between humans and automated systems (Hoff and Bashir, 2015). According to Gefen (2000), trust in automation is the confidence, based on past interactions, in expecting actions from automation that align with one's expectations and benefit oneself. In a similar vein, Lee and See (2004) characterize trust in automation as a quantifiable attitude that determines the extent of reliance on automated agents. Consequently, human-automation trust, akin to interpersonal trust, constitutes a psychological state that influences behaviors.

Trust is crucial for the adoption of automation technologies, and a deficiency of trust in automation can lead to reduced reliance on these systems (Lee and See, 2004). Since the 1980s, with the widespread adoption of automation technology and its increasing complexity, research in human-automation interaction, technology acceptance models, and human-automation trust has drastically expanded. This section offers a brief overview of research in this field.

3.1 Automation

Automation usually refers to the technology of using devices, such as computers, to replace human execution of tasks in modern society, where automated technologies increasingly take over functions for efficiency, accuracy, and safety purposes (Kohn et al., 2021). Based on system complexity, autonomy, and necessary human intervention, automation can be divided into 10 levels, with level 0 signifying full manual control and level 10 denoting complete automation (Parasuraman et al., 2000). Furthermore, Parasuraman et al. (2000) identified four principal functions of automation within a framework of human information processing: information acquisition, information analysis, decision making, and action execution. An automation system may exhibit varying degrees of automation across these distinct functions.

The field of human-automation interaction has evolved alongside advances in computer technology, as reflected in the progress of its terminology: from HCI (Human-Computer Interaction) and HRI (Human-Robot Interaction) to HAI (Human-AI Interaction) (Ueno et al., 2022). Initially, research on automation trust was concentrated in sectors such as military, aviation, banking, and industrial manufacturing. With advancing computer technology, the focus of automation trust research has expanded to encompass office settings and the service sector. Furthermore, in the context of the internet, the pivotal importance of trust in the adoption of e-commerce, e-governance, and social media platforms has also been extensively investigated (Gefen, 2000; Featherman and Pavlou, 2003; Khan et al., 2014).

3.2 Similarities between interpersonal trust and human-automation trust

Being a cornerstone of sustained human cooperation, trust is equally crucial to human-automation collaboration (Xie et al., 2019). Trust in humans and automation shares similarities, supported by both empirical and neurological evidence (Lewandowsky et al., 2000; Hoff and Bashir, 2015). For instance, a three-phase experiment study by Jian et al. (2000) that included tasks of word elicitation and meaning comparison showed that the constructs of trust in human-human and human-automation interactions are analogous. This resemblance may stem from the similar perceptions that individuals hold toward automated agents and fellow humans (Frischknecht, 2021).

Despite their non-human appearance, computers are often subject to social norms and expectations. Nass et al. (1997) demonstrated that assigning male or female voices to computers elicits stereotypical perceptions. In a similar vein, Tay et al. (2014) reported that robots performing tasks aligned with gender or personality stereotypes—such as medical robots perceived as female or extroverted, and security robots as male or introverted—received higher approval ratings. Moreover, studies in economic games like the ultimatum and public goods games have shown that people display prosocial behaviors toward computers, suggesting a level of social engagement (Nielsen et al., 2021; Russo et al., 2021).

The Computers Are Social Actors (CASA) paradigm posits that during human-computer interactions, individuals often treat computers and automated agents as social beings by applying social norms and stereotypes to them (Nass et al., 1997). Such anthropomorphization usually happens subconsciously, leading to automated agents being perceived with human-like qualities (Kim and Sundar, 2012). In reality, intelligent devices exhibit anthropomorphism by mimicking human features or voices, setting them apart from traditional automation (Troshani et al., 2021; Liu and Tao, 2022). Although AI lacks emotions and cannot be held accountable for its actions, it is usually perceived through the lens of social cognition, making it difficult to regard AI as merely a machine or software; instead, AI is often viewed as an entity worthy of trust (Ryan, 2020).

3.3 Importance of trust in automation

Automation differs significantly from machines that operate specific functions entirely and indefinitely without human intervention (Parasuraman and Riley, 1997). For example, traditional vehicle components such as engines, brakes, and steering systems are generally regarded as highly reliable; in contrast, autonomous vehicles often evoke skepticism regarding their capabilities (Kaplan et al., 2021). While tasks performed by automation could also be executed by humans, the decision to rely on automation is contingent upon trust. For instance, individuals may refrain from using a car's autonomous driving feature if they distrust its reliability. Moreover, the complexity of automation technologies may lead to a lack of full understanding by users (Muir, 1987), a gap that trust can help to bridge. Additionally, automation systems are also known to be particularly vulnerable to unexpected bugs (Sheridan, 1988), making the effectiveness of such systems heavily reliant on users' trust in their performance (Jian et al., 2000).

Because of different individual understandings of automation and the complexity of automated systems, people may exhibit inappropriate trust in automation (Lee and See, 2004), which may lead to misuse or disuse. Misuse refers to the inappropriate use of automation, such as automation bias, that is, people relying on the recommendations of automated systems too much instead of exercising careful information search and processing. They may ignore information that contradicts the suggestions of the automation system, even if that information may be correct (Parasuraman and Manzey, 2010). Disuse refers to people refusing to use automation (Venkatesh et al., 2012). For instance, in decision-making tasks, algorithm aversion, which refers to skepticism toward algorithms, the core of automated systems, often takes place (Burton et al., 2020). The lack of public acceptance impedes advanced technology from achieving its full potential and practical application (Venkatesh et al., 2012). Similarly, inappropriate trust compromises the effectiveness of automated systems. Aligning the public's trust level with the developmental stage of automation represents an ideal scenario. Thus, it is crucial to investigate the factors that shape trust in automation.

In human-automation trust, viewing trust as an attitude is widely accepted. However, in the context of interpersonal trust, the term *attitude* is not often used, whereas *willingness* is commonly employed. This distinction likely stems from technology acceptance theories, which posit that attitude shapes behavioral intentions and in turn influences actual behavior (Gefen et al., 2003). Hence, the importance of trust in automation is underscored by its effect on users' behavior toward automated systems.

3.4 Measurement of trust in automation

Kohn et al. (2021) conducted a comprehensive review of methods for measuring human trust in automation, classifying these into self-reports, behavioral measures, and physiological indicators. Self-report methods typically involve questionnaires and scales, while behavioral measures include indicators like team performance, compliance and agreement rate, decision time, and delegation. Despite varied terminologies, these measures are based on the same principle: individuals demonstrating trust in an automation system are more inclined to follow its recommendations, depend on it, comply with its advice, lessen their oversight of the system, and delegate decision-making authority to it. Such behaviors are more evident when the automation system demonstrates high accuracy, potentially improving group performance. In dual-task situations, systems that are trusted usually result in faster decision-making and response times for ancillary tasks, whereas distrust can lead to slower responses.

Physiological indicators include those from skin conductance, EEG (electroencephalography), fMRI (functional magnetic resonance imaging), and fNIRS (functional near-infrared spectroscopy). A notable finding is that a reduction in skin conductance, suggesting lower cognitive load and emotional arousal, is associated with increased trust in automation (Khawaji et al., 2015). Moreover, employing methodologies like EEG, fMRI, and fNIRS to investigate the brain regions engaged in processing trust has demonstrated notable alterations (Ajenaghughrure et al., 2020).

Self-reporting methods, such as questionnaires measuring trust, capture static aspects of trust but cannot reflect its dynamic nature. Ayoub et al. (2023) introduced a dynamic measurement of trust during autonomous driving. Drivers' physiological measures, including galvanic skin response, heart rate indices, and eye-tracking metrics, are recorded in real-time. Machine learning algorithms were then used to estimate trust based on these data. This real-time assessment of trust is critical for capturing its dynamic changes, thereby facilitating trust calibration.

3.5 The relationship between automation and AI

3.5.1 AI: a next generation of automation

AI typically refers to the simulation of human intelligence by computers (Gillath et al., 2021). This simulation process encompasses learning (acquiring information and using it to acquire rules), reasoning (using rules to reach conclusions), and self-correction. In essence, AI represents a sophisticated form of automation, enhancing its domain and efficacy. In this paper, we refer automation as traditional automated technologies that are distinct from AI. The distinction lies in automation being systems that perform repetitive tasks based on static rules or human commands, while AI involves systems skilled in dealing with uncertainties and making decisions in novel situations (Cugurullo, 2020; Kaplan et al., 2021).

Interestingly, in the initial research on human-automation interaction, AI was considered a technology difficult to implement (Parasuraman and Riley, 1997). However, in the 21st century and especially after 2010, AI technology has progressed significantly. Nowadays, the impact of AI technology and its applications pervade daily life and professional environments, encompassing speech and image recognition, autonomous driving, smart homes, among others. AI can autonomously deliver personalized services by analyzing historical human data (Lu et al., 2019). Particularly, the recent advancements in generative AI have provided a glimpse of the potential for achieving general AI. That said, understanding how current AI arrives at specific decisions or outcomes can be complex due to its reliance on vast amounts of data and intricate algorithms.

3.5.2 From trust in automation to trust in AI

With the advancement and widespread applications of AI, trust in AI has indeed become a new focal point in the study of human-automation interaction. This transition redefines relationships between humans and automation, moving from reliance on technologies for repetitive, accuracy-driven tasks to expecting AI to demonstrate capabilities in learning, adapting, and collaborating (Kaplan et al., 2021). This evolution in trust requires AI systems to demonstrate not only technical proficiency but also adherence to ethical standards, legal compliance, and socially responsible behavior. As AI becomes more integrated into daily life and crucial decision making tasks, establishing trust in AI is essential. This necessitates a focus on enhancing transparency, explainability, fairness, and robustness within AI systems, which is also central to trustworthy AI.

While AI represents a new generation of automation technology with unique characteristics, research on trust in earlier forms of automation remains relevant. This historical perspective can inform the development of trust in AI by highlighting the important trust factors. Incorporating lessons from the past, the transition to AI additionally demands a renewed focus on ethical considerations, transparency, and user engagement to foster a deeper and more comprehensive trust.

3.5.3 A framework of trust in automation

Early theoretical models of trust in automation focused on attributes related to automation's performance or competence, such as reliability, robustness, capability, and predictability (Sheridan, 1988; Malle and Ullman, 2021). In these models, trust was primarily grounded in the systems' technical performance and their ability to meet user expectations reliably. Trust varied directly with the system's demonstrated competence in executing tasks (Malle and Ullman, 2021). However, the progress of AI has broadened the scope of research on trust determinants to include considerations of automation's inferred intentions and the ethical implications of its actions (Malle and Ullman, 2021).

Individual differences are critical to human-automation trust. Some individuals apply frequently the *machine heuristic*, which is similar to trust propensity and represents the tendency to perceive automation as safer and more trustworthy than humans (Sundar and Kim, 2019). Moreover, an individual's self-efficacy in using automation—the confidence in their ability to effectively utilize automation technologies—also plays a crucial role in shaping trust (Kraus et al., 2020); a higher sense of self-efficacy correlates with greater trust and willingness to use automated systems (Latikka et al., 2019). Furthermore, the degrees of familiarity and understanding of automated systems contribute to a more accurate evaluation of these systems' competence, promoting a well-calibrated trust (Lee and See, 2004). Trust in automation, therefore, is a dynamic process, continuously recalibrated with the input of new information, knowledge, and experience (Muir, 1987).

Initial research on trust in automation adopted frameworks from interpersonal trust studies, positing a similar psychological structure in them (Muir, 1987). In practical research, factors affecting human-automation trust can also be categorized into trustor (human factors), trustee (automation factors), and the interaction context. This tripartite framework has been validated across various studies, affirming its applicability in understanding trust dynamics (Hancock et al., 2011; Hoff and Bashir, 2015; Drnec et al., 2016). Automation technology has evolved into the era of AI, inheriting characteristics of traditional automation while also exhibiting new features such as learning capabilities and adaptability. Factors influencing trust in AI based on this three-dimension framework are analyzed in detail in the next section from a sociocognitive perspective.

4 A three-dimension framework of trust in AI

AI, nested within the broad category of automation technology, benefits from existing trust research in automation, despite its unique characteristics. The algorithmic black-box nature of AI poses challenges in understanding its operational mechanisms, and its ability for autonomous learning compounds the difficulty in predicting its behavior. According to theories of technology acceptance, trust plays a pivotal role in the development and adoption of AI (Siau and Wang, 2018). Moreover, it is crucial for both enduring human collaborations and effective cooperation with AI (Xie et al., 2019). Furthermore, AI's limitations in understanding human intentions, preferences, and values present additional challenges (Dafoe et al., 2021). Thus, research on trust in AI can guide the development of trustworthy AI, promote its acceptance and human interaction, and reduce the risks of misuse and disuse.

It is evident from our above review that trust, whether in interpersonal relationships or human-automation interactions, operates within a dyadic framework against an interactive context. Kaplan et al. (2021) also validated a similar framework through a meta-analysis of trust in AI, suggesting that the antecedents influencing trust in AI can be classified into three categories: human-related, AI-related, and context-related. In the following, we review factors influencing trust in AI related to these three dimensions, and Figure 1 shows what these factors are and to which dimension each of them belongs.

4.1 Factors related to the trustor

4.1.1 Demographic variables

The impacts of demographic variables on trust in AI are complicated. In a recent worldwide survey study conducted in 17 countries, Gillespie et al. (2023) found that gender differences in trust toward AI were generally minimal, with notable exceptions in a few countries (i.e., the United States, Singapore, Israel, and South Korea). Regarding age, while a trend of greater trust among the younger generation was prevalent, this pattern was reversed in China and South Korea, where the older population demonstrated higher levels of trust in AI. Additionally, the data indicated that individuals possessing university-level education or occupying managerial roles tended to exhibit higher trust in AI, pointing to the significant roles of educational background and professional status in shaping trust dynamics. The study further showed pronounced cross-country variations, identifying a tendency for placing more trust in AI in economically developing nations, notably China and India. Previous research has also found that culture and social groups can influence trust in AI (Kaplan et al., 2021; Lee and Rich, 2021). Therefore, the impact of demographic variables on trust in AI may be profoundly influenced by socio-cultural factors.

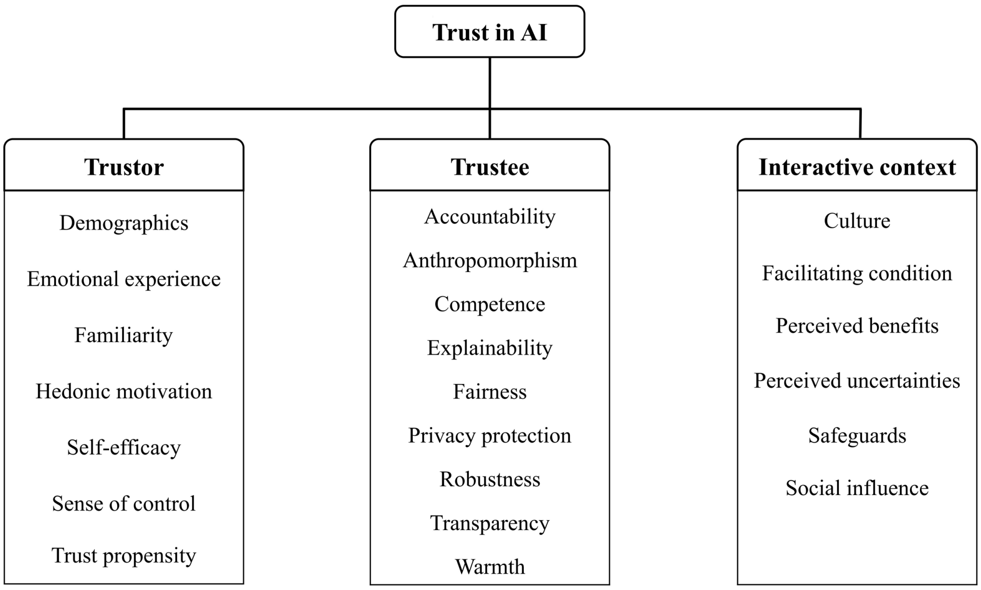

FIGURE 1
A three-dimension framework of trust that specifies the critical factors in each dimension that can affect trust in AI.

4.1.2 Familiarity and self-efficacy

An individual's familiarity with AI, rooted in their knowledge and prior interactive experience, plays a pivotal role in trust formation (Gefen et al., 2003). Such familiarity not only reduces cognitive complexity by supplying essential background information and a cognitive framework, but also enables the formation of concrete expectations regarding the AI's future behavior (Gefen, 2000). Additionally, a deeper understanding of AI can reduce the perceived risks and uncertainties associated with its use (Lu et al., 2019). Relatedly, AI use self-efficacy, or individuals' confidence in their ability to effectively use AI, significantly influences acceptance and trust (REF). Familiarity and self-efficacy are both related to past interactive experience with AI, and both are positively correlated with a precise grasp of AI, thereby facilitating appropriate trust in AI.

4.1.3 Hedonic motivation and emotional experience

Hedonic motivation, an intrinsic form of motivation, can lead people to use AI in pursuit of enjoyment. Venkatesh et al. (2012) recognized this by incorporating hedonic motivation into the expanded Unified Theory of Acceptance and Use of Technology (UTAUT) model as a key determinant of technology acceptance. This form of motivation is instrumental in increasing user satisfaction and enjoyment derived from AI, thereby positively influencing their attitudes toward technology and enhancing their intention to use it (Gursoy et al., 2019).

Emotional experience in the context of AI refers to the sense of security and comfort users feel when relying on AI, often described as emotional trust. It can reduce people's perception of uncertainties and risks, and thus increase trust in AI. The acceptance and utilization of AI are guided by both cognitive judgments and affective responses. As such, it is crucial for trust research in AI to address both the cognitive and the emotional components (Gursoy et al., 2019). Specifically, emotional experience has been shown to directly impact the willingness to adopt AI-based recommendation systems, as seen in the context of travel planning (Shi et al., 2020).

4.1.4 Sense of control

Sense of control represents individuals' perception of their ability to monitor and influence AI decision-making processes. Dietvorst et al. (2018) found that algorithm aversion decreased when participants were allowed to adjust the outcomes of an imperfect algorithm, even if the adjustments were minimal. This finding underscores the importance of a sense of control in enhancing user satisfaction and trust in algorithms that are fundamental components of AI. Aoki (2021) found that AI-assisted nursing care plans that explicitly informed individuals that humans retained control over the decision-making processes significantly boosted trust in AI, compared to those who were not provided with this information. This highlights the importance of communicating human oversight in AI applications to enhance public trust. Similarly, Jutzi et al. (2020) found a favorable attitude toward AI in medical diagnosis when AI acted in a supportive capacity, reinforcing the value of positioning AI as an adjunct to human expertise.

4.1.5 Trust propensity

The propensity to trust refers to stable internal psychological factors affecting an individual's inclination to trust, applicable to both people and technology. Research indicates that individuals with high trust propensity are more inclined to place trust in others, including strangers, and hold a general belief in the beneficial potential of technology (Brown et al., 2004). This tendency enables them to rely on technological systems without extensive evidence of their reliability. Attitudes toward new technologies vary significantly; some individuals readily adopt new technologies, while others exhibit skepticism or caution initially. This variation extends to AI, where trust propensity influences acceptance levels (Chi et al., 2021). Furthermore, trust propensity may intersect with personality traits. For instance, individuals experiencing loneliness may show lower trust in AI, whereas those with a penchant for innovation are more likely to trust AI (Kaplan et al., 2021).

4.2 Factors related to the trustee

4.2.1 Accountability

Because of the complexity and potential wide-ranging impacts of AI, accountability is a key factor in establishing public trust in AI. Fears that AI cannot be held responsible hinder trust in AI. Therefore, people need assurance that clear processes exist to handle AI issues and that specific parties, like developers, providers, or regulators, are accountable.

When people think that AI cannot be held accountable, they are less willing to let AI make decisions and tend to blame it less. Research has found that in the service industry when service providers make mistakes that result in customer losses, participants believe that the robot responsible for the mistake bears less responsibility compared to a human service provider, and the service-providing company should bear more responsibility (Leo and Huh, 2020). This occurs probably because people perceive that robots have poorer controllability over tasks. People are reluctant to allow AI to make moral decisions because AI is perceived to lack mind perception (Bigman and Gray, 2018). Bigman et al. (2023) found that algorithmic discrimination elicits less anger, with people showing less moral outrage toward algorithmic (as opposed to human) discrimination and being less inclined to blame the organization, but it does lower the evaluation of the company. This might be because people perceive algorithms as lacking prejudicial motivation.

4.2.2 Anthropomorphism

Anthropomorphism, the tendency to ascribe human-like qualities to non-human entities such as computers and robots, significantly affects individuals' trust in these technologies (Bartneck et al., 2009). Cominelli et al. (2021) found that robots perceived as highly human-like are more likely to be trusted by individuals. Beyond physical appearance and vocal cues, emotional expression is a crucial aspect of anthropomorphism. Troshani et al. (2021) found that robots exhibiting positive emotions are more likely to receive increased trust and investment from people. Similarly, Li and Sung (2021) demonstrated through a network questionnaire that anthropomorphized AI correlates with more positive attitudes toward the technology. Experimental studies have corroborated these findings, suggesting that psychological distance plays a mediating role in how anthropomorphism influences perceptions of AI (Li and Sung, 2021).

4.2.3 Competence and warmth

The key to evaluating trustworthy AI is whether AI does what it claims to do (Schwartz et al., 2022). The claims of AI can be analyzed from two perspectives: one is whether it fulfills the functional requirements of its users, and the other is whether it demonstrates good intentions toward its users. This directly corresponds to the perceptions of competence and warmth of AI.

In both interpersonal trust and human-automation trust, competence and warmth are pivotal in shaping perceptions of trustworthiness (Kulms and Kopp, 2018). The stereotype content model (SCM) posits that stereotypes and interpersonal perceptions of a group are formed along two dimensions: warmth and competence (Fiske et al., 2002). Warmth reflects how one perceives the intentions (positive or negative) of others, while competence assesses the perceived ability of others to fulfill those intentions.

Trust and stereotype share a foundational link through the attitudes and beliefs individuals hold toward others (Kong, 2018). Moreover, in Mayer et al.'s (1995) trust model, the trustworthiness dimensions of ability and benevolence align closely with the competence and warmth dimensions in the SCM, respectively. Therefore, warmth and competence may be two core components of trustworthiness affecting interpersonal trust. The significance of these two dimensions is evident in their substantial influence on individuals' evaluations and behaviors toward others (Mayer et al., 1995; Fiske, 2012).

Competence is the key factor influencing trust in automation (Drnec et al., 2016). When users observe errors in the automated system, their trust in it decreases, leading them to monitor the system more closely (Muir and Moray, 1996). However, an AI agent that is competent but not warm might not be trusted, because, in certain situations, intention is a crucial influencing factor of trust (Gilad et al., 2021). For instance, although users may recognize the technical proficiency of autonomous vehicles (AVs) in navigating complex environments, concerns that AVs may compromise safety for speed or prioritize self-preservation in emergencies can undermine trust (Xie et al., 2019). Thus, AI needs to demonstrate good intentions to build trust. This is exemplified by AI's social intelligence, such as understanding and responding to user emotions, which significantly bolsters trust in conversational agents (Rheu et al., 2021). Moreover, trust in AI is generally lower in domains traditionally dominated by human expertise, potentially due to concerns about the intentions of AI (Lee, 2018).

4.2.4 Privacy protection

The rapid development of AI, facilitated greatly by network technology, raises privacy concerns, especially when third parties access data through networks without user consent, risking privacy misuse (Featherman and Pavlou, 2003). Additionally, while AI's ability to tailor services to individual needs can enhance user satisfaction, this often requires accessing personal information, thus creating a dilemma between personalization benefits and privacy risks (Guo et al., 2016). Network technologies have amplified privacy risks, resulting in individuals losing control over the flow of their private data. As a result, privacy concerns play a crucial role in establishing online trust, and internet users are highly concerned about websites' privacy policies, actively seeking clues to ensure the protection of their personal information (Ang et al., 2001). Research has found that providing adequate privacy protection measures directly influences people's trust in AI and their willingness to use it (Vimalkumar et al., 2021; Liu and Tao, 2022).

4.2.5 Robustness and fairness

Sheridan (1988) argued that robustness should be an important determinant of trust. The robustness of AI refers to the reliability and consistency of its operations and results, including its performance under diverse and unexpected conditions (High-Level Expert Group on Artificial Intelligence (AI HLEG), 2019). The fairness of AI involves treating all users equitably, making unbiased decisions, and not discriminating against any group (Shin and Park, 2019). Robustness is a key factor influencing trust in AI. Rempel et al. (1985) identified three components of trust from a dynamic perspective, including predictability (the consistency of actions over time), dependability (reliability based on past experience), and faith (belief in future behavior). Based on the definitions, these components also correspond to the formation of the perception of robustness. Compared to trust in humans, building trust in AI takes more time; moreover, when AI encounters problems, the loss of trust in it happens more rapidly (Dzindolet et al., 2003). Furthermore, the simpler the task in which the error occurs, the greater the loss of trust (Madhavan et al., 2006).

Because robustness and fairness are vulnerable to data bias, from both theoretical and practical standpoints, these two factors are closely related. Robustness serves as a crucial foundation for fairness, with the presence of discrimination and bias often signaling a lack of robustness. For example, the training data used for developing large language models often contain biases, and research has found that ChatGPT replicates gender biases in reference letters written for hypothetical employees (Wan et al., 2023). Such disparities underscore the importance of aligning AI with human values, as perceived fairness significantly influences users' trust in AI technologies (Angerschmid et al., 2022).

4.2.6 Transparency and explainability

Users need to understand how and why AI makes specific decisions, which corresponds, respectively, to transparency and explainability, before trusting in it. However, this is not an easy task for AI practitioners and stakeholders. Because of the mechanisms of algorithms, particularly the opacity of neural networks, it is difficult for humans to fully comprehend their decision-making process (Shin and Park, 2019).

Transparent AI models with clear explanations of their decision-making processes help users gain confidence in the system's capabilities and accuracy. Moreover, transparency in AI design and implementation helps identify potential sources of bias, allowing developers to address these issues and ensure the AI system treats all users fairly. The concepts of transparency and explainability are deeply interconnected; explainability, in particular, plays a crucial role in reducing users' perceived risks associated with AI systems (Qin et al., 2020). Additionally, providing reasonable explanations after AI errors can restore people's trust in AI (Angerschmid et al., 2022).

That said, the impact of transparency and explainability on trust in AI shows mixed results. Leichtmann et al. (2022) found that displaying AI's decision-making process through graphical and textual information enhances users' trust in the AI program. However, Wright et al. (2020) found no significant difference in trust levels attributed to varying degrees of transparency in simulated military tasks for target detection. Furthermore, in a task of using AI assistance to rate movies, Schmidt et al. (2020) observed that increased transparency in AI-assisted movie rating tasks paradoxically reduced user trust.

Therefore, the relationship between transparency and trust in AI is intricate. Appropriate levels of transparency and explainability can enhance people's trust in AI, but excessive information might be confusing (Kizilcec, 2016), thereby reducing their trust in AI. The absence of clear operational definitions for AI's transparency and explainability complicates the determination of the optimal transparency levels that effectively build trustworthy AI. In general, lack of transparency indeed hurts trust in AI, but high levels of transparency do not necessarily lead to good results.

4.3 Factors related to the interactive context of trust

4.3.1 Perceived uncertainties and benefits

AI is surrounded by various unknowns, including ethical and legal uncertainties, that are critical evaluations of the interactive environment. Lockey et al. (2020) emphasized uncertainties as a key factor influencing trust in AI. Similarly, Jing et al. (2020) conducted a literature review and discovered a negative correlation between perceived uncertainties and risks with the acceptance of autonomous driving. Furthermore, perceived uncertainties in the application of AI vary across different applications, particularly pronounced in medical expert systems and vehicles (Yang and Wibowo, 2022).

Perceived benefits, such as time savings, cost reductions, and increased convenience, highlight the recognized advantages of using AI (Kim et al., 2008). Liu and Tao (2022) found that perceived benefits, such as usefulness, could facilitate the use of smart healthcare services. Although perceived benefits can be viewed as characteristics of AI the trustee, they can also be highly socially dependent, mainly because the impacts of these benefits are not uniform across all users: For instance, while AI applications may enhance work efficiency for some, they could pose a risk of unemployment for others (Pereira et al., 2023). Therefore, perceived benefits are intricately linked to the social division of labor, underscoring their importance within the broader interactive context of AI usage.

4.3.2 Safeguards

Drawing from the concept of institution-based trust, safeguards are understood as the belief in existing institutional conditions that promote responsible and ethical AI usage (McKnight et al., 1998). Because AI is perceived as lacking agency and cannot be held accountable for its actions (Bigman and Gray, 2018), safeguards play a crucial role in ensuring human trust in AI.

Lockey et al. (2020), however, found that a mere 19–21% of Australians considered the current safety measures adequate for AI's safe application, underscoring a significant trust gap. Their analysis further showed that perception of these safeguards was a strong predictor of trust in AI. In today's AI landscape, establishing legal frameworks to protect human rights and interests is crucial for fostering trust. A prime example is the European Union's General Data Protection Regulation (GDPR). Enacted in 2018, GDPR introduces stringent privacy protections and sets clear standards for algorithmic transparency and accountability (Felzmann et al., 2019).

4.3.3 Social influence and facilitating condition

Social influence is defined by the extent to which an individual perceives endorsement of specific behaviors by their social network, including encouragement from influential members to adopt new technologies (Gursoy et al., 2019). It is a crucial construct in the UTAUT (Venkatesh et al., 2003). Social influence theory posits that individuals tend to conform to the norms and beliefs of their social network (Shi et al., 2020). When individuals perceive that the use of AI is socially acceptable, they are more likely to experience positive emotions toward it, leading to an increase in their emotional trust in AI.

Facilitating condition is another critical variable in the UTAUT model, referring to the extent to which individuals perceive organizational, group, or infrastructural support in using AI (Venkatesh et al., 2003). Chi et al. (2021) found that facilitating robot-use conditions could improve users' trust in social robots in service scenarios.

4.3.4 Culture

Cultural factors can significantly influence trust and acceptance of AI. For instance, cultures with high uncertainty avoidance are more inclined to trust and depend on AI (Kaplan et al., 2021), and the level of trust in AI also varies between individualistic and collectivistic cultures (Chi et al., 2023). Moreover, cultural influences may interact with economic factors to affect AI trust. Gillespie et al. (2023) found that individuals in the emerging economies, such as Brazil, India, China, and South Africa, exhibit higher levels of trust, in comparison to developed nations, such as the United Kingdom, Australia, Japan, and France. Furthermore, the impact of culture on AI trust can be mediated through social influence, highlighting the importance of social norms (Chi et al., 2023).

5 Implications for enhancing trustworthy AI

While intention is pivotal in interpersonal trust, competence is paramount in human-automation trust. Nonetheless, research on trust in AI encompasses both competence and intention, indicating that AI is perceived through a combination of human-like and automated characteristics, reflecting a sociocognitive perspective on trust in AI (Kulms and Kopp, 2018). Understanding how trust in AI forms from this perspective and integrating the resulting knowledge into the design and applications of AI systems will be critical to foster their effective use and the collaboration between humans and AI.

The proposed three-dimension framework of trust in AI not only encompasses the desired characteristics of AI but also emphasizes enhancing AI literacy among users and refining the interactive context. The framework highlights users' expectations of AI and can help developers and managers grasp user concerns and needs. In addition, given the imperative to reduce perceived uncertainties associated with AI, it becomes critical to address concerns related to privacy protection in AI applications, ensure accountability, and meet the demand for enhanced safeguard measures.

The three-dimension framework also provides a solid foundation for developing ethical standards and policies that can enhance trustworthy AI. In social psychology, competence and warmth are critical for assessing trustworthiness. These dimensions are equally vital in evaluating AI. Specifically, robustness and safety illustrate competence, whereas fairness and privacy protection embody warmth. Thus, in formulating ethical standards for trustworthy AI, we would recommend focusing on the key dimensions of competence and warmth. For example, in developing and deploying AI applications, it is critical to conduct an ethical evaluation based on their competence and warmth. This evaluation ensures that the applications are functionally effective and possess benevolent intentions toward humanity. In addition, as AI technology advances, its potential to infringe upon human rights intensifies, underscoring the increasing importance of evaluating its warmth.

Recognizing how individual characteristics influence trust in AI can guide the development of personalized and adaptive AI interaction strategies. These strategies, tailored to meet the specific needs and preferences of diverse users, can foster a sustained and appropriate trust in AI. While some individuals may place excessive trust in AI because of a high trust propensity, others, hindered by limited understanding of AI and a lower sense of self-efficacy, may demonstrate a lack of trust. Lockey et al. (2020) discovered a widespread desire among individuals to learn more about AI. Therefore, in developing trustworthy AI, it is crucial to acknowledge the varying levels of trust people have toward AI and to devise effective communication strategies, engaging in AI education to bridge this knowledge gap.

Moreover, hedonic motivation plays a critical role in shaping trust in AI, with the potential to cause users to overtrust AI systems. For example, algorithms behind short video apps often leverage this motivation, leading to excessive requests for user data (Trifiro, 2023). Despite users' general inclination to protect their privacy, they often adopt a pragmatic approach toward privacy protection, a discrepancy referred to as the *privacy paradox* (Sundar and Kim, 2019). Therefore, it is essential to be vigilant against the overtrust in AI that may result from an excessive pursuit of enjoyment and entertainment.

Furthermore, power asymmetry often results in trust asymmetry. The prevailing trust in AI serves as a pertinent example of this asymmetry, where interactions with AI-driven technologies may engender a perceived sense of power or dominance among users. Such perceptions significantly influence the dynamics of trust in AI (Fast and Schroeder, 2020). Consequently, the influence of this sense of power on human interactions with AI necessitates further investigation.

Gaining insight into the factors affecting AI's trustworthiness enables a more sophisticated approach to identifying and managing the inherent risks associated with its application. Notably, anthropomorphism, the attribution of human-like qualities to AI, significantly influences users' emotional trust, potentially enhancing AI's acceptance and trustworthiness (Glikson and Woolley, 2020). Anthropomorphized AI might be more readily accepted and trusted by users, yet this could also mask its inherent limitations and potential risks. Further, attributing human traits to AI can lead to unrealistic expectations about its capabilities, including agency and moral judgment, thereby fostering misconceptions about its competence. Thus, cultivating trustworthy AI requires ensuring that users possess an accurate understanding of AI's anthropomorphic features.

Policymakers focused on trustworthy AI must recognize the significant influence of social, organizational, and institutional factors in shaping AI perceptions within the interactive context of trust. The mass media plays a pivotal role in influencing public attitudes toward AI, either by highlighting uncertainties or by raising awareness of new technological advancements. A series of studies have shown the significant role of media in promoting emerging technologies (Du et al., 2021). Media headlines can influence people's emotional responses, thereby affecting their willingness to adopt technology (Anania et al., 2018). Mass media can also influence trust in AI by impacting social influence and self-efficacy. Given these dynamics, regulating mass media to ensure accurate representation of AI is crucial. Policymakers should additionally prioritize the establishment of clear laws and regulations, define responsibilities for AI failures, and engage in transparent communication with the public to mitigate perceived uncertainties.

For example, trust in autonomous vehicles is dynamic (Luo et al., 2020) and easily swayed by mass media (Lee et al., 2022). Furthermore, media portrayals often lack objectivity, with companies overstating autonomy levels in promotions, whereas media primarily reports on accidents. Therefore, ensuring balanced and factual media representations is essential to foster an environment where people can develop informed trust in autonomous vehicles. Moreover, implementing sensible legislation and regulations, as well as clarifying responsibility in accidents involving autonomous vehicles, is vital for public endorsement.

High-Level Expert Group on Artificial Intelligence (AI HLEG) (2019) delineated seven crucial requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing, and accountability. The factors influencing trust in AI as we have reviewed (see Figure 1) are generally consistent with these requirements. High-Level Expert Group on Artificial Intelligence (AI HLEG) (2019) also proposes communication, education, and training as key non-technical strategies for promoting trustworthy AI, again consistent with our recommendations derived from the literature.

Generative AI has emerged as the most noteworthy development in AI technologies in recent years, with products such as GPT and Sora showing impressive capabilities in content generation and analysis (Yang et al., 2024). For instance, videos created by Sora can be indistinguishably realistic. Even large language models are being utilized to explain other AI models, enhancing AI's explainability (Bills et al., 2023). As AI's capabilities grow, so does its impact on society, including potential negative effects, such as the ease of generating fraudulent content through generative AI. Concurrently, governments worldwide are introducing laws and regulations to guide AI development responsibly. On March 13, 2024, the European Union passed *The AI Act*, the world's first comprehensive regulatory framework for AI (European Parliament, 2024). It categorizes AI usage by risk levels, banning its use in certain areas such as social scoring systems and the remote collection of biometric information and highlighting the importance of fairness and privacy protection. While the competence of AI is advancing, skepticism about its warmth also grows. Simultaneously, the emphasis on its warmth and the need for safeguards will increase.

Overall, we review factors influencing trust formation from the user's perspective via a three-dimension model of trust in AI. The framework, with its detailed examination of factors impacting the trustor, the trustee, and their interaction context, is instrumental in guiding the creation of targeted educational and training programs that are essential for enabling users to understand and engage with AI more effectively. Furthermore, trustworthy AI could benefit from the adoption of trust measurement methods to assess the effectiveness of these initiatives. These assessments should include both subjective self-report methods and objective indicators of engagement with AI technologies, including reliance, compliance, and decision-making behavior and time.

6 Summary and conclusion

This article provides a comprehensive review and analysis of factors influencing trust in AI and offers insights and suggestions on the development of trustworthy AI. The three-dimension framework of trust is applicable for understanding trust in interpersonal relationships, human-automation interactions, and human-AI systems. The framework can also help understand user needs and concerns, guide the refinement of AI system designs, and aid in the making of policies and guidelines on trustworthy AI. All of these shall lead to AI systems that are more trustworthy, increasing the likelihood for people to accept, adopt, and use them properly.

Author contributions

YL: Writing – original draft. BW: Writing – original draft. YH: Writing – original draft. SL: Writing – review & editing.

Funding

The authors declare financial support was received for the research and publication of this article. This work was supported by research grants from the National Natural Science Foundation of China [grant number 32171074] and the Institute of Psychology, Chinese Academy of Sciences [grant number E1CX0230] awarded to SL.

Acknowledgments

We thank members of the Risk and Uncertainty Management Lab for their helpful comments and suggestions during the study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ajenaghughrure, I. B., Sousa, S. D. C., and Lamas, D. (2020). Measuring trust with psychophysiological signals: a systematic mapping study of approaches used. *Multimodal Technol. Interact.* 4:63. doi: 10.3390/mti4030063

Anania, E. C., Rice, S., Walters, N. W., Pierce, M., Winter, S. R., and Milner, M. N. (2018). The effects of positive and negative information on consumers' willingness to ride in a driverless vehicle. *Transp. Policy* 72, 218–224. doi: 10.1016/j.tranpol.2018.04.002

Ang, L., Dubelaar, C., and Lee, B.-C. (2001). "To trust or not to trust? A model of internet trust from the customer's point of view" in *BLED 2001 proceedings* (Bled), 40–52.

Angerschmid, A., Theuermann, K., Holzinger, A., Chen, F., and Zhou, J. (2022). "Effects of fairness and explanation on Trust in Ethical AI" in *International cross-domain conference for machine learning and knowledge extraction*, 51–67.

Aoki, N. (2021). The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: evidence from an online experiment. *Comput. Hum. Behav.* 114:106572. doi: 10.1016/j.chb.2020.106572

Ayoub, J., Avetisian, L., Yang, X. J., and Zhou, F. (2023). Real-time trust prediction in conditionally automated driving using physiological measures. *IEEE Trans. Intell. Transp. Syst.* 24, 14642–14650. doi: 10.1109/TITS.2023.3295783

Baer, M. D., Matta, F. K., Kim, J. K., Welsh, D. T., and Garud, N. (2018). It's not you, it's them: social influences on trust propensity and trust dynamics. *Pers. Psychol.* 71, 423–455. doi: 10.1111/peps.12265

Bailey, P. E., and Leon, T. (2019). A systematic review and meta-analysis of age-related differences in trust. *Psychol. Aging* 34, 674–685. doi: 10.1037/pag0000368

Bailey, P. E., Slessor, G., Rieger, M., Rendell, P. G., Moustafa, A. A., and Ruffman, T. (2015). Trust and trustworthiness in young and older adults. *Psychol. Aging* 30, 977–986. doi: 10.1037/a0039736

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* 1, 71–81. doi: 10.1007/s12369-008-0001-3

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games Econ. Behav.* 10, 122–142. doi: 10.1006/game.1995.1027

Bigman, Y. E., and Gray, K. (2018). People are averse to machines making moral decisions. *Cognition* 181, 21–34. doi: 10.1016/j.cognition.2018.08.003

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., and Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *J. Exp. Psychol. Gen.* 152, 4–27. doi: 10.1037/xge0001250

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., et al. (2023). Language models can explain neurons in language models. *OpenAI*. Available at: https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html (Accessed April 2, 2024).

Brown, H. G., Poole, M. S., and Rodgers, T. L. (2004). Interpersonal traits, complementarity, and trust in virtual collaboration. *J. Manag. Inf. Syst.* 20, 115–138. doi: 10.1080/07421222.2004.11045785

Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *J. Behav. Decis. Mak.* 33, 220–239. doi: 10.1002/bdm.2155

Chen, Y., He, H., Lin, W., Yang, J., Tan, S., Tao, W., et al. (2023). The connectome-based prediction of trust propensity in older adults: a resting-state functional magnetic resonance imaging study. *Hum. Brain Mapp.* 44, 4337–4351. doi: 10.1002/hbm.26385

Chi, O. H., Chi, C. G., Gursoy, D., and Nunkoo, R. (2023). Customers' acceptance of artificially intelligent service robots: the influence of trust and culture. *Int. J. Inf. Manag.* 70:102623. doi: 10.1016/j.ijinfomgt.2023.102623

Chi, O. H., Jia, S., Li, Y., and Gursoy, D. (2021). Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Comput. Hum. Behav.* 118:106700. doi: 10.1016/j.chb.2021.106700

Cominelli, L., Feri, F., Garofalo, R., Giannetti, C., Meléndez-Jiménez, M. A., Greco, A., et al. (2021). Promises and trust in human–robot interaction. *Sci. Rep.* 11:9687. doi: 10.1038/s41598-021-88622-9

Cugurullo, F. (2020). Urban artificial intelligence: from automation to autonomy in the smart city. *Front. Sustain. Cities* 2:38. doi: 10.3389/frsc.2020.00038

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., and Graepel, T. (2021). Cooperative AI: machines must learn to find common ground. *Nature* 593, 33–36. doi: 10.1038/d41586-021-01170-0

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. *Manag. Sci.* 64, 1155–1170. doi: 10.1287/mnsc.2016.2643

Dietz, G. (2011). Going back to the source: why do people trust each other? *J. Trust Res.* 1, 215–222. doi: 10.1080/21515581.2011.603514

Dirks, K. T., and Ferrin, D. L. (2002). Trust in leadership: Meta-analytic findings and implications for research and practice. *J. Appl. Psychol.* 87, 611–628. doi: 10.1037/0021-9010.87.4.611

Dittrich, M. (2015). Gender differences in trust and reciprocity: evidence from a large-scale experiment with heterogeneous subjects. *Appl. Econ.* 47, 3825–3838. doi: 10.1080/00036846.2015.1019036

Drnec, K., Marathe, A. R., Lukos, J. R., and Metcalfe, J. S. (2016). From trust in automation to decision neuroscience: applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction. *Front. Hum. Neurosci.* 10:290. doi: 10.3389/fnhum.2016.00290

Du, H., Zhu, G., and Zheng, J. (2021). Why travelers trust and accept self-driving cars: an empirical study. *Travel Behav. Soc.* 22, 1–9. doi: 10.1016/j.tbs.2020.06.012

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *Int. J. Hum. Comput. Stud.* 58, 697–718. doi: 10.1016/S1071-5819(03)00038-7

European Parliament (2024). Artificial intelligence act: MEPs adopt landmark law. Available at: https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law (Accessed April 2, 2024).

Evans, A. M., and Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *J. Res. Pers.* 42, 1585–1593. doi: 10.1016/j.jrp.2008.07.011

Fast, N. J., and Schroeder, J. (2020). Power and decision making: new directions for research in the age of artificial intelligence. *Curr. Opin. Psychol.* 33, 172–176. doi: 10.1016/j.copsyc.2019.07.039

Featherman, M. S., and Pavlou, P. A. (2003). Predicting e-services adoption: a perceived risk facets perspective. *Int. J. Hum. Comput. Stud.* 59, 451–474. doi: 10.1016/S1071-5819(03)00111-3

Felzmann, H., Villaronga, E. F., Lutz, C., and Tamò-Larrieux, A. (2019). Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Soc.* 6:205395171986054. doi: 10.1177/2053951719860542

Fiske, S. T. (2012). Warmth and competence: Stereotype content issues for clinicians and researchers. *Can. Psychol.* 53, 14–20. doi: 10.1037/a0026054

Fiske, S. T., Cuddy, A. J. C., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *J. Pers. Soc. Psychol.* 82, 878–902. doi: 10.1037/0022-3514.82.6.878

Frazier, M. L., Johnson, P. D., and Fainshmidt, S. (2013). Development and validation of a propensity to trust scale. *J. Trust Res.* 3, 76–97. doi: 10.1080/21515581.2013.820026

Frischknecht, R. (2021). A social cognition perspective on autonomous technology. *Comput. Hum. Behav.* 122:106815. doi: 10.1016/j.chb.2021.106815

Gefen, D. (2000). E-commerce: the role of familiarity and trust. *Omega* 28, 725–737. doi: 10.1016/S0305-0483(00)00021-9

Gefen, D., Karahanna, E., and Straub, D. W. (2003). Trust and TAM in online shopping: an integrated model. *MIS Q.* 27, 51–90. doi: 10.2307/30036519

Gilad, Z., Amir, O., and Levontin, L. (2021). The effects of warmth and competence perceptions on users' choice of an AI system. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, (pp. 1–13).

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., and Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Comput. Hum. Behav.* 115:106607. doi: 10.1016/j.chb.2020.106607

Gillespie, N., Lockey, S., Curtis, C., Pool, J., and Akbari, A. (2023). Trust in artificial intelligence: a global study. The University of Queensland and KPMG Australia. doi: 10.14264/00d3c94,

Glikson, E., and Woolley, A. W. (2020). Human trust in artificial intelligence: review of empirical research. *Acad. Manag. Ann.* 14, 627–660. doi: 10.5465/annals.2018.0057

Guo, X., Zhang, X., and Sun, Y. (2016). The privacy–personalization paradox in mHealth services acceptance of different age groups. *Electron. Commer. Res. Appl.* 16, 55–65. doi: 10.1016/j.elerap.2015.11.001

Gursoy, D., Chi, O. H., Lu, L., and Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *Int. J. Inf. Manag.* 49, 157–169. doi: 10.1016/j.ijinfomgt.2019.03.008

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Factors* 53, 517–527. doi: 10.1177/0018720811417254

High-Level Expert Group on Artificial Intelligence (AI HLEG). (2019). Ethics guidelines for trustworthy AI [European Commission report]. Available at: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1.

Hoff, K. A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Factors* 57, 407–434. doi: 10.1177/0018720814547570

Ito, T. (2022). Effects of general trust as a personality trait on willingness to communicate in a second language. *Personal. Individ. Differ.* 185:111286. doi: 10.1016/j.paid.2021.111286

Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* 4, 53–71. doi: 10.1207/S15327566IJCE0401_04

Jing, P., Xu, G., Chen, Y., Shi, Y., and Zhan, F. (2020). The determinants behind the acceptance of autonomous vehicles: a systematic review. *Sustain. For.* 12:1719. doi: 10.3390/su12051719

Jones, S. L., and Shah, P. P. (2016). Diagnosing the locus of trust: a temporal perspective for trustor, trustee, and dyadic influences on perceived trustworthiness. *J. Appl. Psychol.* 101, 392–414. doi: 10.1037/apl0000041

Jutzi, T. B., Krieghoff-Henning, E. I., Holland-Letz, T., Utikal, J. S., Hauschild, A., Schadendorf, D., et al. (2020). Artificial intelligence in skin cancer diagnostics: the patients' perspective. *Front. Med.* 7:233. doi: 10.3389/fmed.2020.00233

Kaplan, A. D., Kessler, T. T., Brill, J. C., and Hancock, P. A. (2021). Trust in artificial intelligence: Meta-analytic findings. *Hum. Factors* 65, 337–359. doi: 10.1177/00187208211013988

Khan, G. F., Swar, B., and Lee, S. K. (2014). Social media risks and benefits: a public sector perspective. *Soc. Sci. Comput. Rev.* 32, 606–627. doi: 10.1177/0894439314524701

Khawaji, A., Zhou, J., Chen, F., and Marcus, N. (2015). Using galvanic skin response (GSR) to measure trust and cognitive load in the text-chat environment. In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (pp. 1989–1994).

Kim, D. J., Ferrin, D. L., and Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: the role of trust, perceived risk, and their antecedents. *Decis. Support. Syst.* 44, 544–564. doi: 10.1016/j.dss.2007.07.001

Kim, Y., and Sundar, S. S. (2012). Anthropomorphism of computers: is it mindful or mindless? *Comput. Hum. Behav.* 28, 241–250. doi: 10.1016/j.chb.2011.09.006

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 2390–2395). doi: 10.1145/2858036.2858402

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., and Shaw, T. H. (2021). Measurement of trust in automation: a narrative review and reference guide. *Front. Psychol.* 12:604977. doi: 10.3389/fpsyg.2021.604977

Kong, D. T. (2018). Trust toward a group of strangers as a function of stereotype-based social identification. *Personal. Individ. Differ.* 120, 265–270. doi: 10.1016/j.paid.2017.03.031

Korsgaard, M. A., Brower, H. H., and Lester, S. W. (2015). It isn't always mutual: a critical review of dyadic trust. *J. Manag.* 41, 47–70. doi: 10.1177/0149206314547521

Kraus, J., Scholz, D., Messner, E.-M., Messner, M., and Baumann, M. (2020). Scared to trust? – predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Front. Psychol.* 10:2917. doi: 10.3389/fpsyg.2019.02917

Kulms, P., and Kopp, S. (2018). A social cognition perspective on human-computer trust: the effect of perceived warmth and competence on trust in decision-making with computers. *Front. Digit. Humanit.* 5:14. doi: 10.3389/fdigh.2018.00014

Latikka, R., Turja, T., and Oksanen, A. (2019). Self-efficacy and acceptance of robots. *Comput. Hum. Behav.* 93, 157–163. doi: 10.1016/j.chb.2018.12.017

Lee, M. K. (2018). Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. *Big Data Soc.* 5:205395171875668. doi: 10.1177/2053951718756684

Lee, J., Baig, F., and Li, X. (2022). Media influence, trust, and the public adoption of automated vehicles. *IEEE Intell. Transp. Syst. Mag.* 14, 174–187. doi: 10.1109/MITS.2021.3082404

Lee, M. K., and Rich, K. (2021). Who is included in human perceptions of AI?: trust and perceived fairness around healthcare AI and cultural mistrust. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–14.

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50_30392

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., and Mara, M. (2022). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Comput. Hum. Behav.* 139:107539. doi: 10.1016/j.chb.2022.107539

Leo, X., and Huh, Y. E. (2020). Who gets the blame for service failures? Attribution of responsibility toward robot versus human service providers and service firms. *Comput. Hum. Behav.* 113:106520. doi: 10.1016/j.chb.2020.106520

Lewandowsky, S., Mundy, M., and Tan, G. P. A. (2000). The dynamics of trust: comparing humans to automation. *J. Exp. Psychol. Appl.* 6, 104–123. doi: 10.1037/1076-898X.6.2.104

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., et al. (2023). Trustworthy AI: from principles to practices. *ACM Comput. Surv.* 55, 177:1–177:46. doi: 10.1145/3555803

Li, X., and Sung, Y. (2021). Anthropomorphism brings us closer: the mediating role of psychological distance in user–AI assistant interactions. *Comput. Hum. Behav.* 118:106680. doi: 10.1016/j.chb.2021.106680

Liu, K., and Tao, D. (2022). The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. *Comput. Hum. Behav.* 127:107026. doi: 10.1016/j.chb.2021.107026

Lockey, S., Gillespie, N., and Curtis, C. (2020). Trust in Artificial Intelligence: Australian insights. The University of Queensland and KPMG. doi: 10.14264/b32f129,

Lu, L., Cai, R., and Gursoy, D. (2019). Developing and validating a service robot integration willingness scale. *Int. J. Hosp. Manag.* 80, 36–51. doi: 10.1016/j.ijhm.2019.01.005

Luo, R., Chu, J., and Yang, X. J. (2020). "Trust dynamics in human-AV (automated vehicle) interaction" in *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, 1–7.

Madhavan, P., Wiegmann, D. A., and Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Hum. Factors* 48, 241–256. doi: 10.1518/001872006777724408

Magd, H., Jonathan, H., Khan, S. A., and El Geddawy, M. (2022). "Artificial intelligence—the driving force of industry 4.0" in *A roadmap for enabling industry 4.0 by artificial intelligence*, 1–15.

Malle, B. F., and Ullman, D. (2021). "A multidimensional conception and measure of human-robot trust" in *Trust in human-robot interaction*. (Cambridge, MA: Academic Press), 3–25.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 709–734. doi: 10.5465/amr.1995.9508080335

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Acad. Manag. J.* 38, 24–59. doi: 10.5465/256727

McKnight, D. H., Choudhury, V., and Kacmar, C. (2002). Developing and validating trust measures for e-commerce: an integrative typology. *Inf. Syst. Res.* 13, 334–359. doi: 10.1287/isre.13.3.334.81

McKnight, D. H., Cummings, L. L., and Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Acad. Manag. Rev.* 23, 473–490. doi: 10.5465/amr.1998.926622

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *Int. J. Man Mach. Stud.* 27, 527–539. doi: 10.1016/S0020-7373(87)80013-5

Muir, B. M., and Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 429–460. doi: 10.1080/00140139608964474

Nass, C., Moon, Y., and Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *J. Appl. Soc. Psychol.* 27, 864–876. doi: 10.1111/j.1559-1816.1997.tb00275.x

National Governance Committee for the New Generation Artificial Intelligence. (2019). Governance principles for the new generation artificial intelligence — developing responsible artificial intelligence [press release]. Available at: https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7_10.html.

Nielsen, Y. A., Thielmann, I., Zettler, I., and Pfattheicher, S. (2021). Sharing money with humans versus computers: on the role of honesty-humility and (non-) social preferences. *Soc. Psychol. Personal. Sci.* 13, 1058–1068. doi: 10.1177/19485506211055622

Parasuraman, R., and Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Hum. Factors* 52, 381–410. doi: 10.1177/0018720810376055

Parasuraman, R., and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Hum. Factors* 39, 230–253. doi: 10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man Cybern. Syst. Hum.* 30, 286–297. doi: 10.1109/3468.844354

Pereira, V., Hadjielias, E., Christofi, M., and Vrontis, D. (2023). A systematic literature review on the impact of artificial intelligence on workplace outcomes: a multi-process perspective. *Hum. Resour. Manag. Rev.* 33:100857. doi: 10.1016/j.hrmr.2021.100857

Qin, F., Li, K., and Yan, J. (2020). Understanding user trust in artificial intelligence-based educational systems: evidence from China. *Br. J. Educ. Technol.* 51, 1693–1710. doi: 10.1111/bjet.12994

Rempel, J. K., Holmes, J. G., and Zanna, M. P. (1985). Trust in close relationships. *J. Pers. Soc. Psychol.* 49, 95–112. doi: 10.1037/0022-3514.49.1.95

Rheu, M., Shin, J. Y., Peng, W., and Huh-Yoo, J. (2021). Systematic review: trust-building factors and implications for conversational agent design. *Int. J. Hum. Comput. Interact.* 37, 81–96. doi: 10.1080/10447318.2020.1807710

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *J. Pers.* 35, 651–665. doi: 10.1111/j.1467-6494.1967.tb01454.x

Russo, P. A., Duradoni, M., and Guazzini, A. (2021). How self-perceived reputation affects fairness towards humans and artificial intelligence. *Comput. Hum. Behav.* 124:106920. doi: 10.1016/j.chb.2021.106920

Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Sci. Eng. Ethics* 26, 2749–2767. doi: 10.1007/s11948-020-00228-y

Schmidt, P., Biessmann, F., and Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *J. Decis. Syst.* 29, 260–278. doi: 10.1080/12460125.2020.1819094

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., and Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. NIST Special Publication 1270. doi: 10.6028/NIST.SP.1270,

Sheridan, T. B. (1988). Trustworthiness of command and control systems. *IFAC Proc. Volumes* 21, 427–431. doi: 10.1016/S1474-6670(17)53945-2

Shi, S., Gong, Y., and Gursoy, D. (2020). Antecedents of trust and adoption intention toward artificially intelligent recommendation systems in travel planning: a heuristic–systematic model. *J. Travel Res.* 60, 1714–1734. doi: 10.1177/0047287520966395

Shin, D., and Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Comput. Hum. Behav.* 98, 277–284. doi: 10.1016/j.chb.2019.04.019

Siau, K., and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Bus. Tech. J.* 31, 47–53.

Sundar, S. S., and Kim, J. (2019). Machine heuristic: when we trust computers more than humans with our personal information. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–9).

Tay, B., Jung, Y., and Park, T. (2014). When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. *Comput. Hum. Behav.* 38, 75–84. doi: 10.1016/j.chb.2014.05.014

The White House. (2023). Fact sheet: president Biden issues executive order on safe, secure, and trustworthy artificial intelligenc. Available at: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/ (Accessed April 2, 2024).

Thielmann, I., Spadaro, G., and Balliet, D. (2020). Personality and prosocial behavior: a theoretical framework and meta-analysis. *Psychol. Bull.* 146, 30–90. doi: 10.1037/bul0000217

Trifiro, B. M. (2023). Breaking your boundaries: how TikTok use impacts privacy concerns among influencers. *Mass Commun. Soc.* 26, 1014–1037. doi: 10.1080/15205436.2022.2149414

Troshani, I., Rao Hill, S., Sherman, C., and Arthur, D. (2021). Do we trust in AI? Role of anthropomorphism and intelligence. *J. Comput. Inf. Syst.* 61, 481–491. doi: 10.1080/08874417.2020.1788473

Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., and Seaborn, K. (2022). Trust in human-AI interaction: scoping out models, measures, and methods. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts (pp. 1–7).

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User acceptance of information technology: toward a unified view. *MIS Q.* 27, 425–478. doi: 10.2307/30036540

Venkatesh, V., Thong, J. Y. L., and Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS Q.* 36:157. doi: 10.2307/41410412

Vimalkumar, M., Sharma, S. K., Singh, J. B., and Dwivedi, Y. K. (2021). 'Okay google, what about my privacy?': User's privacy perceptions and acceptance of voice based digital assistants. *Comput. Hum. Behav.* 120:106763. doi: 10.1016/j.chb.2021.106763

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., and Peng, N. (2023). "Kelly is a warm person, Joseph is a role model". Gender biases in LLM-generated reference letters. *arXiv:2310.09219*. doi: 10.48550/arXiv.2310.09219

Westjohn, S. A., Magnusson, P., Franke, G. R., and Peng, Y. (2022). Trust propensity across cultures: The role of collectivism. *J. Int. Mark.* 30, 1–17. doi: 10.1177/1069031X211036688

Wright, J. L., Chen, J. Y. C., and Lakhmani, S. G. (2020). Agent transparency and reliability in human-robot interaction: the influence on user confidence and perceived reliability. *IEEE Trans. Hum. Mach. Syst.* 50, 254–263. doi: 10.1109/THMS.2019.2925717

Xie, Y., Bodala, I. P., Ong, D. C., Hsu, D., and Soh, H. (2019). Robot capability and intention in trust-based decisions across tasks. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 39–47).

Yang, D., Hu, L., Tian, Y., Li, Z., Kelly, C., Yang, B., et al. (2024). WorldGPT: a Sora-inspired video AI agent as Rich world models from text and image inputs. *arXiv:2403.07944*. doi: 10.48550/arXiv.2403.07944

Yang, R., and Wibowo, S. (2022). User trust in artificial intelligence: a comprehensive conceptual framework. *Electron. Mark.* 32, 2053–2077. doi: 10.1007/s12525-022-00592-6