Human-AI Trust in the Era of Artificial General Intelligence

Authors: Qi Yue, Chen Junting, Qin Shaotian, Du Feng, Qi Yue, Du Feng

Date: 2024-08-02T00:00:00+00:00

Abstract

With technological development, Artificial General Intelligence (AGI) is beginning to emerge, and human-computer interaction as well as human-computer relationships will enter a new era. The trust relationship between humans and artificial intelligence (AI) is also poised to gradually transition from unidirectional human-to-AI trust to mutual trust between humans and AI. Building upon a review of interpersonal trust models in social psychology and human-machine trust models in engineering psychology, this study proposes a dynamic human-AI mutual trust model from an interpersonal trust perspective. This model conceptualizes humans and AI as equal trust-building parties, and constructs a basic theoretical framework for dynamic human-AI mutual trust by integrating influencing factors from both trustor and trustee, outcome feedback, and behavioral adjustment, thereby emphasizing two crucial characteristics of human-AI trust: "mutual trust" in the relational dimension and "dynamic" nature in the temporal dimension. The model incorporates, for the first time, AI's trust in humans and the dynamic interaction process of mutual trust into the analysis, offering a novel theoretical perspective for human-AI trust research. Future research should devote greater attention to how AI's trust in humans is established and maintained, quantitative models of human-AI mutual trust, and human-AI mutual trust in multi-agent interactions.

Full Text

Preamble

Human-AI Mutual Trust in the Era of Artificial General Intelligence

Qi Yue[1,2], Chen Junting[1,2], Qin Shaotian[1,2], Du Feng[3,4]
[1] (Department of Psychology, Renmin University of China, Beijing 100872, China)
[2] (Laboratory of the Department of Psychology, Renmin University of China,

Beijing 100872, China)
[3] (CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China)
[4] (Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: With technological advancement, artificial general intelligence (AGI) is beginning to take shape, ushering in a new era for human-machine interaction and relationships. The trust relationship between humans and artificial intelligence (AI) is poised to shift gradually from unidirectional human trust in AI to mutual trust between humans and AI. This study reviews interpersonal trust models from social psychology and human-machine trust models from engineering psychology, proposing a dynamic mutual trust model for human-AI relationships from an interpersonal trust perspective. The model treats humans and AI as equal parties in trust establishment, combining influencing factors from both trustor and trustee, outcome feedback, and behavioral adjustment to construct a fundamental theoretical framework for dynamic human-AI mutual trust. It emphasizes two critical features: "mutual trust" in the relational dimension and "dynamics" in the temporal dimension of human-AI trust. This model is the first to incorporate AI's trust in humans and the dynamic interactive process of mutual trust, providing a new theoretical perspective for human-AI trust research. Future studies should focus on how AI's trust in humans is established and maintained, quantitative modeling of human-AI mutual trust, and mutual trust in multi-agent interactions.

Keywords: trust, human-machine mutual trust, trust calibration, human-machine relationship, human-AI

With the rapid development of next-generation information technology, artificial intelligence (AI) has permeated multiple domains of our daily work and life [**?**, **?**], evolving from smartphone assistants and autonomous vehicles to next-generation chatbots. AI is no longer merely a cold machine system but has become an assistant [**?**, **?**], companion [**?**, **?**], or even romantic partner [**?**, **?**] in people's daily lives, learning, and work, playing increasingly important roles. Since 2023, with ChatGPT entering the public eye, researchers have found that AI is becoming increasingly human-like [**?**, **?**, **?**], proposing that the latest AI model GPT-4 represents a significant step toward artificial general intelligence (AGI) [**?**, **?**]. The relationship between humans and AI is about to transform from a tool-user relationship to a collaborative partnership. In human-AI collaboration, the maturity of AI technology is a prerequisite, but whether humans trust AI becomes a key factor moderating the collaboration [**?**, **?**, **?**, **?**].

Trust is central to human-AI interaction [**?**, **?**, **?**], directly affecting interaction success and user experience. For instance, in autonomous driving system evaluation, trust influences user experience [**?**, **?**, **?**]. This occurs for two primary reasons. First, as technology advances, AI algorithms have become increasingly complex, forming a "black box." People can observe the data input and results output but cannot understand what happens inside [**?**, **?**, **?**]. This makes it dif-

ficult for users to comprehend the decision-making process [?, ?, ?] and predict AI's final decisions. In such cases, users' trust in AI determines whether they will use the algorithm's results. Second, to improve AI performance, users must provide personal data to AI systems [?, ?], which may lead to privacy risks. Therefore, whether users trust AI and are willing to entrust their personal data becomes a critical prerequisite for usage intentions.

Maintaining appropriate trust levels also affects human-AI interaction outcomes, specifically the quality of collaborative task completion. In autonomous driving, trust is a key factor influencing human-machine coordination efficiency and driving safety [?, ?, ?]. If drivers do not trust the AI system, they may ignore its assistance features, failing to effectively reduce risky driving behaviors such as fatigue or distraction. Conversely, if drivers overtrust the AI system, they may completely abandon vehicle monitoring, overlooking the system's limitations and creating significant traffic safety hazards [?, ?, ?, ?]. In military domains, trust between humans and AI teammates is crucial for mission completion [?, ?]. The widespread application of AI systems like drones in human-machine collaborative operations has also drawn increasing research attention to human-AI trust relationships [?, ?]. With the arrival of the AGI era, human-AI trust has become the foundation for harmonious coexistence and collaborative development.

Currently, human-AI interaction relationships have begun to transform, but existing human-AI trust research has not accurately understood this novel trust relationship. This inadequate understanding manifests in three aspects. First, existing research lacks a clear definition of human-AI trust, leading to inconsistent understanding and application among researchers. Second, traditional trust models have primarily discussed interpersonal trust and human-machine trust separately, but as AI technology advances, human-AI interaction will increasingly resemble human-human interaction, making it valuable to integrate these two distinct research domains in psychology. Third, existing trust models only address human trust in AI, neglecting AI's trust in humans and lacking understanding of the bidirectional trust process in human-machine interaction. To address these limitations, this paper examines the definition of human-AI trust and the evolution of trust models, proposes and elaborates on the dynamic mutual trust model for human-AI relationships, and concludes with future research prospects.

To obtain relevant literature on human-AI trust, this review employed the following search strategy. We conducted keyword searches in CNKI, Web of Science, IEEE Xplore, and Elsevier ScienceDirect using terms including "Human-Machine Trust," "trust in AI or trust in artificial intelligence," "trust in automation," and "trust in robot." The search covered literature from 1994 to January 2024, including journal articles and conference papers to ensure comprehensive coverage of nearly 30 years of research.

1. Defining Human-AI Trust

Trust is a common research topic across many disciplines, extensively studied in psychology, sociology, philosophy, political science, and economics. Trust is a complex and ambiguous concept, with over 300 definitions provided across different research fields. Inconsistent trust descriptions prevent researchers from building upon previous work to establish a research system for human-AI trust. Therefore, a clear definition of human-AI trust is crucial for both theoretical and practical research. This paper proposes a definition of human-AI trust based on a review of previous relevant research and definitions.

In the human-machine trust domain, the definition proposed by Lee and See (2004) is widely accepted [**?**, **?**, **?**]. They define trust from an attitudinal perspective, proposing that vulnerability and uncertainty are prerequisites for trust, and define human-machine trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Subsequently, in human-automation system interaction, researchers began proposing automation trust. For example, Billings et al. (2012) reviewed 302 trust definitions, including 220 interpersonal trust definitions and 82 automation trust definitions, finding that numerous automation trust definitions involve user expectations of automation, confidence, risk, vulnerability, dependence, attitudes, and cooperation. These definitions reveal three core characteristics of automation trust required for human-automation collaborative task completion. First, regarding the trust subject, both parties in the trust relationship must be included: a trustor (operator/user) to give trust and a trustee (automation) to receive trust. Second, the matter to be jointly completed must involve certain risks, with the possibility that the trustee may fail to execute and complete the task, thereby triggering uncertainty and risk [**?**, **?**]. Third, the trustee (automation) must have the motivation and capability to execute and complete the task.

Automation trust is the predecessor of human-AI trust. In previous human-machine trust and interaction research, automation and AI have often been used interchangeably [**?**, **?**]. Automation refers to situations where computers follow pre-programmed rules to perform repetitive and monotonous tasks previously executed by humans [**?**, **?**]. Traditional automation produces pre-programmed behaviors and results, allowing users to well understand the automation system's decisions. Traditional automation is deterministic and does not include any learning process [**?**, **?**]. AI, however, can not only achieve automation—for instance, machine learning algorithms can establish rules for automated processes—but can also learn and adjust based on experience and feedback. Summarizing the above elements of automation trust reveals that automation trust exists as a necessary condition for collaboration between trustor and trustee in uncertain cooperative relationships. This construction of automation trust elements applies equally to human-human trust relationships and human-AI trust relationships.

However, unlike automation trust, the establishment of human-AI trust relationships may occur unintentionally. In many AI system usage scenarios, people are not even aware of AI's presence. For example, research on embedded AI has found that people may not realize they are using an AI-supported algorithm application. In a survey of Facebook users, researchers discovered that over half (62%) of users were unaware that an AI algorithm was managing information on their pages, deciding which information to present and which to hide [**?**, **?**]. Although participating Facebook users felt displeased, surprised, or even angry upon learning about the AI algorithm's use, they continued using the platform after understanding how it worked. Hiding AI algorithm usage did not significantly impact long-term user trust. This demonstrates that human trust in AI is not necessarily affected by prior awareness.

More importantly, as AI intelligence levels increase, the relationship between humans and intelligent systems will gradually transform from unidirectional human trust in AI to bidirectional human-AI mutual trust [**?**, **?**, **?**]. Therefore, the definition of human-AI trust itself should evolve with the times, and updated definitions will give rise to new research questions. Based on this, this paper proposes a new definition of human-AI trust: regardless of awareness of AI algorithms' existence, the attitude and confidence held between people and AI systems that the other party can help achieve specific goals, and the willingness to accept the other's uncertainty and vulnerability and bear corresponding risks during interaction.

This new definition synthesizes previous human-machine and automation trust definitions, encompassing not only Lee and See's (2004) attitude-based perspective on human-machine trust but also aligning with the three core characteristics of automation trust summarized by Billings et al. (2012): two trust subjects, risk in the matter to be completed, and the trustee's motivation and capability to complete the task. Building on previous perspectives, the new definition fully considers contemporary characteristics of human-AI interaction: on one hand, it addresses the hidden nature of AI technology use by emphasizing that the definition can extend to situations where users are unaware of AI involvement; on the other hand, considering the transformation of trust roles between humans and AI, it proposes the existence of a mutual trust relationship, where trust includes both the user's trust in AI as trustor and AI's dependence on and adaptation to user input as trustor. This mutual trust relationship also implicitly reveals the dynamic process of human-AI trust, where both humans and AI act as trustors during interaction and continuously calibrate their trust in trustees based on the other's behavior.

2. The Evolution of Human-AI Trust Models: From Human-Human to Human-AI

Human-AI trust research originated from interpersonal trust. With technological development, people increasingly face interactions with AI, such as in healthcare [**?**, **?**] and social domains [**?**, **?**]. In social platform chat windows,

people can no longer easily distinguish whether the other party is human or AI based solely on interaction design, form, and content. In domains accustomed to human cooperation, people increasingly replace human interaction partners with AI, raising the question: Do we trust AI as we trust humans? Therefore, given that trust is a fundamental prerequisite for effective human-AI interaction [?, ?, ?], it is valuable to study human-AI trust by drawing on interpersonal trust theories. Many studies have already transformed human interaction theories and models into theories and models for human-computer interaction (HCI) and human-robot interaction (HRI) research (e.g., [?, ?, ?, ?, ?]). Researchers have applied the human stereotype content model [?, ?, ?] to human-robot research, finding that competence and warmth can positively influence people's trust in robots [?, ?]. Therefore, these determinants of interpersonal trust can be transferred to human-robot interaction and human-AI trust development.

2.1 Interpersonal Trust Models

In interpersonal trust research, scholars propose that trust is essentially a choice to entrust matters we consider important to others, a response to risks beyond our control [?, ?, ?]. In Mayer et al.'s (1995) trust model, interpersonal trust judgments consider three main characteristics: the potential trustee's ability to do what the trustor needs, their benevolence in deciding whether to do it, and their integrity in respecting the trustor and honoring any agreements about whether they will do it.

Early models like Mayer et al.'s focused primarily on the trustee's key characteristics. As trust research deepened, McKnight and Chervany (1996) built upon Mayer et al. (1995) and incorporated the trust decision-making process to propose a trust concept relational model. They noted that despite similar risk and trustee trustworthiness, people often trust others more in some situations than others. Therefore, the new model emphasizes that trust propensity derives not only from trustee characteristics but also from potential trustor attitudes (such as optimism) and decision-making contexts. The decision-making context refers to the broader (i.e., non-personal) social situation in which trust decisions are made, meaning individual trust decisions are also influenced by system trust. For example, some social cultures tend to cultivate more general trust among people than others, and the existence of certain institutions and social norms may increase or decrease system trust tendencies [?, ?]. Unlike Mayer's early model, which only emphasized trustee traits, the trust concept relational model's emphasis on trustor and situational factors provides an appropriate theoretical framework for proposing human-machine trust three-dimensional models.

In 2011, Sanders et al. proposed a four-factor model for human-robot trust, identifying robot performance, robot dependence, individual differences, and collaboration as influencing factors in human-robot interaction (HRI) [?, ?]. These factors, ranging from robot attributes (such as anthropomorphism, animacy, affinity, perceived intelligence, and perceived safety) to human interaction factors (such as usability, social acceptance, user experience, and social influence),

comprehensively summarize antecedents of trust in previous research. Building on this work and based on meta-analysis results, researchers first summarized antecedents of human-robot trust into three factors: robot-related factors (including robot performance and characteristics), human-related factors (including capability and personal traits), and environment-related factors (including team collaboration and task-related factors) [**?**, **?**]. This work laid the foundation for subsequent classic three-factor models.

In 2014, Schaefer et al. developed a three-factor model of human-machine trust based on a review of human-machine trust literature and the human-robot trust model [**?**, **?**] [**?**, **?**]. They categorized factors influencing human-machine trust into operator factors, machine system factors, and environmental factors, further subdividing operator-related factors into operator traits, operator states, cognitive factors, and affective factors; machine system-related factors into machine system characteristics and capabilities; and environment-related factors into task-related and team-related categories.

Building on this, Chinese researchers proposed a three-factor model influencing AI trust. Human trust in AI relates to three aspects: individual operator characteristics, as trust in AI originates from humans; situational characteristics, as a good trust social system and institutions provide a favorable context for AI trust; and AI system characteristics, such as technical performance and effects as essential elements of AI trust [**?**, **?**, **?**]. In short, factors influencing human trust in AI should include individual, technology, and environment.

In 2022, Lewis and Marsh proposed an integrated model in their review of AI trust research. This model applies not only to trust between peers but also to heuristic trust decision-making, laying the foundation for the dynamic mutual trust model for human-AI relationships. The model suggests that people's trustworthiness judgments heavily depend on the amount and type of available information, which influences perceptions of four main trustworthiness characteristics: ability, predictability, honesty and integrity, and willingness and benevolence [**?**, **?**]. When a trustee possesses the ability to complete corresponding tasks, behaves consistently and predictably, is willing to fulfill commitments, and has the intention to meet needs, trust decisions can be fulfilled. When information on these four main characteristics is difficult to obtain, people can complete decisions through proxy trust (i.e., trust in other relevant parties). For example, trust in a new product may be influenced by the proxy—the manufacturer. The manufacturer's reputation and past product quality affect consumers' trustworthiness perceptions of new products. In actual human-AI relationships, an AI system may be capable, but people find it difficult to discern whether it is honest or benevolent. In most cases, these influencing factors intertwine and jointly affect people's trust in AI.

However, the integrated model focuses more on the trustee—AI's own characteristics—and their impact on perceived trustworthiness and trust decisions and behaviors, while neglecting the influence of user states.

2.5 Comprehensive Comparison and Analysis of Previous Trust Models

The evolution of these models shows that researchers' understanding of trust is continuously deepening. Early trust models were discussed in interpersonal interaction contexts. Mayer et al.'s (1995) model innovatively proposed that trust depends on three trustee characteristics—ability, benevolence, and integrity—but only considered trustee traits. McKnight and Chervany's (1996) model expanded to include trustor and situational factors, more comprehensively explaining interpersonal trust influencing factors and providing a general framework for human-machine trust models. In human-machine interaction, Sanders et al. (2011) first summarized human-machine trust research, proposing a four-factor model, but this model could not broadly generalize all trust antecedents. Hancock et al. (2011) revised existing models, summarizing antecedents into three factors, but this model was based only on human-robot interaction research, where supporting evidence for human-related and environment-related factors was limited, resulting in insufficient exploration of these two factor categories. Schaefer et al. (2014) revised the three-factor model based on meta-analysis results of human-automation interaction research, providing more detailed categorization of each factor's content. Lewis et al. (2022) further developed the trust model; their integrated model considers proxy trust's influence and emphasizes the dynamic adjustment process of trust, providing a general analytical framework for different types of trust relationship research. The development of previous trust models shows a trend from static to dynamic, from simple dimensional division to more comprehensive and detailed factor consideration, but still exhibits the limitation of neglecting the bidirectional mutual trust relationship between humans and AI.

3. The Dynamic Mutual Trust Model for Human-AI Interaction

3.1 Proposal of the Dynamic Mutual Trust Model

Against the backdrop of the AGI era, human-AI interaction relationships are becoming increasingly complex. Previous human-machine trust models, despite their theoretical contributions, have limitations in explaining the dynamic and bidirectional trust relationship between humans and AI and are insufficient to comprehensively describe the trust interaction process. Therefore, this paper proposes a new model to fill the theoretical gap in existing human-AI trust research. This model fully references existing trust model content to comprehensively capture factors influencing the trust process. In framework, it draws upon the interpersonal trust model [**?**, **?**], including trustor-related factors, trustee-related factors, and situational factors. Specific content for each factor category references the general integrated model [**?**, **?**] while further considering the uniqueness of human-AI interaction. Therefore, the new model's characteristics are: it emphasizes that trust is not just unidirectional human evaluation

of AI but an interactive process involving both humans and AI, where both parties continuously adjust their trust levels and behavioral strategies based on each other's actions and feedback. In summary, based on existing trust models (including interpersonal trust models, four-factor human-machine trust models, three-factor human-machine trust models, and integrated models of human trust in AI), this paper proposes a new human-machine mutual trust model for the novel bidirectional mutual trust relationship in the AGI era: the dynamic mutual trust model for human-AI relationships, as shown in Figure 1 [Figure 1: see original paper].

Figure 1 illustrates the dynamic mutual trust model for human-AI relationships. The roles of trustor and trustee are dynamic; human-AI trust is influenced by perceived other-party states (blue), self-states (green), and situational factors (yellow), and adjusts based on outcome feedback (blue lines).

The model proposes two important features in human-AI trust: "mutual trust" and "dynamics." "Mutual trust" emphasizes the relational dimension, while "dynamics" focuses on the temporal dimension of the human-AI trust relationship.

The mutual trust relationship between humans and AI represents a new type of human-machine relationship in the AGI era. Unlike previous research focusing on unidirectional human trust in AI, "mutual trust" emphasizes AI's subject status similar to humans in the trust relationship. As AGI technology continues to develop, intelligent machines will evolve from auxiliary tools supporting human operations to autonomous intelligent agents with certain cognitive, independent execution, and adaptive capabilities, possessing human-like behavioral abilities to some extent [**?**, **?**]. AI will actively perceive human user states and its own system states, evaluate trust levels in human users, and determine control allocation. Although no real-world examples exist yet, science fiction works have depicted scenarios where AI distrusts users and refuses tool usage [**?**, **?**]. At this point, human-machine trust will no longer be unidirectional human trust in machine systems but will gradually transform into bidirectional human-machine mutual trust [**?**, **?**, **?**, **?**]. Human-AI mutual trust is actually based on an interpersonal trust perspective, treating humans and AI as equal trust-building parties. Therefore, both humans and AI can serve as trustors (entrusting parties) or trustees (entrusted parties).

The human-AI mutual trust relationship also determines that its "dynamic" changes differ from previous unidirectional trust in the temporal dimension. In unidirectional trust, researchers have divided human-automation system trust into several stages according to the temporal sequence of trust development: dispositional trust, situational trust, and learned trust [**?**, **?**, **?**, **?**, **?**]. In interpersonal trust establishment, researchers propose that trust is a feedback loop of trust, where trustors form initial trust based on their own experiences and dispositions, make trust decisions and behaviors based on perceptions of trustees, and then use feedback outcomes to influence subsequent trust [**?**, **?**, **?**]. The human-AI mutual trust interaction process is a dynamic process where trustors and trustees continuously adjust their own behaviors and calibrate trust levels

in trustees based on each other's states and behaviors during the trust process and the final trust outcomes. Therefore, this framework proposes that dynamic human-AI mutual trust can be divided into three stages: the initial stage before human-AI interaction, the perception stage during human-AI interaction, and the behavior stage, forming a closed loop. The initial stage is the beginning of human-AI trust, where humans and AI have not yet interacted, relying on inherent trust dispositions, system trust, and relevant trust experience from previous interactions to set the tone for subsequent trust. Among these, trust experience will be corrected after receiving outcome feedback from the current interaction, participating in the dynamic process of human-AI mutual trust, while system trust and trust dispositions remain relatively stable and do not participate in subsequent dynamic processes. In the perception stage, human-AI trust is influenced by perceived other-party states, perceived self-states, and situational states, forming trust decisions. In the behavior stage, trustors complete trust behaviors and calibrate trust experience from the initial stage based on behavioral outcome feedback, generating new trust experience while updating perceived states of trustees to influence subsequent trust behaviors. Outcome feedback contains two meanings: on one hand, the trustee's trust behavior itself—whether the trustee executed the trustor's decision; on the other hand, the system operation results after the trustee executed or failed to execute the trustor's decision. Human-AI mutual trust continuously calibrates through this process, achieving dynamic interaction. For example, when humans serve as trustees, if AI signals distrust (such as fatigue warnings), humans will adjust their own states (trusting AI's decisions) to regain AI trust or choose to trust AI to take over the system. If AI receives human distrust signals, it will also adjust its system state through self-checking (trusting human decisions) or allowing designers to debug the system to gain human trust, thereby achieving normal system operation. The dynamic interaction process of human-AI mutual trust actually reflects the trust calibration process [**?**, **?**]. Although the ideal state of human-AI mutual trust is appropriate trust, in reality, overtrust [**?**, **?**, **?**] and trust insufficiency [**?**, **?**, **?**] are common in human-machine interaction. Therefore, this model proposes that human-AI mutual trust should, like interpersonal trust, involve trust updating [**?**, **?**, **?**]. In dynamic human-AI mutual trust, trust updating depends on previous trust behavior outcomes.

In summary, the dynamic mutual trust model for human-AI relationships includes three stages (initial, perception, and behavior) and two subjects (humans and AI). In human-AI mutual trust, the two subjects—humans and AI—have similar and different trust influencing factors in the first two stages, which will be discussed separately below.

3.2 Factors Influencing Human Trust in AI

Factors influencing human trust in AI are primarily proposed based on Lewis and Marsh's (2022) integrated model framework combined with previous literature. In the initial stage, human trust in AI is mainly influenced by individual

trust disposition, previous trust experience, and system trust. Individual trust disposition, also called dispositional trust in other studies, is affected by inherent individual traits such as age [?, ?, ?], personality [?, ?], and education level [?, ?]. Trust experience refers to prior experience or expertise gained from using AI-related systems and products. This experience helps individuals predict system behavior [?, ?], thereby changing human trust in AI. For example, Dikmen and Burns (2017) experimentally tested user trust in Tesla's autonomous driving system. Results showed that drivers who had experienced vehicle accidents had lower trust in the autonomous driving system, while drivers familiar with Tesla's autonomous driving system had higher trust. Human trust in AI is also influenced by institutional trust. For instance, some social cultures tend to cultivate more general trust among individuals than others [?, ?].

In the perception stage, human trust in AI is influenced by three factors. First, perceived individual state—whether individuals perceive themselves as capable of handling current tasks. Second, perceived system state, including perceived trustworthiness and perceived risk. Perceived trustworthiness includes multidimensional perceptions of the trustee's ability, predictability, integrity, benevolence, and proxy trust (such as brand) [?, ?, ?]. Perceived risk refers to the assessment of the trustee's vulnerability and the risk level associated with completing current tasks [?, ?, ?]. Third, situational state—humans need to evaluate the nature of the situation and task difficulty. Research shows that when human-AI collaborative task workload increases, human trust in AI systems decreases, and people tend to prefer completing tasks alone [?, ?]. Conversely, Atoyan et al. (2006) demonstrated through experiments that when human-AI collaborative tasks are too complex and numerous for humans to complete independently, people may develop overtrust in the collaborative AI system.

3.3 Factors Influencing AI's Trust in Humans

In the initial stage, factors influencing AI's trust in humans include the AI system's own trust disposition and prior experience formed from previous interactions with users. Currently, AI's trust disposition primarily reflects system designers' trust disposition toward human users. Considering the current lack of national-level AI-related legal systems [?, ?] and difficulty in assigning relevant responsibilities, AI in high-risk tasks currently tends to trust human users (e.g., autonomous driving). In the AGI era, AI's trust disposition may become more similar to humans', depending more on AI's inherent traits (such as personalized design for specific user groups, primary tasks, form, safety guarantees, etc.) rather than solely on initial settings.

In the perception stage, AI's trust in humans is also influenced by three factors. First, user state—AI needs to build monitoring systems to monitor user states (cognitive, physiological, intentional, emotional, values, moral levels, etc.) in real time. When users are in untrustworthy states (such as fatigue or distraction), AI will proactively take over to avoid accidents [?, ?]. Second, system state—AI needs active monitoring and evaluation systems for its own state, in-

cluding monitoring its own performance and stability and assessing whether its current state can complete tasks. Taking autonomous driving as an example, autonomous vehicles are equipped with numerous internal sensors to continuously monitor vehicle internal state data, and researchers continue developing effective automatic fault diagnosis and health monitoring algorithms [**?**, **?**] to evaluate system states. When the system detects it is unreliable (such as system failure or tasks exceeding system capability), it will make judgments to trust humans and prompt human users to take over control. Third, situational state—AI needs to evaluate the risk and complexity levels of the situation, such as environmental conditions and emergency occurrences, to determine whether to trust users. Again using autonomous driving as an example, vehicles use cameras, LiDAR, ultrasonic sensors, etc., to perceive traffic conditions, lighting conditions, obstacles, and other external situations [**?**, **?**], taking corresponding trust behaviors based on perceived situations. When the system detects high-risk situations (such as an impending rear-end collision), AI may become more cautious, reduce trust in humans, and take emergency measures like braking or emergency lane changes. In low-risk situations, AI will trust humans more, granting them greater autonomy.

4. Future Directions for Human-AI Trust Research

Currently, the AGI era is about to begin, yet research on human-AI trust remains insufficient. Future research can proceed in the following three directions:

4.1 AI's Trust in Humans

Currently, AI system designers' trust in human operators is subtle. In a study of Uber drivers, Möhlmann and Zalmanson (2017) noted that continuous individual performance evaluation and feedback (only possible through continuous tracking) violated drivers' sense of autonomy and reduced their trust. This continuous monitoring was viewed as a form of micromanagement, indicating that those deploying AI (AI system designers) lacked trust in AI operators, which in turn reduced drivers' trust in autonomous driving AI.

In the intelligent era, machines gradually acquire human-like behavioral capabilities [**?**, **?**], and trust will no longer be unidirectional human trust in automated systems but bidirectional human-machine mutual trust [**?**, **?**, **?**]. Under the human-machine mutual trust framework, research in two areas urgently needs to be conducted. First, system designers' trust in users. Due to human operators' limitations (such as physiological and psychological factors), system designers often trust AI judgments more in some scenarios while neglecting monitoring of user states. For example, researchers could model appropriate trust models of the system in drivers based on data such as drivers' current states (fatigue, distraction, etc.), system states (reliability, etc.), and situational conditions (environmental risk, etc.), enabling the system to proactively intervene when drivers are in untrustworthy states to avoid accidents. Second, AI's trust in users. If in the weak AI era, AI's trust in humans could be equated with

AI designers' trust in users, then in the strong AI era, AI will possess self-awareness and autonomous judgment, and human-AI cooperative relationships will depend on mutual trust. A novel question then becomes how humans can earn AI's trust and what special characteristics this might entail compared to human-human trust. As AI intelligence levels increase, this question will gradually affect human-AI interaction.

4.2 Quantitative Models of Human-AI Mutual Trust

Currently, most studies have theoretically proposed research frameworks and qualitative models, yet lack quantitative models to guide system design. Establishing quantitative models depends on two prerequisites: (1) Trust measurement. The most commonly used trust measurement method is self-reporting [**?**, **?**, **?**, **?**]. Self-report measurement methods are easy to use and can effectively reflect operators' human-machine trust levels if questionnaires or scales are properly constructed. However, self-report measurement methods are intrusive to interaction tasks and cannot capture dynamic changes in human-AI trust in real time, limiting their application in real-world environments. Additionally, this method has inevitable flaws: subjects may be unable or unwilling to accurately report their true attitudes and cannot describe the impact of implicit attitudes on their trust levels [**?**, **?**]. To compensate for self-report measurement defects, some researchers have begun inferring human-machine trust levels from observable behaviors. Behavioral indicators for measuring human-machine trust primarily rely on concepts of compliance and dependence: when operators are more inclined to comply with or depend on the system, their human-machine trust level is higher, and vice versa. Compliance refers to operators responding when the machine system signals, measurable by the degree to which operators accept suggestions or actions provided by the system [**?**, **?**]. Dependence refers to operators not responding when the machine system is silent or operating normally, measurable by the proportion of time (or number of times) operators use the automated system out of total time (or total task count) [**?**, **?**, **?**]. Additionally, reaction time measurement methods can be used, measuring the speed at which operators take over automated system control after detecting system risk [**?**, **?**, **?**], where faster operator reaction times indicate lower trust in the automated system. Physiological and neural measurements aim to measure human-machine trust in real time by measuring related physiological and neural indicators. Although this method is still in its infancy, existing literature shows it is highly effective in capturing real-time dynamic changes in human-machine trust [**?**, **?**].

Based on existing measurement methods, further identification of inconsistencies among multiple measurement methods is needed, seeking correspondences between behavioral indicators and physiological/neural indicators and subjective trust levels to determine more accurate real-time measurement indicators for identifying basic states of dynamic human-AI trust (appropriate trust, trust insufficiency, and overtrust). How to accurately characterize AI's trust in humans

will also become important content in human-AI trust research. (2) Determining weights of trust influencing factors. Different studies have examined various influencing factors in human-AI mutual trust, yet integrated model research is still lacking. How to accurately measure various trust-related influencing factors, analyze and quantify each factor's weight and conditions of effect in actual human-AI interaction processes, cover environmental and individual factors that have significant impacts, and consider the adverse effects of unmodeled factors on model performance to construct trust computational models that meet different design stage requirements will become key to improving human-AI mutual trust and enhancing model application value.

4.3 Human-AI Mutual Trust in Multi-Agent Interactions

The model proposed in this paper applies to common scenarios where AI serves as a human assistant or collaborative partner and human-AI collaboration completes tasks [**?**, **?**]. However, the model only addresses the mutual trust process between a single human and a single AI. As AI usage scenarios become more complex, interactions involving multiple humans and multiple AI will emerge. Previous researchers believe that in multi-agent interactions, each member's role and interaction methods are key factors influencing trust [**?**, **?**]. In such environments, the dynamic trust construction process becomes more complex. Based on this model, the identity roles of each agent should be further incorporated, considering their weights in the dynamic mutual trust process. For example, Figure 2 [Figure 2: see original paper], based on distributed cognition [**?**, **?**], incorporates role allocation in the human-AI dynamic mutual trust process. When an "opinion leader" emerges in multi-agent interactions (the blue agent in the figure), the opinion leader's (human or AI) trust experience will influence other agents (the gray agents in the figure) through communication, thereby affecting other humans' trust in AI. Only by studying AI in complex groups (such as teams or networks) can researchers truly understand how people establish "partnership" relationships with AI and how AI changes relationships between people and between people and other machines. Moreover, since AI behavior is not stable and unchanging, scholars need to study how it changes based on human-AI interaction [**?**, **?**] to promote understanding of relationship changes. Future research should consider interactions between multiple people and multiple AI, which will provide better support for establishing human-AI partnerships and forming human-AI mutual trust under the human-in-the-loop framework.

References

高在峰, 李文敏, 梁佳文, 潘晗希, 许为, 沈模卫. (2021). 自动驾驶车中的人机信任. 心理科学进展, 29(12),

何积丰. (2019). 安全可信人工智能. 信息安全与通信保密, 10, 5–8.

许为, 高在峰, 葛列众. (2024). 智能时代人因科学研究的新范式取向及重点. 心理学报, 56(3), 363–382.

许为, 葛列众. (2020). 智能时代的工程心理学. 心理科学进展, 28(9), 1409–1425.

闫宏秀. (2019). 用信任解码人工智能伦理. 人工智能, (4), 7.

赵竟, 孙晓军, 周宗奎, 魏华, 牛更枫. (2013). 网络交往中的人际信任. 心理科学进展, 21(8), 1493–1501.

Ajenaghughrure, I. B., da Costa Sousa, S. C., & Lamas, D. (2020, June). Risk and trust in artificial intelligence technologies: A case study of autonomous vehicles. In 2020 13th International Conference on Human System Interaction (pp. 118–123), Tokyo, Japan. doi: 10.1109/HSI49210.2020.9142686

Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. ACM Transactions on Interactive Intelligent Systems, 8(4), 1–20. doi:10.1145/3132743

Aly, A., & Tapus, A. (2016). Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction. Autonomous Robots, 40(2), 193–209. doi:10.1007/s10514-015-9444-1

Atoyan, H., Duquet, J.-R., & Robert, J.-M. (2006, April). Trust in new decision aid systems. In Proceedings of the 18th Conference I'Interaction Homme-Machine 115–122), Montreal, Canada. doi:10.1145/1132736.1132751

Bartneck, C., & Forlizzi, J. (2004, September). A design-centered framework for social human-robot interaction. In IEEE International Workshop on Robot & Human Interactive Communication (pp. 591–594), Kurashiki, Japan. doi: 10.1109/ROMAN.2004.1374827

Biddle, L., & Fallah, S. (2021). A novel fault detection, identification and prediction approach for autonomous vehicle controllers using SVM. Automotive Innovation, 4(3), 301-314. doi: 10.1007/s42154-021-00138-0

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

Billings, D. R., Schaefer, K. E., Llorens, N., & Hancock, P. A. (2012). What is trust? Defining the construct across domains. In Poster presented at the American Psychological Association Conference (Division 21, pp. 1-7), Orlando, FL, USA.

Bindewald, J. M., Rusnock, C. F., & Miller, M. E. (2018). Measuring human trust behavior in human-machine teams. In Advances in Human Factors in Simulation and Modeling (vol. 591, pp. 47–58), Los Angeles, USA. Springer International Publishing. doi:10.1007/978-3-319-60591-3_5

Binz, M. & Eric Schulz. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120(6), e2218523120. doi:10.1073/pnas.2218523120

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E.,

… & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arxiv preprint arxiv:2303.12712.

Chen, I.-R., Bastani, F. B., & Tsao, T.-W. (1995). On the reliability of AI planning software in real-time applications. IEEE Transactions on Knowledge and Data Engineering, 7(1), 4–13. doi:10.1109/69.368522

Chen, J. Y. C., Barnes, M. J., & Harper-Sciarini, M. (2011). Supervisory control of multiple robots: Human-performance issues and user-interface design. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41(4), 435–454. doi:10.1109/TSMCC.2010.2056682

Christoforakos, L., Gallucci, A., Surmava-Große, T., Ullrich, D., & Diefenbach, S. (2021). Can robots earn our trust the same way humans do? A systematic exploration of competence, warmth, and anthropomorphism as determinants of trust development in HRI. Frontiers in Robotics and AI, 8, 640444. doi:10.3389/frobt.2021.640444

Cofta, P. (2007). Trust, complexity and control: Confidence in a convergent world. John Wiley & Sons, Ltd. doi:10.1002/9780470517857

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. Journal of Experimental Psychology: Applied, 22(3), 331–349. doi:10.1037/xap0000092

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. International Journal of Human-Computer Studies, 58(6), 719–735. doi:10.1016/S1071-5819(03)00039-9

Deutsch, M. (1962). Cooperation and trust: Some theoretical notes. In Jones, M.R., (Ed.), Nebraska Symposium on Motivation (pp. 275–320). University of Nebraska Press.

Dikmen, M., & Burns, C. (2017, October). Trust in autonomous vehicles: The case of Tesla Autopilot and Summon. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 1093–1098), Banff, Canada. doi:10.1109/SMC.2017.8122757

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015, April). "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 153–162), Seoul, Republic of Korea. doi:10.1145/2702123.2702556

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology, 82(6), 878–902. doi:10.1037/0022-3514.82.6.878

Fiske, S. T., Xu, J., Cuddy, A. C., & Glick, P. (1999). (Dis)respecting versus (Dis)liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. Journal of Social Issues, 55(3), 473–489. doi:10.1111/0022-4537.00128

Fogg, B. J., & Tseng, H. (1999, May). The elements of computer credibility. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 80–87), Pittsburgh, USA. doi:10.1145/302979.303001

Forcier, M. B., Khoury, L., & N Vézina. (2020). Liability issues for the use of artificial intelligence in health care in canada: and medical decision-making. Dalhousie Medical Journal, 46(2), 7–11. doi: 10.15273/dmj.Vol46No2.10140

French, B., Duenser, A., Heathcote, A. (2018). Trust in automation – A literature review. Commonwealth Scientific and Industrial Research Organisation Report, EP184082.

Frison, A.-K., Wintersberger, P., Riener, A., Schartmüller, C., Boyle, L. N., Miller, E., & Weigl, K. (2019, May). In UX we trust: Investigation of aesthetics and usability of driver-vehicle interfaces and their impact on the perception of automated driving. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13), Glasgow, UK. doi:10.1145/3290605.3300374

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660. doi:10.5465/annals.2018.0057

Gockley, R., Simmons, R., & Forlizzi, J. (2006, September). Modeling affect in socially interactive robots. In ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication (pp. 558–563), Hatfield, UK. doi:10.1109/ROMAN.2006.314448

Gremillion, G. M., Metcalfe, J. S., Marathe, A. R., Paul, V. J., Christensen, J., Drnec, K., Haynes, B., & Atwater, C. (2016). Analysis of trust in autonomy for convoy operations. In Micro and nanotechnology sensors, systems, and applications, 9836, 356–365. doi:10.1117/12.2224009

Groom, V., & Nass, C. (2007). Can robots be teammates?: Benchmarks in human–robot teams. Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems, 8(3), 483–500. doi: 10.1075/is.8.3.10gro

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human Factors: The Journal of the Human Factors and Ergonomics Society, 53(5), 517–527. doi:10.1177/0018720811417254

Hancock, P. A., Nourbakhsh, I., & Stewart, J. (2019). On the future of transportation in an era of automated and autonomous vehicles. Proceedings of the National Academy of Sciences, 116(16), 7684–7691. doi:10.1073/pnas.1805770115

Hardin, R. (2006). Trust and trustworthiness (the russell sage foundation series on trust, vol. 4). American Handgunner (March-April).

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors: The Journal of the Human Factors and Ergonomics Society, 57(3), 407–434. doi:10.1177/0018720814547570

Ignatious, H. A., & Khan, M. (2022). An overview of sensors in autonomous vehicles. Procedia Computer Science, 198, 736-741. doi: 10.1016/j.procs.2021.12.315

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), 10.1207/S15327566IJCE0401_{04}

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2017). Calibrating trust to increase the use of automated systems in a vehicle. In Advances in Human Aspects of Transportation: Proceedings of the AHFE 2016 International Conference on Human Factors in Transportation, 484, 535–546. Springer International Publishing. doi:10.1007/978-3-319-41682-3_{45}

Kim, M., Park, B. K., & Young, L. (2020). The psychology of motivated versus rational impression updating. Trends in Cognitive Sciences, 24(2), 101–111. doi: 10.1016/j.tics.2019.12.001

Kulms, P., & 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. Frontiers in Digital Humanities, 5, 14. doi:10.3389/fdigh.2018.00014

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. doi: 10.1518/hfes.46.1.50_{30392}

Lewis, P. R., & Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness artificial intelligence. Cognitive Systems Research, doi:10.1016/j.cogsys.2021.11.001

Liao, T., & MacDonald, E. F. (2021). Manipulating users' trust of autonomous products with affective priming. Journal of Mechanical Design, 143(5), 051402. doi:10.1115/1.4048640

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. Journal of Consumer Research, 46(4), 629–650. doi:10.1093/jcr/ucz013

Luhmann, & N. (1990). Technology, environment and social risk: A systems perspective. Organization & Environment, 4(3), 223–231. doi: 10.1177/108602669000400305

Ma, Y., Li, S., Qin, S., & Qi, Y. (2020, December). Factors affecting trust in the autonomous vehicle: A survey of primary school students and parent

perceptions. In 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (pp. 2020–2027), Guangzhou, China. doi: 10.1109/TrustCom50675.2020.00277

Madsen, M., & Gregor, S. (2000, December). Measuring human-computer trust. In 11th australasian conference on information systems, 53, 6–8.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. Academy of Management Review, 20(3), 709–734. doi:10.5465/amr.1995.9508080335

Mcknight, D. H., & Chervany, N. L. (1996). The Meaning of Trust [Technical Report]. Management Information Systems Research Center, University of Minnesota.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. Social Cognitive and Affective Neuroscience, 8(6), 623–631. doi: 10.1093/scan/nss040

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in humanautomation interactions. Human Factors: The Journal of the Human Factors and Ergonomics Society, 50(2), 194–210. doi: 10.1518/001872008X288574

Mohanty, S., & Vyas, S. (2018). Putting it all together: Toward a human-machine collaborative ecosystem. In S. Mohanty & S. Vyas (Eds.), How to Compete in the Age of Artificial Intelligence: Implementing a collaborative human-machine strategy for your business (pp. 215-229), Apress, Berkeley, CA, USA. doi: 10.1007/978-1-4842-3808-0_{11}

Möhlmann, M., & Zalmanson, L. (2017, December). Hands on the wheel: Navigating algorithmic management and Uber drivers'. In Autonomy', in proceedings of the international conference on information systems (pp. 10-13), Seoul, Republic of Korea.

Molnar, L. J., Ryan, L. H., Pradhan, A. K., Eby, D. W., St. Louis, R. M., & Zakrajsek, J. S. (2018). Understanding trust and acceptance of automated vehicles: An exploratory simulator study of transfer of control between automated and manual driving. Transportation Research Part F: Traffic Psychology and Behaviour, 58, 319–328. doi:10.1016/j.trf.2018.06.004

Noah, B. E., Gable, T. M., Schuett, J. H., & Walker, B. N. (2016, October). Forecasted affect towards automated and warning safety features. In Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces Interactive Vehicular Applications 123–128), Ann Arbor, USA. doi:10.1145/3004323.3004337

Noah, B. E., Wintersberger, P., Mirnig, A. G., Thakkar, S., Yan, F., Gable, T. M., Kraus, J., & McCall, R. (2017, September). First workshop on trust in the age of automated driving. Proceedings of the 9th International Conference

on Automotive User Interfaces and Interactive Vehicular Applications Adjunct (pp. 15–21), Oldenburg, Germany. doi:10.1145/3131726.3131733

Oleson, K. E., Billings, D. R., Kocsis, V., Chen, J. Y. C., & Hancock, P. A. (2011, February). Antecedents of trust in human-robot collaborations. In 2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA) (pp. 175–178), Miami Beach, USA. doi:10.1109/COGSIMA.2011.5753439

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors: The Journal of the Human Factors and Ergonomics Society, 39(2), 230–253. doi:10.1518/001872097778543886

Payre, W., Cestac, J., & Delhomme, P. (2016). Fully automated driving: Impact of trust and practice on manual control recovery. Human Factors: The Journal of the Human Factors and Ergonomics Society, 58(2), 229–241. doi:10.1177/0018720815612319

Perry, M. (2003). Distributed cognition. In J.M. Carroll (Ed.), HCI models, theories, and frameworks: Toward a multidisciplinary science (pp. 193–223), Morgan Kaufmann.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., … Wellman, M. (2019). Machine behaviour. Nature, 568(7753), 477–486. doi:10.1038/s41586-019-1138-y

Raj, M., & Seamans, R. (2019). Primer on artificial intelligence and robotics. Journal of Organization Design, 8(1), 11. doi:10.1186/s41469-019-0050-0

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016, March). Overtrust of robots in emergency evacuation scenarios. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (pp. 101–108), Christchurch, New Zealand. doi:10.1109/HRI.2016.7451740

Rödel, C., Stadler, S., Meschtscherjakov, A., & Tscheligi, M. (2014, September). Towards autonomous cars: The effect of autonomy levels on acceptance and user experience. Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 1–8), Seattle, USA. doi:10.1145/2667317.2667330

Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2018). The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. Paladyn, Journal of Behavioral Robotics, 9(1), 137–154. doi:10.1515/pjbr-2018-0010

Sanders, T., Oleson, K. E., Billings, D. R., Chen, J. Y. C., & Hancock, P. A. (2011). A model of human-robot trust: Theoretical model development. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55(1), 1432–1436. doi:10.1177/1071181311551298

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 [Technical Report]. Army Research Lab, Aberdeen Proving Ground, Maryland, Human Research Engineering Directorate. doi:10.21236/ADA607926

Scopelliti, M., Giuliani, M. V., & Fornara, F. (2005). Robots in a domestic setting: A psychological approach. Universal Access in the Information Society, 4(2), 146–155. doi:10.1007/s10209-005-0118-1

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. Proceedings of the National Academy of Sciences, 120(10), e2300963120.

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) ethics: Ethics of AI and ethical AI. Journal of Database Management, 31(2), 74–87. doi:10.4018/JDM.2020040105

Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y. C., Dong, J., Duffy, V. G., … Zhou, J. (2019). Seven HCI grand challenges. International Journal of Human–Computer Interaction, 35(14), 1229–1269. doi:10.1080/10447318.2019.1619259

Stokes, C. K., Lyons, J. B., Littlejohn, K., Natarian, J., Case, E., & Speranza, N. (2010, May). Accounting for the human in cyberspace: Effects of mood on trust in automation. In 2010 International Symposium on Collaborative Technologies and Systems (pp. 180–187), Chicago, USA. doi:10.1109/CTS.2010.5478512

Sullins, J. P. (2010). Love and sex with robots: The evolution of human-robot relationships [Book review]. Industrial Robot: An International Journal, 37(4), 401–402. doi:10.1108/ir.2010.04937dae.001

Urban, G. L., Amyx, C., & Lorenzon, A. (2009). Online trust: State of the art, new frontiers, and research potential. Journal of Interactive Marketing, 23(2), 179–190. doi:10.1016/j.intmar.2009.03.001

van Pinxteren, M. M. E., Wetzels, R. W. H., Rüger, J., Pluymaekers, M., & Wetzels, M. (2019). Trust in humanoid robots: Implications for services marketing. Journal of Services Marketing, 33(4), 507–518. doi:10.1108/JSM-

Walter, S., Wendt, C., Böhnke, J., Crawcour, S., Tan, J.-W., Chan, A., Limbrecht, K., Gruss, S., & Traue, H. C. (2014). Similarities and differences of emotions in human–machine and human–human interactions: What kind of emotions relevant future companion systems? Ergonomics, 57(3), doi:10.1080/00140139.2013.822566

Wang, W., & Siau, K. (2019). Artificial Intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. Journal of Database Management, 30(1), 61–79. doi:10.4018/JDM.2019010104

Wikipedia contributors. (2024, January 13). Psycho-Pass. In Wikipedia, The

Free Encyclopedia. from https://en.wikipedia.org/w/index.php?title=Psycho-Pass&oldid=1195338833

Wintersberger, P., Noah, B. E., Kraus, J., McCall, R., Mirnig, A. G., Kunze, A., Thakkar, S., & Walker, B. N. (2018, September). Second workshop on trust in the age of automated driving. Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 56–64), Toronto, Canada. doi:10.1145/3239092.3239099

Wright, P., McCarthy, J., & Meekison, L. (2003). Making sense of experience. In Blythe, M.A., Overbeeke, K., Monk, A.F., Wright, P.C. (Eds.), Funology: From usability to enjoyment (vol. 3, pp. 43–53), Springer Netherlands. doi:10.1007/1-4020-2967-5_5

Yagoda, R. E., & Gillan, D. J. (2012). You want me to trust a robot? The development of a human–robot interaction trust scale. International Journal of Social Robotics, 4(3), 235–248. doi:10.1007/s12369-012-0144-0

Author Contribution Statement:

Qi Yue and Du Feng: Proposed the research proposition;
Qi Yue, Chen Junting, and Qin Shaotian: Drafted the paper;
Qi Yue, Chen Junting, and Du Feng: Revised the final version of the paper.

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