Dual Pathways for Human-Machine Trust Calibration: Trust Suppression and Trust Enhancement

Authors: Huang Xinyu, Li Ye, Li Ye

Date: 2023-12-25T00:00:00+00:00

Abstract

Trust is the foundation of successful human-robot cooperation. However, individuals in human-robot interaction do not always maintain appropriate trust levels, and trust biases may emerge: over-trust and under-trust. Trust biases can impede human-robot cooperation, thus necessitating trust calibration. Trust calibration is often achieved through two pathways: trust suppression and trust enhancement. Trust suppression focuses on reducing individuals' excessively high trust levels in robots, whereas trust enhancement emphasizes increasing individuals' low trust levels in robots. Future research could further optimize measurement methods for evaluating calibration effectiveness, reveal the mechanisms of cognitive changes in individuals during and after trust calibration, explore the boundary conditions of trust calibration, and examine personalized and refined trust calibration strategies, with the aim of facilitating human-robot collaboration.

Full Text

Preamble

Trust Dampening and Trust Promoting: A Dual-Pathway Approach to Trust Calibration in Human-Robot Interaction

HUANG Xinyu, LI Ye
(School of Psychology, Central China Normal University & Key Laboratory of Adolescent Cyberpsychology and Behavior, Ministry of Education, Wuhan 430079, China)

Abstract: Trust forms the foundation of successful human-robot collaboration. However, individuals do not always maintain appropriate trust levels in human-robot interaction and may instead exhibit trust bias—encompassing both over-

trust and under-trust. Such bias impedes effective collaboration, necessitating trust calibration. This process is typically achieved through two pathways: trust dampening, which focuses on reducing excessively high trust in robots, and trust promoting, which aims to elevate insufficient trust levels. Future research should optimize measurement methods for evaluating calibration effectiveness, uncover the cognitive mechanisms underlying trust calibration processes and their aftermath, explore boundary conditions for trust calibration, and develop personalized and refined trust calibration strategies to enhance human-robot cooperation.

Keywords: trust calibration, trust bias, trust dampening, trust promoting, human-robot interaction

Trust permeates the establishment and development of all relationships, including intimate relationships (Rempel et al., 1985), consumer relationships (Kwon et al., 2021), organizational relationships (Meng & Berger, 2019), and doctor-patient relationships (Petrocchi et al., 2019). It serves not only as a crucial factor in interpersonal communication but also as a lubricant for social development (乐国安, 韩振华, 2009). As robots increasingly integrate into daily life, researchers have discovered that trust also exists in human-robot interaction (Hoff & Bashir, 2015; Khavas, 2021). Drawing on previous research (高在峰 et al., 2021; Lee & See, 2004; Mayer et al., 1995), this paper defines human-robot trust as an individual's confidence and psychological expectation that a robot will help achieve their goals or will not exploit their vulnerabilities under conditions of uncertainty or risk. Trust is vital for human-robot interaction, serving as both a prerequisite for human acceptance of algorithms (Sanders et al., 2019) and the foundation for human-robot collaboration (Esterwood & Robert, 2021).

This paper focuses primarily on interactions between humans and intelligent robots, algorithms, and artificial intelligence, with particular emphasis on human-robot interaction. Using intelligent robots as an example, individuals in human-robot interaction may exhibit either excessively high trust (over-trust) or insufficient trust (under-trust). Over-trust leads to inappropriate reliance and misuse of intelligent robots, while under-trust results in disuse. Both undermine the value of human-robot interaction systems (Hancock et al., 2011), necessitating accurate calibration between perceived and actual reliability to maintain appropriate trust levels (Madhavan & Wiegmann, 2007). Well-calibrated trust enables individuals to know when to trust robots and when not to (Ali et al., 2022). Human-robot trust is typically calibrated through two pathways: trust dampening, which reduces unrealistic high trust levels, and trust promoting, which elevates low trust levels. Notably, we term the pathway for elevating low trust "trust promoting" rather than the commonly used "trust repair." While trust repair emphasizes restoring trust after a robot commits a trust violation, it does not address initially low trust levels. In contrast, trust promoting better captures the full scope of this calibration pathway.

Foreign researchers have conducted extensive studies on human-robot trust calibration (Alarcon et al., 2020; de Visser et al., 2020; Ososky et al., 2013), ex-

amining the causes of trust bias and proposing corresponding calibration strategies. However, these studies remain fragmented, lacking systematic integration of empirical research in this field. Moreover, controversies persist regarding the effectiveness of trust calibration strategies, and most previous research has focused on only one aspect of trust bias (either under-trust or over-trust), neglecting the necessity and importance of integrating trust bias and calibration research from a holistic perspective. Therefore, this paper examines the causes of trust bias in human-robot interaction, exploring how robots, individuals, and contextual factors influence trust bias and how trust can be calibrated through dampening and promoting pathways to correct deviations (see Figure 1 [Figure 1: see original paper]). We also aim to clarify the boundary conditions of trust calibration strategies and propose future research directions.

2 Human-Robot Trust Bias

In this paper, we define human-robot trust bias as the deviation of trust from its calibrated value due to individuals' misestimation of robot capabilities, encompassing both over-trust and under-trust. Trust bias causes individuals to trust algorithms that are less reliable than humans or to distrust algorithms that are more reliable than humans (Dzindolet et al., 2003).

2.1 Harms of Human-Robot Trust Bias

Over-trust typically emerges when individuals believe robots possess functions that humans lack or expect robots to help reduce risks (Borenstein et al., 2018; Parasuraman & Riley, 1997). It directly leads to overestimation of robot capabilities and often carries the risk of decision-making errors, such as blindly accepting all decisions proposed by robotic agents without considering their reasonableness (Khavas, 2021; Khavas et al., 2020). In extreme cases, over-trust can threaten human lives. Borenstein et al. (2018) found that although state-of-the-art robotic exoskeletons can only provide limited assistance for children with mobility disabilities under low-speed walking conditions, many parents believed the exoskeletons would protect their children from injury during risky activities like climbing. Similar dangers from over-trusting machines appear in transportation: drivers with high trust in autonomous vehicles are more likely to fall asleep while driving (Kundinger et al., 2019), increasing accident risk.

While less harmful than over-trust, under-trust causes individuals to underestimate algorithmic capabilities (Parasuraman & Riley, 1997), preventing them from leveraging algorithms effectively and enjoying their benefits (Ali et al., 2022), ultimately degrading overall performance and reducing human-robot team efficiency (Okamura & Yamada, 2020).

2.2.1 Robot-Related Factors

Reliability. Performance-related robot characteristics significantly influence initial trust in human-robot interaction (Robinette et al., 2017b). Reliability

refers to the consistency of robot performance (Hancock et al., 2021). A reliable robot should be predictable and stable. However, reliability can induce either over-trust or under-trust (Shi et al., 2020). When people perceive robots as reliable, stable, and predictable, they may relax real-time monitoring and exhibit over-trust; conversely, they may show under-trust. Robot errors trigger trust violations, reducing the trustor's trust intentions or beliefs toward the trustee (严瑜, 吴霞, 2016; Kim et al., 2009). Errors undermine trust primarily because they raise doubts about algorithmic reliability (Alarcon et al., 2020; Correia et al., 2018; Lee & Moray, 1992) and because people are highly sensitive to algorithmic errors, often abandoning them after a single mistake (Dietvorst et al., 2015). Error frequency, severity, and quantity also affect trust dynamics—more frequent, severe, and numerous errors cause faster and greater trust declines (Rossi et al., 2017). Additionally, unexpected, unanticipated behaviors can also violate trust. Lyons et al. (2023) found that when robots deviated from participants' preset routes, both trust perceptions and perceived trustworthiness decreased.

Interestingly, some researchers argue that errors do not necessarily reduce trust. Sarkar et al. (2017) found that errors did not affect perceived robot trustworthiness or subsequent task performance, though they acknowledged influences from task difficulty and error type (cognitive errors that provided incorrect guidance but did not prevent task completion). Fascinatingly, robot errors are sometimes perceived as endearing. After errors, people may find robots more human-like and likable (Mirnig et al., 2017; Salem et al., 2013), while perfect robots may seem unnatural (Biswas & Murray, 2015). This reflects the "pratfall effect" in human-robot interaction. In a rock-paper-scissors game, when robots cheated verbally (claiming victory when they lost) or behaviorally (changing their answer after seeing the human's move), participants showed increased social interaction with the robot and were more amused by its cheating behavior compared to a control group with no cheating, despite subjectively considering cheating unfair (Short et al., 2010).

Embodiment. Embodiment—whether a robot has a physical or virtual form (van Maris et al., 2017)—significantly impacts trust. Physical embodiment refers to robots with tangible, three-dimensional bodies that can move and manipulate the environment (Haring et al., 2021), while virtual embodiment (e.g., virtual robots) appears only on screens with limited mobility. People prefer interacting with physically present robots over virtual agents. Physical embodiment influences trust through social presence, evoking positive attitudes and prompting individuals to treat robots as social actors (Jung & Lee, 2004). Prominently positioned robots increase dependency likelihood (Robinette et al., 2017a). Bainbridge et al. (2011) found that participants were more likely to obey unusual commands from physically present robots than from telepresence robots. In the physical condition, 12 of 22 participants hesitantly but ultimately obeyed instructions to throw books in the trash, compared to only 2-3 participants in the telepresence condition. In this study, obedience was interpreted as trust.

2.2.2 Individual-Related Factors

Motivation. Over-trust may stem from social loafing—individuals exert less effort in human-robot teams than when working alone (Onnasch & Panayotidis, 2020; Parasuraman & Manzey, 2010). Responsibility diffusion in human-robot collaboration can create "free-riding" effects (Dzindolet et al., 2002). Cymek et al. (2023) found that although participants in both solo and human-robot teams self-reported high effort, solo workers demonstrated significantly better performance, suggesting that robot team members relaxed vigilance after observing high robot reliability during the first three-quarters of the task, failing to detect subsequent robot errors.

Self-confidence. When self-confidence exceeds trust in automation, individuals rely on themselves; when self-confidence is low, they depend on automation (Lee & Moray, 1994). The latter situation readily induces over-trust, not only because algorithms are perceived as more authoritative while human agency is weaker (Shank et al., 2021), but also because algorithmic decisions are seen as more reliable and accurate than human decisions (Mosier & Skitka, 1996). In Dijkstra's (1999) study, an algorithmic expert system consistently ruled defendants guilty regardless of case specifics. Despite having better alternatives (e.g., listening to human lawyers), participants ultimately preferred following the algorithmic system's advice, even when incorrect. Participants who complied with the algorithmic system evaluated it more positively and scored higher on authority compliance. Xu et al. (2018) similarly found that people trust robot therapists more than human therapists, with associated over-trust risks.

Algorithmic attitudes. Algorithmic attitudes encompass individuals' cognitive, affective, and behavioral tendencies toward algorithms. Algorithm appreciation and algorithm aversion represent two typical attitudes. Algorithm appreciation prompts individuals to approach algorithmic decisions, leading to over-trust. Logg et al. (2019) found that even when unable to judge decision correctness, participants more readily depended on algorithms than humans when they believed decisions came from algorithms—even when content was identical. This algorithm appreciation effect is consistent across subjective and objective tasks. Excessive trust in robots also implies high performance expectations (Lyons et al., 2020; Shin et al., 2020), which may relate to algorithm appreciation. High expectations lead to higher initial trust, stemming from both appearance-based cognition (e.g., anthropomorphism enhances trust; van Pinxteren et al., 2019) and lack of real interaction experience. In one study, when robots reported "Q-values" (numerical codes) during task completion, participants with and without AI knowledge perceived AI as more reliable, believing that more incomprehensible AI was smarter (Ehsan et al., 2021).

Low trust levels relate to algorithm aversion. Chiarella et al. (2022) found that two paintings by the same artist using different colors received lower aesthetic ratings when attributed to AI rather than humans. Algorithm aversion may result from limited real-world robot exposure combined with sensationalist media

reports about AI threats (e.g., AI world domination, future human-robot wars) that intensify negative attitudes (Demir et al., 2019). It may also stem from negative trust transfer—if individuals have poor experiences with computers or phones, this negative attitude migrates to new algorithm-related products (Lee & Kolodge, 2020; Okuoka et al., 2022).

Mental models. Mental models are organized knowledge structures and cognitive representations of the work environment that people use to predict, explain, and construct expectations about their surroundings (杨正宇 et al., 2003). In human-robot interaction, mental models help individuals infer robots' internal states and predict their capabilities through cues. However, because mental models are based on personal experience and change with new experiences, they vary across individuals (Müller et al., 2023). Trust calibration requires individuals to correctly, comprehensively, and objectively understand robots' strengths and weaknesses—in other words, to possess appropriate mental models for representing and understanding robot capabilities. For example, trust calibration depends on humans appropriately interpreting signals from robots to predict and understand their behavior (Breazeal, 2003). Without appropriate mental models, individuals misestimate robot capabilities, leading to trust bias (Ososky et al., 2013).

2.2.3 Situation-Related Factors

Risk and time pressure. High-risk conditions may increase trust in robots. In Robinette et al.'s (2016) study, participants followed a robot to a conference room via either an inefficient, circuitous route (low-ability robot) or a direct route. After arriving, an alarm sounded, requiring immediate evacuation within one minute. All participants followed the robot, ignoring its previously demonstrated low ability. Time pressure also exacerbates over-trust—when participants perceive time constraints, they are more likely to seek robot assistance despite previously observed errors (Xu & Howard, 2018).

Decision domain characteristics. Researchers suggest that algorithm aversion or appreciation depends on the expertise underlying the intelligent agent (Hou & Jung, 2021). If individuals perceive algorithms as less professionally capable than humans in a domain, algorithm aversion may emerge. In medical diagnosis, for instance, people generally prefer human decisions, partly because human decisions preserve dignity while algorithmic decisions create dehumanizing experiences (Formosa et al., 2022), and partly because people trust humans more for self-disclosure tasks (Barfield, 2021). Decision domain certainty also influences algorithm use. As uncertainty increases, performance differences between humans and algorithms narrow, and people become less tolerant of "perfect" algorithm errors than "imperfect" human errors, leading them to rely on riskier, more error-prone human judgments (Dietvorst & Bharti, 2020). This algorithm bias may also stem from perceptions that algorithmic decisions are less fair, trustworthy, and provoke stronger negative emotions when errors occur (Lee, 2018).

3 Pathways for Human-Robot Trust Calibration

Human-robot trust calibration involves two pathways: trust dampening and trust promoting (see Table 1). Trust dampening refers to activities that reduce unrealistic high trust levels when robots make undetected errors or unexpectedly correct decisions (de Visser et al., 2020). Trust promoting involves activities that positively adjust the trustor's low trust beliefs and intentions, either during initial interaction or after trust violations (Kim et al., 2004). Below, we introduce specific calibration strategies from robot, individual, and environmental perspectives.

3.1 Robot-Related Trust Calibration Strategies

Transparency enhancement. Increasing robot transparency can both reduce over-trust (de Visser et al., 2020) and elevate under-trust (Lyons et al., 2017), though it is more commonly used to correct over-trust. Transparency involves providing users with information about how models work to aid system understanding (Seong & Bisantz, 2008). Algorithms must be understandable, enabling users to comprehend underlying mechanisms and use systems correctly. Transparency also includes revealing robots' inner speech—displaying their reasoning, motivational processes, goals, and action plans to users (Chen et al., 2018; Geraci et al., 2021), thereby dampening trust. Robots can also provide performance feedback, such as verbally communicating uncertainty about decision correctness (Okamura & Yamada, 2020). As noted earlier, predictability is a key component of reliability. When robot performance is unstable or unpredictable, individuals cannot accurately assess reliability. Communicating performance uncertainty—implying potential future performance degradation—helps dampen excessive trust. Beller et al. (2013) used a driving simulation task to test uncertainty effects (an autonomous vehicle displaying a hesitant emoji under uncertain performance). Compared to a control group, the uncertainty group reduced dependence on the autonomous vehicle, prepared users for automation failures, and prompted more active, faster fault handling. Uncertainty group participants also maintained better focus and were less distracted. These results align with Kunze et al. (2019), showing that uncertainty feedback helps users allocate attention and calibrate trust. However, designers must carefully consider uncertainty presentation methods, as while beneficial for trust calibration, they may increase workload and reduce task performance (Kunze et al., 2019).

Displaying confidence indices is another transparency strategy (de Visser et al., 2020). Confidence indices represent the probability of AI making correct decisions—theoretically, people should rely on AI when confidence is high and rely on themselves when it is low. McGuirl and Sarter (2006) found that dynamic confidence indices helped pilots make better decisions about task allocation and compliance with automation, leading to more accurate system accuracy estimates. Similarly, if robots frequently provide confidence indices for task completion, individuals can allocate tasks appropriately.

Explanation. Explainable AI (XAI) is essential for proper trust calibration (Adadi & Berrada, 2018) and a primary trust dampening strategy (Buçinca et al., 2021). XAI provides meaningful explanations and can solicit explanations from users (de Visser et al., 2020), helping users understand AI decision processes so they can identify and reject erroneous decisions. Excessively high expectations of robots may stem from their "black box" nature. Opening this black box before interaction may reduce unrealistic trust and help build appropriate mental models. Wang et al. (2018) found that explanations helped calibrate trust and improve decision-making. Without explanations, participants over-trusted robots and made decision errors; with explanations, compliance decreased. Additionally, communicating robot limitations can correct high reliability expectations, such as explicitly informing users about task and functional capabilities to prevent misuse (de Visser et al., 2020).

When robots make errors, appropriate explanations help individuals understand error mechanisms, strengthen persuasiveness with evidence, and enhance trust. Explanations include describing error causes (Correia et al., 2018), acknowledging errors and providing causal reasoning (Bhatt et al., 2020; Lyons et al., 2023), and proposing solutions (Hald et al., 2021). Explanations must match users' knowledge backgrounds (Adadi & Berrada, 2018; Kim & Hinds, 2006)—overly technical explanations confuse users and reduce transparency. However, explanations can backfire (Papenmeier et al., 2019), with effectiveness depending on error severity. In Correia et al.'s (2018) tangram puzzle task, robot voice failures paused the game. Explanations were only effective when participants could continue the task; when game restart was required, explanations were useless and even decreased trust.

Commitment. Commitment applies to integrity-based or competence-based violations—the former involves honesty issues reducing trust, while latter involves capability failures (严瑜, 吴霞, 2016). Commitment includes both post-violation robot promises to humans and pre-interaction human promises to robots. Esterwood and Robert (2022) found that commitment most effectively repaired trust when individuals held high prior positive attitudes toward robots. By validating individuals' attitudes, commitment reduced cognitive dissonance and facilitated trust repair. Sebo et al. (2019) found that pre-interaction reciprocal commitments—agreeing to fair competition without harming each other—maintained higher trust even when robots cheated during tasks.

Apology. Apology is the most common trust repair method in human-robot interaction, suitable for competence-based violations (Quinn, 2018). Defined as acknowledging responsibility for trust violations and expressing regret (Kim et al., 2004), apology often involves attribution. Kim and Song (2021) found that post-violation, human-like virtual agents using internal attribution apologies repaired trust better than external attribution, while the opposite pattern emerged for machine-like agents. When robots expressed human-like emotions such as regret, trust increased dramatically compared to robots showing no regret; trust increased most when apologies combined regret expression with explanation

(Kox et al., 2021). Timing also matters—Robinette et al. (2015) found that immediate post-violation apologies and explanations were ineffective in simulated fire emergencies, but apologies combined with promises during crises led most participants to follow robots to emergency exits. However, Quinn (2018) questions apology effectiveness, suggesting repeated guilt expression and perceived low sincerity may reduce trust repair efficacy.

Denial. Denial effectively repairs integrity-based violations (Sebo et al., 2019), involving rejection of external causality without admitting responsibility or expressing regret (Kim et al., 2004). Denial gives violators opportunities to refute and question rather than simply admit fault, though it may signal unwillingness to change behavior, raising concerns about future trustworthiness (Kim et al., 2004). However, denial may be safer than apology under high workload when individuals cannot verify robot integrity or clarify fault causes (Quinn, 2018). Interestingly, when robots denied integrity violations, subsequent trust reports showed no differences from other conditions, yet 60% of participants retaliated against the robot (Sebo et al., 2019).

Blame attribution. Blame is a high-risk trust repair strategy. Like apology, it involves attribution, with internal blame (attributing failure to robot-internal causes) preferred over external blame (algorithm designers, third-party algorithms, human partners) (Groom et al., 2010). Post-violation internal blame increases perceived integrity and benevolence compared to external blame, despite no behavioral trust differences (Jensen et al., 2019). However, not all blame is effective—blame emphasizing only errors without explaining causes decreases trust (Kaniarasu & Steinfeld, 2014), as robots blaming participants cause anger, while self-pitying robots seem untrustworthy despite honest error admission.

Anthropomorphism. Anthropomorphism is the psychological process or individual difference of attributing human characteristics, motivations, intentions, or mental states to non-human objects (许丽颖 et al., 2017; Epley et al., 2007), and can promote trust. Since algorithms are perceived as cold and lacking warmth, adding human-like emotional features—such as female robots representing high warmth—may reduce dehumanization perceptions (Borau et al., 2021). Toader et al. (2019) confirmed that participants interacting with female chatbots showed stronger personal information disclosure intentions and higher social perception and service satisfaction than those interacting with male chatbots. Anthropomorphism may increase trust resilience by creating "robots are as fallible as humans" cognition (Aroyo et al., 2021), helping form mental models (Ososky et al., 2013) and slowing trust decline after errors (de Visser et al., 2016). In one study, a communicative robot expressing emotions (e.g., looking 委屈 after dropping an egg) and occasionally making mistakes (dropping one egg during transport) was more favored than silent, efficient robots. Even after errors, participants trusted this robot as much as the efficient, silent robot (Hamacher et al., 2016).

However, anthropomorphism can also induce over-trust. Users may over-trust

highly anthropomorphized robots because they are perceived as more reliable, benevolent, and honest, creating a false sense of familiarity and human-like expectations (Wagner et al., 2018). Therefore, reducing anthropomorphic features can dampen trust in individuals with initially high trust levels.

3.2 Individual-Related Trust Calibration Strategies

Increased contact. Contact can change attitudes toward robots. In trust promotion, the exposure effect exists in human-robot interaction (Jessup et al., 2020; Wullenkord et al., 2016). Face-to-face robot interaction reduces vigilance (Haring et al., 2013), decreases uncertainty and risk perception (Kraus, Scholz, Messner, et al., 2020), and enhances liking and initial trust. Interestingly, simply helping a robot press a button increases trust compared to no-button conditions (Ullman & Malle, 2017). Overall, actual robot contact reduces negative bias and anxiety, corrects inappropriate threat perceptions, and increases future interaction intentions (Wullenkord et al., 2016).

Contact can also reduce inappropriate algorithm appreciation and dampen overtrust. Human-robot interaction experience correlates with automation reliance (Goddard et al., 2012). Haring et al. (2013) found that pre-interaction perceptions of robot intelligence decreased after actual interaction. Wullenkord et al. (2016) replicated this: pre-interaction beliefs about robots' emotional experience decreased after interaction as participants realized robots were less advanced and emotional than imagined, while control groups maintained high capability beliefs. Actual robot contact normalizes capability perceptions (Sanders et al., 2017), achieving trust calibration.

Expectation reduction. Reducing expectations is a trust dampening method. Trust's dynamic nature prompts continuous expectation calibration during interaction, updating robot cognition (Kraus, Scholz, Stiegemeier, et al., 2020). Pop et al. (2015) found that high-expectation users were more sensitive to automation reliability changes but did not necessarily show better trust calibration—calibration was good when automation capability improved but poor when it declined. Therefore, forewarning is effective for high-expectation individuals. Pre-warning about task difficulty and potential poor performance helps reset expectations (de Visser et al., 2020; Lee et al., 2010).

Enhanced algorithmic literacy. Algorithmic literacy comprises four aspects: (1) understanding how apps and platform algorithms are used, (2) knowing how algorithms operate, (3) critically evaluating algorithmic decisions, and (4) effectively handling algorithmic problems (Dogruel et al., 2022). Good algorithmic literacy enables smooth robot interaction, knowledge extraction from explanations, and mental model improvement (Naiseh et al., 2021). Literacy can be enhanced through learning—e.g., user manuals highlighting over-trust risks, robot operators developing training courses (Aroyo et al., 2021)—and through self-learning to update AI knowledge and calibrate trust most effectively.

3.3 Situation-Related Trust Calibration Strategies

Cognitive intervention and resource enhancement. Situational characteristics affect cognitive load. High-risk and time-pressure conditions often cause cognitive overload. According to cognitive load theory, working memory is limited, with intrinsic load from learning tasks and extraneous load from environmental sources (Sweller, 2011). High cognitive load during human-robot interaction prevents accurate automation error detection, while continuous monitoring creates task-unrelated intrinsic load (Lyell & Coira, 2017). Fewer cognitive resources increase algorithm over-trust tendencies (Chien et al., 2016). Optimizing interaction environments may improve resource utilization and dampen trust, such as simplifying user interfaces (Naiseh et al., 2023) and providing clear, understandable instructions (Wickens, 1995).

Fast decision-making contexts are also vulnerable to cognitive heuristics. Buçinca et al. (2021) proposed cognitive intervention strategies based on dual-process theory, which posits two cognitive systems: heuristic thinking (using heuristics and mental shortcuts to reduce cognitive resource consumption) and analytical thinking (slow, resource-intensive, rarely activated). They trained participants in cognitive forcing by requiring them to (1) decide before AI, (2) slow decision-making by increasing AI response time, and (3) choose whether and when to view AI advice. Results showed cognitive intervention increased analytical thinking motivation and reduced AI over-reliance.

Enhancing robot (algorithm) advantages in decision domains. Hou and Jung (2021) argue that humans do not uniformly prefer algorithms or humans but rather the expertise behind them. Injecting expert power behind algorithmic decisions can improve negative trust attitudes. Matching robot appearances to task domains also aids acceptance—e.g., child-like, high-warmth robots are preferred in hedonic service contexts, while adult-like, high-competence robots are preferred in utilitarian contexts (Liu et al., 2022). People generally dislike algorithms for subjective tasks, but emphasizing objective, fact-based components can reduce subjectivity and increase trust. For instance, informing participants that music recommendations (a subjective task) can be determined by personality traits (objective factors) enhances trust in algorithmic decisions (Castelo et al., 2019).

4 Future Research Directions

Human-robot interaction has permeated daily life, and trust is crucial for team cohesion (Perkins et al., 2021). However, only appropriately calibrated trust promotes effective collaboration—excessive or insufficient trust threatens cooperation, necessitating calibration. Despite substantial research progress, current work has limitations and improvement opportunities.

4.1 Optimizing Measurement Methods for Calibration Effectiveness

First, regarding measurement, researchers have begun using implicit methods to examine trust in automation (Merritt et al., 2013), but these remain limited to pre-calibration trust measurement, not addressing post-calibration implicit trust after bias correction. Second, post-calibration evaluation relies primarily on explicit trust indicators like scale scores or trust behaviors. We argue that both explicit and implicit trust attitudes should be examined to better test calibration strategy effectiveness. For trust dampening, future research could use implicit association tests to examine whether dampening strategies reduce implicit trust levels and compare differences between explicit and implicit trust reduction.

4.2 Revealing Cognitive-Neural Processes in Trust Calibration

Most human-robot trust research has focused on behavioral experiments, though some have begun examining cognitive-neural perspectives (Eloy et al., 2022; Oh et al., 2020; Walker et al., 2019; Yen & Chiang, 2021). For example, increased neural activation in medial and right dorsolateral prefrontal cortex and decreased functional connectivity were observed after robot errors (Hopko & Mehta, 2022), with increased negative waves in anterior cingulate cortex (de Visser et al., 2018). Trust is a continuous process where establishment, growth, damage, and dissolution powerfully and persistently affect all members' current and future behaviors (Hancock et al., 2011). Complete trust calibration cycles involve trust establishment-growth/damage-calibration stages. Previous cognitive-neural research has focused on the first two stages, but the cognitive-neural processes involved during and after calibration—particularly after trust promotion or dampening—remain underexplored. This is crucial for revealing physiological mechanisms and optimizing calibration strategies. Future research could use physiological indicators for real-time, continuous monitoring of cognitive-neural activity from initial trust establishment through post-calibration, revealing trust dynamics at the physiological level.

4.3 Fine-Grained Research by Trust Development Stage

Trust exhibits dynamic development, yet most calibration research uses static cross-sectional designs examining only current-stage trust elevation or dampening, not dynamic trust development factors. For example, pre-interaction negative algorithm attitudes and low capability expectations may be disrupted when individuals perceive algorithmic reliability during interaction, creating expectation gaps that increase trust (Washburn et al., 2020). Filiz et al. (2021) found that in 40 rounds of stock price prediction, participants initially trusting themselves gradually shifted to trusting algorithms as they observed higher algorithmic accuracy. This was replicated when robot journalists' article quality exceeded expectations—positive disconfirmation increased acceptance and satisfaction (Kim & Kim, 2021). High trust from such positive disconfirmation may differ from gradually accumulated high trust during interaction, warrant-

ing different dampening strategies that previous research has not distinguished. Future work should compare calibration strategy effectiveness for different bias causes and identify optimal strategies for specific trust biases.

4.4 Exploring Boundary Conditions of Trust Calibration

First, most research examines trust changes with humanoid or mechanized robots, rarely addressing animal-like robots, especially "cute" ones that may automatically evoke naive, kind trait inferences and positive emotions (许丽颖 et al., 2019). Baby-faced robots with large, round eyes may also seem more trustworthy (Song & Luximon, 2020). Animal-like robots are more likable than mechanized robots (Li et al., 2010). Since human uniqueness may contribute to low initial trust, cute animal-like robots might reduce threat perception and increase initial trust; after trust violations, they may also slow trust decline and facilitate repair. For trust dampening, animal-like robots may reduce expectations through familiarity while avoiding racial bias issues in humanoid design (Löffler et al., 2020), thus dampening over-trust. They may also reduce mental model inferences and dampen trust. Future research should compare humanoid and animal-like robots in trust calibration, though animal-like robots should have high or low animal similarity to avoid the uncanny valley effect (Löffler et al., 2020).

Second, research has begun examining trust development in groups rather than individual interactions (Montague & Xu, 2012; Montague et al., 2014; Xu & Montague, 2013). Martinez et al. (2023) found that while individual trust in food delivery robots increased with contact, group trust showed more variation without consistent growth. Volante et al. (2019) found people prefer following others' robot evaluations rather than direct interaction. These studies explore group trust changes but not group trust calibration. Future cross-cultural research could compare Chinese and Western participants' group trust differences, examine group trust calibration, and compare individual versus group trust bias to identify appropriate group calibration strategies.

Finally, calibration success depends on individual factors, with strategy effectiveness showing individual differences. Lee et al. (2010) found that relationship-oriented participants preferred apology strategies, while utilitarian-oriented participants preferred compensation strategies. Future research could model trust-related behaviors during interaction (Pynadath et al., 2019) to calibrate trust personalized to user characteristics.

Acknowledgments

We thank Dr. Zhao Wen from Southwest University of Science and Technology for polishing the English abstract, and two anonymous reviewers and the editorial board for their valuable comments and suggestions.

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), 2172 2183.

许丽颖, 喻丰, 邬家骅, 韩婷婷, 赵靓. (2017). 拟人化: 从“它”到“他”. 心理科学进展, 25(11),1942 1954.

许丽颖, 喻丰, 周爱钦, 杨沈龙, 丁晓军.(2019). 萌: 感知与后效. 心理科学进展, 27(4), 689 699.

严瑜, 吴霞.(2016). 从信任违背到信任修复: 道德情绪的作用机制. 心理科学进展, 24(4), 633 642.

杨正宇, 王重鸣, 谢小云. (2003). 团队共享心理模型研究新进展. 人类工效学, 9(3), 34 37.

乐国安, 韩振华. (2009). 信任的心理学研究与展望. 西南大学学报 (社会科学版), 35(2), 1 5.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138 52160.

Alarcon, G. M., Gibson, A. M., & Jessup, S. A. (2020, September). Trust repair in performance, process, and purpose factors of human-robot trust. In 2020 IEEE International Conference on Human-Machine Systems (ICHMS) (pp. 1 6). Rome, Italy.

Ali, A., Tilbury, D. M., & Jr, L. R. (2022). Considerations for task allocation in human-robot teams. arXiv preprint arXiv:2210.03259.

Aroyo, A. M., De Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., … & Tamò-Larrieux, A. (2021). Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. Paladyn, Journal of Behavioral Robotics, 12(1), 423 436.

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. International Journal of Social Robotics, 3, 41−52.

Barfield, J. K. (2021, August). Self-disclosure of personal information, robot appearance, and robot trustworthiness. In 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN) (pp. 67 72). Vancouver, BC, Canada.

Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the driver-automation interaction: An approach using automation uncertainty. Human Factors, 55(6), 1130 1141.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., … & Eckersley, P. (2020, January). Explainable machine learning in deployment. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 648 657). https://doi.org/10.1145/3351095.3375624.

Biswas, M., & Murray, J. C. (2015, September). Towards an imperfect robot for long-term companionship: Case studies using cognitive biases. In 2015

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 5978 5983). Hamburg, Germany.

Borau, S., Otterbring, T., Laporte, S., & Fosso Wamba, S. (2021). The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. Psychology & Marketing, 38(7), 1052 1068.

Borenstein, J., Wagner, A. R., & Howard, A. (2018). Overtrust of pediatric health-care robots: A preliminary survey of parent perspectives. IEEE Robotics & Automation Magazine, 25(1), 46 54.

Breazeal, C. (2003). Toward sociable robots. Robotics and Autonomous Systems, 42, 167 175.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5, CSCW1, 1 21.

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

Chen, J. Y., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theoretical Issues in Ergonomics Science, 19(3), 259 282.

Chiarella, S. G., Torromino, G., Gagliardi, D. M., Rossi, D., Babiloni, F., & Cartocci, G. (2022). Investigating the negative bias towards Artificial Intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings. Computers in Human Behavior, 137, 107406.

Chien, S. Y., Lewis, M., Sycara, K., Liu, J. S., & Kumru, A. (2016, October). Influence of cultural factors in dynamic trust in automation. In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 2884 2889). Budapest, Hungary.

Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., & Paiva, A. (2018, July). Exploring the impact of fault justification in human-robot trust. In Proceedings of the 17th international conference on autonomous agents and multiagent systems (pp. 507 513). Stockholm, Sweden.

Cymek, D. H., Truckenbrodt, A., & Onnasch, L. (2023). Lean back or lean in? Exploring social loafing in human-robot teams. Frontiers in Robotics and AI, 10, 1249252. doi: 10.3389/frobt.2023.1249252.

Demir, K. A., Döven, G., & Sezen, B. (2019). Industry 5.0 and human-robot co-working. Procedia Computer Science, 158, 688 695.

de Visser, E. J., Beatty, P. J., Estepp, J. R., Kohn, S., Abubshait, A., Fedota, J. R., & McDonald, C. G. (2018). Learning from the slips of others: Neural

correlates of trust in automated agents. Frontiers in Human Neuroscience, 12, 309.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., 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.

de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. International Journal of Social Robotics, 12(2), 459−478.

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychological Science, 31(10), 1302−1314.

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

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. Behaviour & Information Technology, 18(6), 399−411.

Dogruel, L., Masur, P., & Joeckel, S. (2022). Development and validation of an algorithm literacy scale for internet users. Communication Methods and Measures, 16(2), 115 133.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. International Journal of Human-Computer Studies, 58(6), 697−718.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. Human Factors, 44(1), 79−94.

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I., Muller, M., & Riedl, M. O. (2021). The who in explainable AI: How AI background shapes perceptions of AI explanations. arXiv preprint arXiv:2107.13509.

Eloy, L., Doherty, E. J., Spencer, C. A., Bobko, P., & Hirshfield, L. (2022). Using fNIRS to identify transparency-and reliability-sensitive markers of trust across multiple timescales in collaborative human-human-agent triads. Frontiers in Neuroergonomics, 3, 838625.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. Psychological Review, 114, 864 886.

Esterwood, C., & Robert, L. P. (2021, August). Do you still trust me? Human-robot trust repair strategies. Proceedings of 30th IEEE International Conference on Robot and Human Interactive Communication. Vancouver, BC, Canada.

Esterwood, C., & Robert, L. P. (2022, March). Having the right attitude: How attitude impacts trust repair in human-robot interaction. In 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 332 341). Sapporo, Japan.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021). Reducing algorithm aversion through experience. Journal of Behavioral and Experimental Finance, 31, 100524.

Formosa, P., Rogers, W., Griep, Y., Bankins, S., & Richards, D. (2022). Medical AI and human dignity: Contrasting perceptions of human and artificially intelligent (AI) decision making in diagnostic and medical resource allocation contexts. Computers in Human Behavior, 133, 107296.

Geraci, A., D'Amico, A., Pipitone, A., Seidita, V., & Chella, A. (2021). Automation inner speech as an anthropomorphic feature affecting human trust: Current issues and future directions. Frontiers in Robotics and AI, 8, 620026.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121 127.

Groom, V., Chen, J., Johnson, T., Kara, F. A., & Nass, C. (2010, March). Critic, compatriot, or chump? Responses to robot blame attribution. In 2010 5th ACM/IEEE international conference on human-robot interaction (HRI) (pp. 211 217). IEEE.

Hald, K., Weitz, K., André, E., & Rehm, M. (2021, November). "An Error Occurred!" Trust repair with virtual robot using levels of mistake explanation. In Proceedings of the 9th International Conference on Human-Agent Interaction (pp. 218 226). Virtual Event Japan.

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

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2021). Evolving trust in robots: Specification through sequential and comparative meta-analyses. Human Factors, 63(7), 1196—1229.

Hamacher, A., Bianchi-Berthouze, N., Pipe, A. G., & Eder, K. (2016, August). Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. In 2016 25th IEEE international symposium on robot and human interactive communication(RO-MAN) (pp. 493 500). New York, NY, USA.

Haring, K. S., Matsumoto, Y., & Watanabe, K. (2013). How do people perceive and trust a lifelike robot. In Proceedings of the world congress on engineering and computer science (pp. 425 430). San Francisco, USA.

Haring, K. S., Satterfield, K. M., Tossell, C. C., De Visser, E. J., Lyons, J. R.,

Mancuso, V. F., … & Funke, G. J. (2021). Robot authority in human-robot teaming: Effects of human-likeness and physical embodiment on compliance. Frontiers in Psychology, 12, 625713.

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

Hopko, S. K., & Mehta, R. K. (2022). Trust in shared-space collaborative robots: Shedding light on the human brain. Human Factors, in press. https://doi.org/10.1177/00187208221109039.

Hou, Y. T. Y., & Jung, M. F. (2021). Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. Proceedings of the ACM on Human-Computer Interaction, 5, CSCW2, 477.

Jensen, T., Albayram, Y., Khan, M. M. H., Fahim, M. A. A., Buck, R., & Coman, E. (2019, June). The apple does fall far from the tree: User separation of a system from its developers in human-automation trust repair. In Proceedings of the 2019 on Designing Interactive Systems Conference (pp. 1071 1082). San Diego, CA, USA.

Jessup, S. A., Gibson, A., Capiola, A. A., Alarcon, G. M., & Borders, M. (2020, January). Investigating the effect of trust manipulations on affect over time in human-human versus human-robot interactions. Proceedings of the 53rd Hawaii International Conference on System Sciences(pp. 1 10).

Jung, Y., & Lee, K. M. (2004). Effects of physical embodiment on social presence of social robots. Proceedings of PRESENCE, 80 87.

Kaniarasu, P., & Steinfeld, A. M. (2014, August). Effects of blame on trust in human robot interaction. In The 23rd IEEE international symposium on robot and human interactive communication (pp. 850 855). Edinburgh, Scotland, UK.

Khavas, Z. R. (2021). A review on trust in human-robot interaction. arXiv preprint arXiv:2105.10045.

Khavas, Z. R., Ahmadzadeh, S. R., & Robinette, P. (2020, November). Modeling trust in human-robot interaction: A survey. In Social Robotics: 12th International Conference, ICSR (pp. 529 541). https://doi.org/10.1007/978-3-030-62056-1_{44}.

Kim, D., & Kim, S. (2021). A model for user acceptance of robot journalism: Influence of positive disconfirmation and uncertainty avoidance. Technological Forecasting and Social Change, 163, 120448.

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. Academy of Management Review, 34, 401 422.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing

competence-versus integrity-based trust violations. Journal of Applied Psychology, 89(1), 104−118.

Kim, T., & Hinds, P. (2006, September). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In ROMAN 2006-The 15th IEEE international symposium on robot and human interactive communication (pp. 80 85). Hatfield, UK.

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. Telematics and Informatics, 61, 101595.

Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust repair in human-agent teams: The effectiveness of explanations and expressing regret. Autonomous Agents and Multi-Agent Systems, 35(2), 30.

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

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. Human Factors, 62(5), 718 736.

Kundinger, T., Wintersberger, P., & Riener, A. (2019, May). (Over) Trust in automated driving: The sleeping pill of tomorrow? In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1 6). Glasgow, Scotland UK.

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. Ergonomics, 62(3), 345 360.

Kwon, J. H., Jung, S. H., Choi, H. J., & Kim, J. (2021). Antecedent factors that affect restaurant brand trust and brand loyalty: Focusing on US and Korean consumers. Journal of Product & Brand Management, 30(7), 990 1015.

Lee, J. D., & Kolodge, K. (2020). Exploring trust in self-driving vehicles through text analysis. Human Factors, 62(2), 260 277.

Lee, J.D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35(10), 1243−1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. International Journal of Human-Computer Studies, 40(1), 153−184.

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

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society, 5(1), 1 16.

Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010, March). Gracefully mitigating breakdowns in robotic services. In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 203 210). Osaka, Japan.

Lee, S. L., Lau, I. Y. M., Kiesler, S., & Chiu, C. Y. (2005, April). Human mental models of humanoid robots. In Proceedings of the 2005 IEEE international conference on robotics and automation (pp. 2767 2772). Barcelona, Spain.

Li, D., Rau, P. P., & Li, Y. (2010). A cross-cultural study: Effect of robot appearance and task. International Journal of Social Robotics, 2, 175−186.

Liu, X. S., Yi, X. S., & Wan, L. C. (2022). Friendly or competent? The effects of perception of robot appearance and service context on usage intention. Annals of Tourism Research, 92, 103324.

Löffler, D., Dörrenbächer, J., & Hassenzahl, M. (2020, March). The uncanny valley effect in zoomorphic robots: The U-shaped relation between animal likeness and likeability. In Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction (pp. 261 270). Cambridge, United Kingdom.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90−103.

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. Journal of the American Medical Informatics Association, 24(2), 423 431.

Lyons, J. B., Hamdan, I., & Vo, T. Q. (2023). Explanations and trust: What happens to trust when a robot partner does something unexpected? Computers in Human Behavior, 138, 107473.

Lyons, J. B., Nam, C. S., Jessup, S. A., Vo, T. Q., & Wynne, K. T. (2020, September). The role of individual differences as predictors of trust in autonomous security robots. In 2020 IEEE International Conference on Human-Machine Systems (ICHMS) (pp. 1 5). Rome, Italy.

Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., … & Shively, R. (2017). Shaping trust through transparent design: theoretical and experimental guidelines. In: Savage-Knepshield, P., & Chen, J(Eds.), Advances in Human Factors in Robots and Unmanned Systems(pp.127 136). Springer International Publishing.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. Theoretical Issues in Ergonomics Science, 8(4), 277−301.

Martinez, J. E., VanLeeuwen, D., Stringam, B. B., & Fraune, M. R. (2023, March). Hey? ! What did you think about that robot? Groups polarize users' acceptance and trust of food delivery robots. In Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (pp. 417 427). https://doi.org/10.1145/3568162.3576984.

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

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Human Factors, 48(4), 656−665.

Meng, J., & Berger, B. K. (2019). The impact of organizational culture and leadership performance on PR professionals' job satisfaction: Testing the joint mediating effects of engagement and trust. Public Relations Review, 45(1), 64 75.

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

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. Frontiers in Robotics and AI, 4, 21.

Montague, E., & Xu, J. (2012). Understanding active and passive users: The effects of an active user using normal, hard and unreliable technologies on user assessment of trust in technology and co-user. Applied Ergonomics, 43(4), 702 712.

Montague, E., Xu, J., & Chiou, E. (2014). Shared experiences of technology and trust: An experimental study of physiological compliance between active and passive users in technology-mediated collaborative encounters. IEEE Transactions on Human-Machine Systems, 44(5), 614 624.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other?. In Parasuraman, R., & Mouloua. M(Eds.), Automation and Human Performance(pp.201 220). CRC Press.

Müller, R., Schischke, D., Graf, B., & Antoni, C. H. (2023) How can we avoid information overload and techno-frustration as a virtual team? The effect of shared mental models of information and communication technology on information overload and techno-frustration. Computers in Human Behavior, 138, 107438.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. International Journal of Human-Computer Studies, 169, 102941.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2021). Explainable recommendation: When design meets trust calibration. World Wide Web, 24(5), 1857 1884.

Oh, S., Seong, Y., Yi, S., & Park, S. (2020). Neurological measurement of human trust in automation using electroencephalogram. International Journal of Fuzzy Logic and Intelligent Systems, 20(4), 261 271.

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. Plos One, 15(2), e0229132. https://doi.org/10.1371/journal.pone.0229132.

Okuoka, K., Enami, K., Kimoto, M., & Imai, M. (2022). Multi-device trust transfer: Can trust be transferred among multiple devices? Frontiers in Psychology, 13, 920844.

Onnasch, L., & Panayotidis, T. (2020, December). Social loafing with robots-An empirical investigation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64(1), 97 101.

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013, March). Building appropriate trust in human-robot teams. In 2013 AAAI spring symposium series.

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. arXiv preprint arXiv:1907.12652.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381−410.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230−253.

Perkins, R., Khavas, Z. R., & Robinette, P. (2021). Trust calibration and trust respect: A method for building team cohesion in human robot teams. arXiv preprint arXiv:2110.06809.

Petrocchi, S., Iannello, P., Lecciso, F., Levante, A., Antonietti, A., & Schulz, P. J. (2019). Interpersonal trust in doctor-patient relation: Evidence from dyadic analysis and association with quality of dyadic communication. Social Science & Medicine, 235, 112391.

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. Human Factors, 57(4), 545 556.

Pynadath, D. V., Wang, N., & Kamireddy, S. (2019, September). A Markovian Method for Predicting Trust Behavior in Human-Agent Interaction. In Proceedings of the 7th International Conference on Human-Agent Interaction (pp. 171 178). Kyoto, Japan.

Quinn, D. B. (2018). Exploring the efficacy of social trust repair in human-automation interactions (Doctoral dissertation). Clemson University, Lawton.

Ragni, M., Rudenko, A., Kuhnert, B., & Arras, K. O. (2016, August). Errare humanum est: Erroneous robots in human-robot interaction. In 2016 25th

IEEE International symposium on robot and human interactive communication (RO-MAN) (pp. 501 506). New York, NY, USA.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. Journal of Personality and Social Psychology, 49(1), 95 112.

Robinette, P., Howard, A. M., & Wagner, A. R. (2015, October). Timing is key for robot trust repair. In Social Robotics: 7th International Conference, ICSR, Paris, France.

Robinette, P., Howard, A. M., & Wagner, A. R. (2017a). Conceptualizing overtrust in robots: Why do people trust a robot that previously failed?. In: Lawless, W., Mittu, R., Sofge, D., & Russell, S (Eds), Autonomy and Artificial Intelligence: A Threat or Savior?(pp. 129 155). Springer, Cham.

Robinette, P., Howard, A. M., & Wagner, A. R. (2017b). Effect of robot performance on human-robot trust in time-critical situations. IEEE Transactions on Human-Machine Systems, 47(4), 425−436.

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 (HRI) (pp. 101 108). Christchurch, New Zealand.

Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2017, November). Human perceptions of the severity of domestic robot errors. In Social Robotics: 9th International Conference(ICSR)(pp. 647 656). Tsukuba, Japan.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. International Journal of Social Robotics, 5, 313−323.

Sanders, T. L., Kaplan, A., Koch, R., Schwartz, M., & Hancock, P. A. (2019). The relationship between trust and use choice in human-robot interaction. Human Factors, 61(4), 614−626.

Sanders, T. L., MacArthur, K., Volante, W., Hancock, G., MacGillivray, T., Shugars, W., & Hancock, P. A. (2017, September). Trust and prior experience in human-robot interaction. In Proceedings of the human factors and ergonomics society annual meeting (pp. 1809 1813). Sage CA: Los Angeles, CA.

Sarkar, S., Araiza-Illan, D., & Eder, K. (2017). Effects of faults, experience, and personality on trust in a robot co-worker. arXiv preprint arXiv:1703.02335.

Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019, March). "I don't believe you": Investigating the effects of robot trust violation and repair. In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 57 65). Daegu, Korea (South).

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. International Journal of Industrial Ergonomics, 38(7-8), 608−625.

Shank, D. B., Bowen, M., Burns, A., & Dew, M. (2021). Humans are perceived as better, but weaker, than artificial intelligence: A comparison of affective impressions of humans, AIs, and computer systems in roles on teams. Computers in Human Behavior Reports, 3, 100092.

Shi, Y., Azzolin, N., Picardi, A., Zhu, T., Bordegoni, M., & Caruso, G. (2020). A Virtual reality-based platform to validate HMI design for increasing user' s trust in autonomous vehicle. Computer-Aided Design and Applications, 18(3), 502 518.

Shin, D., Zaid, B., & Ibahrine, M. (2020, November). Algorithm appreciation: Algorithmic performance, developmental processes, and user interactions. In 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI) (pp. 1 5). Sharjah, United Arab Emirates.

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010, March). No fair! An interaction with a cheating robot. In 2010 5th ACM/IEEE international conference on human-robot interaction (HRI) (pp. 219 226). Osaka, Japan.

Song, Y., & Luximon, Y. (2020). Trust in AI agent: A systematic review of facial anthropomorphic trustworthiness for social robot design. Sensors, 20(18), 5087.

Sweller, J. (2011). Cognitive load theory. Psychology of Learning and Motivation, 55, 37 76. https://doi.org/10.1016/B978-0-12-387691-1.00002-8.

Tam, K. Y., & Ho, S. Y. (2005). Web personalization as a persuasion strategy: An elaboration likelihood model perspective. Information Systems Research, 16(3), 271 291.

Toader, D. C., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., & Rădulescu, A. T. (2019). The effect of social presence and chatbot errors on trust. Sustainability, 12(1), 256.

Ullman, D., & Malle, B. F. (2017, March). Human-robot trust: Just a button press away. In Proceedings of the companion of the 2017 ACM/IEEE international conference on human-robot interaction (pp. 309 310). Vienna, Austria.

van Maris, A., Lehmann, H., Natale, L., & Grzyb, B.(2017, March). The influence of a robot' s embodiment on trust: A longitudinal study. In Proceedings of the Companion of the 2017 ACM/IEEE International Conference on human-robot interaction (pp. 313 314). Vienna, Austria.

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

Volante, W. G., Sosna, J., Kessler, T., Sanders, T., & Hancock, P. A. (2019). Social conformity effects on trust in simulation-based human-robot interaction. Human Factors, 61(5), 805 815.

Wagner, A. R., Borenstein, J., & Howard, A. (2018). Overtrust in the robotic age. Communications of the ACM, 61(9), 22 24.

Walker, F., Wang, J., Martens, M. H., & Verwey, W. B. (2019). Gaze behaviour and electrodermal activity: Objective measures of drivers' trust in automated vehicles. Transportation research part F: Traffic psychology and behaviour, 64, 401 412.

Wang, N., Pynadath, D.V., Rovira, E., Barnes, M.J., Hill, S.G. (2018). Is it my looks? Or something i said? The impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In: Ham, J., Karapanos, E., Morita, P., & Burns, C(Eds), Persuasive Technology (pp. 56 69). Springer, Cham.

Washburn, A., Adeleye, A., An, T., & Riek, L. D. (2020). Robot errors in proximate HRI: How functionality framing affects perceived reliability and trust. ACM Transactions on Human-Robot Interaction (THRI), 9(3), 1 21.

Wickens, C. D. (1995). Designing for situation awareness and trust in automation. IFAC Proceedings Volumes, 28(23), 365 370.

Wullenkord, R., Fraune, M. R., Eyssel, F., & Šabanović, S. (2016, August). Getting in touch: How imagined, actual, and physical contact affect evaluations of robots. In 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 980 985). New York, USA.

Yen, C., & Chiang, M. C. (2021). Trust me, if you can: A study on the factors that influence consumers' purchase intention triggered by chatbots based on brain image evidence and self-reported assessments. Behaviour & Information Technology, 40(11), 1177 1194.

Xu, J., De'Aira, G. B., & Howard, A. (2018, August). Would you trust a robot therapist? Validating the equivalency of trust in human-robot healthcare scenarios. In 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 442 447). Nanjing, China.

Xu, J., & Howard, A. (2018, August). The impact of first impressions on human-robot trust during problem-solving scenarios. In 2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN) (pp. 435 441). Nanjing, China.

Xu, J., & Montague, E. (2013, September). Group polarization of trust in technology. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (pp. 344 348). Sage CA: Los Angeles, CA.

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