Trust(worthiness) Issues with Trust in Human-Robot Interaction

LINDA ONNASCH*, Technische Universität Berlin, Germany
EILEEN ROESLER*, George Mason University, United States
LIONEL ROBERT, University of Michigan, United States
EWART DE VISSER, United States Air Force Academy, United States

Trust is a very popular concept in human-robot interaction (HRI) to explain why and how people interact with robots. However, the definition of trust and the methods used to study the concept vary widely, often leading to confusion instead of insight. In this position paper, we discuss possible reasons for the confusion and disagreement by reviewing theory and methods. Our main criticism is that HRI researchers have recently taken an oversimplified approach by not adhering to the process model of trust. Instead, we have primarily measured perceived trustworthiness (often as a proxy for trust) and assumed that it accurately predicts behavior or is satisfactory as an end goal. In addition, many experimental paradigms fail to account for the critical elements of risk and vulnerability that are essential for trust to guide behavior. With this position paper, we aim to shed light on these "trust issues" in HRI and to improve the HRI community's approach by providing suggestions for enhancing the quality of future research.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; • **Hardware →** *Analysis and design of emerging devices and systems*; • **Computer systems organization → Robotics**;

1 Trust Research: Been There, Done That, Still Struggling

Pointing to research on trust in automation [75], early work already suggested trust as an important construct for human-robot interaction (HRI) [48, 124]. Since then, it appears that trust as a topic of inquiry has steadily grown. This is primarily related to the assumption that trust is critical to understanding how and why humans (do not) interact with robots, regardless of whether we focus on industrial or social HRI. Trust in robots is associated with safe, reliable, and meaningful interactions, resulting in successful integration and user satisfaction [14, 29, 46, 50, 110, 139]. In line with the concept's key role, books, book chapters, workshops, numerous scientific papers, and review articles have addressed the topic of how to build trust into HRI [33, 35, 54, 68, 74, 77, 93, 111]. And so do these authors. We have all been researching trust in human-human and human-automation interaction for years [23, 28, 57, 62, 67, 82, 101, 103, 107, 129], and extended this work — more or less recently — to robots. In the past decades, we have explored the impact of several factors that we and others believe to influence trust, such as failures and anthropomorphism [27, 94, 96, 108], aspects of the human [54] as well as strategies to repair trust [41], to name a few.

Yet, we often struggle to reach a consensus when discussing what trust is about and whether (and why) we measure it at all. The term "trust" is so common and universally understood in everyday language that it might lead to the assumption that the scientific exploration of trust is just as straightforward. The fact that this is not

*Shared first authorship.

Authors' Contact Information: Linda Onnasch*, Technische Universität Berlin, Berlin, Germany, linda.onnasch@tu-berlin.de; Eileen Roesler*, George Mason University, Fairfax, United States, eroesle@gmu.edu; Lionel Robert, University of Michigan, Ann Arbor, United States, lprobert@umich.edu; Ewart de Visser, United States Air Force Academy, Colorado Springs, United States, ewartdevisser@gmail.com.

ACM 2573-9522/2025/12-ART
https://doi.org/10.1145/3778865

the case becomes obvious in many places and is at the heart of the problem. Looking back at our own research provides several examples for the fuzziness of trust research in HRI. For example, we used to write about trust behavior, but how can trust, conceptually defined as a latent construct, be represented in behavioral measures [26, 27, 96, 108]? We investigated trust in situations that had little risk and vulnerability, although these make trust necessary in the first place [40]. We claimed to have measured trust, or was it trustworthiness [107]?

We believe that we are not the only ones having "trust issues", especially when we recall several discussions with other researchers and consider the extensive literature on trust in HRI. In addition, a recent commentary has questioned the usefulness of trust [10, 11], while some have called for better measurement [20, 67, 120, 126], and yet others have listed many unresolved research questions and inconsistencies in trust research [6].

You may disagree with these "trust issues", just as some of the authors disagreed with each other, which was the impetus for this article. Following our initial heated debate during an HRI trust workshop [35], we faced the question posed by [32]: "How is it possible for individuals with similar levels of training, experience, and background, working with the same dataset, to arrive at different conclusions on a particular issue?" (p. 6011). This paper tries a response to that question. After reviewing the origins of our disagreements, we have reached the consensus that our HRI trust research may indeed have underlying issues. To resolve, or at least better understand these issues, we first revisit the original trust model proposed by Mayer et al. [83] to give necessary background and context information, before discussing the "trust issues" in detail and suggesting possible solutions.

2 Back To The Roots: Revisiting The Original Trust Model

In this paper, our aim is not to introduce a new trust model, as many esteemed colleagues in HRI and related disciplines have already done [16, 54, 71, 72]. Adding more variables may expand use cases and account for exceptions, but it does not address the fundamental issues.

To clarify our conceptual basics we draw on the most widely accepted and established model by Mayer et al. [83]. The model was originally developed for human-human trust but readily applies to other human-agent relationships, such as human-automation or HRI. Most trust models proposed for these specific domains are based on this model [28, 75, 81]. To avoid the need to describe all, we focus entirely on this trust model (see Fig.1) to illustrate the confusion the HRI community faces.

The model predicts trust via trustworthiness and propensity to trust. The formed trust leads to risk-taking behavior in relationships which leads to outcomes. The outcome of this process in turn influences again the perceived trustworthiness and the following trust formation. The model highlights perceived situational risk as a moderator of the relationship between trust and risk-taking behavior.

Although the model is widely cited, it often seems not to be thoroughly examined. To address this, we first seek to establish a shared knowledge base around the issues we later identify and briefly describe each aspect of the model, already in relation to HRI.

2.1 Perceived Trustworthiness

The first part of the model describes a trustworthiness assessment concerning features of a robot that may impact trust and risk downstream. Perceived trustworthiness is a subjective perception of the trustee based on the robot's characteristics (actual trustworthiness) [117, 118]. The original model included the trustworthiness factors ability, benevolence, and integrity. These factors were adapted for automation into performance, process, and purpose [75] and later adapted for HRI by re-classifying them into performance (capable, reliable) and moral (ethical, transparent, benevolent) dimensions [81]. Regardless of their different labels, these factors are highly consistent across human-human, human-automation, and human-robot interaction fields.

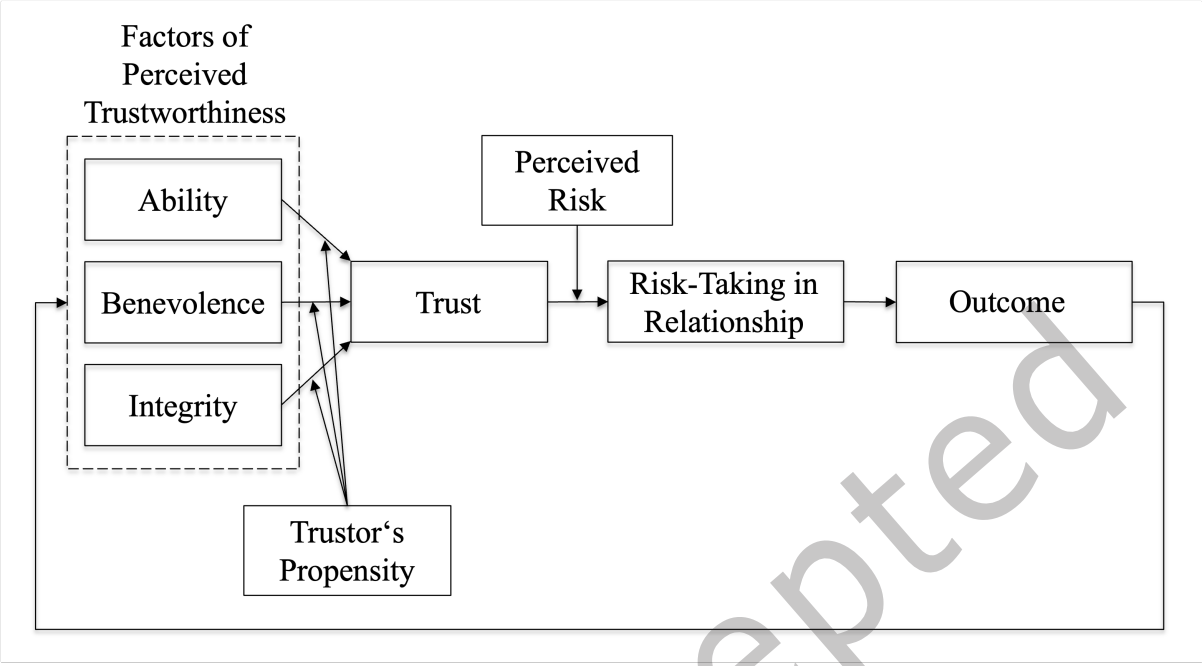

Fig. 1. The originally proposed trust model [83].

2.2 Trustor's Propensity

The relationship between trustworthiness and trust is moderated by the general human tendency to trust. People with a high propensity may have more trust in a specific robot. The defining characteristic of this factor is that it is a relatively stable trait over time, unlike trust. A review by Hoff and Bashir [59] revealed four primary sources of dispositional variability for trust propensity: culture, age, gender, and personality. What becomes evident here is that the propensity to trust is an exclusively individual characteristic of the trustor and completely independent of robot characteristics and situational factors. Trust propensity may not only directly influence trust, but also alter the relationship between perceived trustworthiness and trust.

2.3 Trust

Trust itself is defined [83, p. 712] as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party". Similarly, Lee & See [75, p. 51] also define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability". The adaptation of these definitions to HRI is straightforward as the willingness of a human to be vulnerable to the actions of a robot given that the action is important to that human irrespective of the ability to monitor the robot. Trust happens before any observable human behavior as a consequence of this attitude.

2.4 Perceived Risk

The path from trust to observable behavior is influenced by the situational factor perceived risk. Even though the level of trust (as determined by perceived trustworthiness and propensity to trust) may be constant, the specific

consequences of trust will be determined by this situational factor. For example, no physical risk (at least from the robot) is involved when working next to a caged industrial robot. Consequently, trust might be irrelevant in this situation for behavior. In contrast, when interacting in close proximity with a collaborative industrial robot (cobot), there is a physical risk of collisions [3, 92, 96]. In such cases, the worker's trust in the robot might determine their behavior (being extremely alert when trust is low, or not being vigilant towards the robot's actions when trust is high).

2.5 Risk-Taking In Relationship

Before the actual outcomes manifest, trust together with perceived situational risk determine whether people are taking risks in the interaction. Mayer et al. [83] provide a straightforward calculation for risk-taking based on trust and perceived risk. If trust is higher than the perceived risk, the trustor will engage in risk-taking behavior. Conversely, if trust is lower than the perceived risk, the trustor will not engage in risk-taking behavior. This suggests that the decision to take a risk is binary. In actual behavior, however, this decision can also be more nuanced, e.g. resulting in less frequent monitoring of the robot than the situation would actually suggest. Risk-taking behaviors may take several different forms, including engaging with the robot, revealing secrets, collaborating in team behavior, and being more physically close to the robot [5, 55].

2.6 Outcomes

Outcomes are the model's final stage. HRI research outcomes address overall performance, the human-robot relationship, and the flow and safety of interactions. For example, positive outcomes involve increased cooperation or collaboration [114], performance and mutual benefits [97], while negative outcomes involve betrayal, disappointment, and emotional harm [39]. These outcomes represent the main objectives behind most HRI research questions.

2.7 Trust Evolution

The last important part of the model is the feedback loop where the outcomes inform the perceived trustworthiness of the trustee and where the process starts again. Trustworthiness, trust, and risk-taking are then all adjusted with each loop through the system. This also gives rise to long-term effects and the evolution of the relationship between trustor and trustee.

In sum, this model highlights the fact that trust is just one part of the puzzle. To understand why people do or do not trust robots (and other agents), we need to understand the contextual embedding. Moreover, Mayer et al. (1995) [83] define trust as a dynamic construct that is influenced by characteristics of the robot, the human, and the situation [54, 75, 83]. In addition, the idea of reinforcement via a feedback loop from outcomes to trustworthiness again underlines the dynamic essence of the trust concept.

3 Five Common Confusions In Trust Research

The original trust model [83] works out specific aspects of the development of trust, makes clear statements about hypothesized relationships, and is well known to trust researchers in HRI (and human-human trust researchers, of course). Yet, as trust researchers, we still struggle and often fail to stick with the conceptual precision of trust. In the following, we describe five common confusions we believe are at the core of this struggle. These are also summarized in Table 1 in relation to Mayer et al.'s model [83], possible causes for confusion and our suggestions for improvement, that are discussed in section IV.

3.1 Confusion 1: Trustworthiness Is Not Trust

On a surface level, trustworthiness and trust appear very similar and they are also often highly related [21, 121]. When talking about trust it is common to immediately talk about trustworthiness factors and/or antecedents of trust [54]. But trustworthiness and trust are empirically and conceptually distinct constructs [75, 83]. Yet, the distinction between both constructs is often overlooked in HRI research. Perceived trustworthiness is comprised of the perceived characteristics of the trustee such as ability, benevolence, and integrity [83] or performance, process, and purpose [75]. Approaches to measuring perceived trustworthiness can range from multidimensional evaluations—assessing how reliable, capable, ethical, transparent, or benevolent a human perceives a social robot like Furhat during verbal interaction [89, 91]—to simpler methods, such as user ratings of a robot's reliability in a hand-over task [108]. Actual trustworthiness are the objectively measurable aspects of a trustee. For example, it could be measured how well a robot actually performs. The process of matching perceived trustworthiness with actual trustworthiness is part of the trustworthiness assessment [117]. Trust, in contrast, is the attitude of being vulnerable to that trustee at a specific moment in time for a specific situation. This is in part based on trustworthiness assessments but includes more than just that. Trustworthiness assessments may contain a high degree of uncertainty, may be incomplete or inaccurate. The trust attitude then is the attitude on how comfortable a person is with that uncertainty. Trust further includes how comfortable that person is in general with trusting someone or something else (trust propensity) and is mediated by the perceived risk at the time for a specific situation.

As HRI researchers, we often mistake and falsely refer to trustworthiness measures as trust measures. For example, the Multidimensional Measure of Trust questionnaire (MDMT) [81, 130] is rather assessing dimensions of perceived trustworthiness (reliable, capable, sincere, ethical) than trust. The same is true for the trust in industrial human-robot collaboration questionnaire with the dimensions of robot's motion and pick-up speed, safe cooperation, and robot and gripper reliability [17], Schaefer's Trust Perception Scale [115] (Items are queried starting with the question "What % of the time will this robot be…" indicating that the items are about trustworthiness, not trust), and even one of the current authors has just recently validated a "trust" questionnaire with the dimensions performance, utility, purpose, and transparency [107] (we will rename the scale after writing this paper).

The trend of measuring perceived trustworthiness instead of trust is not unique to HRI but also prevalent in human-automation interaction [67]. So, this issue is not new but a known problem repackaged in a new guise.

3.2 Confusion 2: Defining Trust on the Conceptual Level

Mayer et al. [83] have defined trust as a willingness (to be vulnerable), i.e. an intention, whereas Lee and See [75] have described trust as an attitude. Although attitudes and intentions are both latent constructs related to consequent actions, the definition on either level has different implications. Much speaks in favor of defining trust as an attitude. Attitudes are defined as a psychological tendency that is expressed by evaluating a particular entity or behavioral response with some degree of favor or disfavor [1, 36, 37]. For example, a person could trust a robot vacuum cleaner to keep their home tidy. An intention, on the other hand, results from this attitude [1]. In our example, the person could have the intention to turn on the robot to clean the dining room. Intentions are therefore more specific and related to goals (a clean dining room), resulting actions (turn on the robot), and context (use of robot at home) [13]. Lewin [76] also refers to intentions as a tension that is built up through intention formation and has motivational and action-guiding characteristics. However, as soon as the goal is achieved, the tension and therefore also the intention dissolves (after the dining room is cleaned there is no intention to turn on the robot). The progression from trust attitude to intention formation is expressed in the Lee and See model [75]. Trust refers to the willingness to become vulnerable because of the assumption that another agent is favorably disposed towards oneself (which would be expressed as an attitude, too). Although

trust is always placed in a very specific context with a clear goal commitment, it does not dissolve after a specific situation. Therefore, a conceptualization of trust at the attitude level seems more appropriate than on the intention level. Yet, trust can be very specific. Referring to our example, we might trust the robot to vaccuum the floor, but we might not trust it to evaluate whether the result is sufficient (Is it really clean?). Lee and See [75] have referred to this as the specificity of trust, i.e. "the degree to which trust is associated with a particular component or aspect of the trustee" (p. 56). They further distinguish functional specificity, which addresses the differentiation of functions or modes, and temporal specificity, which describes "changes in trust as a function of the situation or over time" (p. 56). Other authors have referred to the specificity of trust as situational trust [59] or swift trust [60]. Adding to the conceptual fuzziness is the fact, that the trust model not only comprises the trust concept itself, but describes the meta-process of trust formation, that contains attributional processes and involves information processing stages in the context of judging another agent (from perceptions and expectations about an agent's trustworthiness to trust attitude to behavioral outcomes). It is therefore not surprising that there are so many different trust definitions: Depending on the purpose of a study or model, a different aspect of this meta-process takes center stage. Accordingly, researchers (us included) have defined trust as an expectation, an attitude, an intention, or sometimes as a behavior [38, 56, 96, 108].

3.3 Confusion 3: Overlooking Risk And Vulnerability

Risk is difficult to study given the generally risk-averse institutional review boards (IRB) at most universities. But "the need for trust only arises in risky situations" (p. 711, [83]). Whether trust becomes relevant for behavior is dependent on the situational and relational risk and, in consequence, on the vulnerability of the trustor [84]. Risk and vulnerability are crucial for the connection between trust and trust-related behavior to be established at all. Risk can be defined as the product of the probability multiplied by the severity of a negative event (German Institute for Standardization 2010; Yes, our German co-authors are indeed very fond of standardization).

We need more risk in our trust research because it is pivotal for construct validity and the necessity of trust [125]. HRI applications offer an intuitive way of implementing risk through physical proximity to an embodied robot [92]. Specifically, when interacting with mobile and / or industrial robots, this proximity represents a potential threat to the human as the robot could collide with them leading to physical harm [92]. In contrast, HRI online studies often lack this credible risk as there is no interaction between robots and study participants. But other forms of risk can also be realized in online studies, for example, with regard to performance (task engagement can have negative consequences), privacy (an activity may compromise the personal information), finance (situation may lead to monetary loss, often realized with pay-off matrices), and security (evaluation that an activity could be susceptible to criminal interference) [84]. But irrespective of the actual nature of risk, it is at least ethically, not easy to realize in research.

An excellent but rare example of implementing physical risk was the work on over-trust in robots [104, 105]. Researchers went to great lengths to create a realistic (and IRB-approved) study of simulating a fire at a university. In this real, risky situation, students trusted and followed an emergency robot, even if it navigated to nonsensical places. This is a great example of trust-related behavior in a situation characterized by physical risk and vulnerability.

3.4 Confusion 4: Trust Is Not Behavior

Research in HRI has often used subjective trust(worthiness) as a proxy or even substitute to make inferences about people's behavior (without actually measuring this behavior). But trust is not a synonym for behaviors like reliance, compliance and verification behaviors but precedes them [75]. We ourselves have regularly experienced this when observing effects on subjective trust(worthiness) but not on assumed trust-related behavior in HRI [65, 96, 106, 108] and vice versa [132]. Revisiting the initial models [75, 83], trust is fundamentally connected to

the relationship's risk-taking and, consequentially, outcomes. For example, compliance and reliance are behavioral consequences that flow from trust, but they are not trust itself [87]. Moreover, the assumed relationship between trust and behavior is not as straightforward as expected and still needs more empirical validation. A recent review [99] showed a weak relationship between subjective trust and dependent behavior (use, reliance, compliance, verification behaviors). Trust and such assumed behavioral outcomes (i.e. risk-taking in relationship) might deviate based on the perceived risk of the situation, which moderates this relation (see Fig.1). For example, in the service domain, a person might use a delivery robot although they might not trust it because it allows them to do other things in parallel, like cleaning the house. Or they might use a robot vacuum cleaner while they are doing the shopping. In both cases, the costs of robot failures would be relatively low, as it would just mean, the desired time saving is lost as the person would have to do the tasks themselves. So whether people in these scenarios trust robots or not would not necessarily be predictive of their behavior. In both cases, the low perceived risk (moderator) could weaken or even dissolve the relation between trust and risk-taking behavior.

3.5 Confusion 5: Why Trust Might Not Be What We're Really After

Despite common confusions regarding the different aspects of the trust model, there may still be a larger issue with trust — namely, does it deserve the prominence it currently holds? If we frequently investigate trust due to its association with important outcomes, questions arise of why we do not measure those outcomes directly and what additional value trust provides. One of the reasons why trust is measured instead of behavioral outcomes might be that it is relatively easy to implement [124]. Measurements of trust and behavior have different demands regarding the experimental design. Whereas trust can be measured using online studies with hypothetical vignettes and robot depictions, measuring realistic behavior needs actual interaction between a human and a robot. This can be a challenge as it concerns financial and time resources as well as technical skills to program a real robot[44]. But even in lab experiments, it is still much more convenient to measure trust with questionnaires compared to behavior. For example, using video-based data as behavioral measures (e.g., for proximity [85]) again requires technical skills and time that might exceed the resources of some researchers. By using trust as proxy for behavioral outcomes, and instead perhaps seeing trust as a facilitator, we may be ignoring other important outcomes such as collaboration quality and performance. By overemphasizing trust and using it as a one-fits-all concept, a further gain in knowledge could be hindered, as other constructs receive less attention in research.

4 What Can We Do To Improve Things?

The identified confusions (summarized in Table 1) mainly relate to the trust model's concepts of perceived trustworthiness, trust, the contextual factor of risk, risk-taking in relationship, and finally the resulting outcomes in HRI. In the following, we would like to make some suggestions for addressing these issues with greater (conceptual and methodological) clarity in future research.

4.1 Actually Measure Trust

Perceived trustworthiness has been extensively investigated even though it was coined as trust. As a result, we have a range of well-developed questionnaires on trustworthiness [20, 67, 81, 107]. Similarly, some studies have already assessed the propensity to trust robots [131], and we have validated questionnaires for that, too [47]. So, we can evaluate both trustworthiness and propensity to trust, provided we label and assess them correctly.

This leads us to the more challenging aspect: assessing trust. Fortunately, Schoormann et al. [119] developed a trust questionnaire specifically designed to capture the concept, which has already been adapted to the HRI context [40, 42]. To evaluate trust, these questionnaires ask about "the respondent's willingness to let the trustee influence issues important to them without the respondent being able to monitor or control" (p. 81, [22]). This highlights the close connection to trust definitions and clearly differentiates these questionnaires from other

Table 1. Confusions, causes for confusions, and suggestions for future research in relation to the components of Mayer et al.'s trust model [83]

Trust Model Component	Confusion	Cause for Confusion	Suggestions
Perceived Trustworthiness	1. Trustworthiness and trust are used interchangeably. They are different constructs.	Concepts are often highly related, "trust" questionnaires are incorrectly framed.	Critically check whether "trust" questionnaires are about trust before using them; check questionnaires against trust definition by Mayer et al. [34] and Lee & See [1].
			Explicitly state which element of the model is measured and why, relate measures back to the model.
Trust	2. Various trust definitions exist. There is no consensus.	The process model of trust includes concepts on different levels: perception, expectation, attitude, willingness, behavioral intention, behavior.	Explicitly state which element of the model is measured and why, relate measures back to the model.
Perceived Risk	3. Risk and vulnerability are overlooked contextual factors.	Risk is difficult to study and to implement in study designs.	Explore creative cover stories and experimental paradigms to include risk, e.g. framing as emergency situations, use physical proximity, robot speed, trajectories as inherent risks in HRI.
Risk-Taking in Relationship	4. Trust is measured as a substitute for behavior. Trust is not a behavior.	Trust is often used as a substitute for behavior because it is easier to study via questionnaires, more convenient.	Investigate, not assume relationships between model elements.
			Create novel context-specific measures for risk-taking behavior with robots.
Outcomes	5. Trust might not be what we are interested in.	Trust is frequently studied although researchers are interested in behavior and other outcomes.	If interested in risk-taking behavior and outcomes, measure them directly.

(potentially inappropriate) attempts that focus on the robot's performance, reliability or capability, all concepts that are related to perceived trustworthiness, not trust.

In addition to trust questionnaires, other authors have operationalized trust with indirect indices, which are especially useful to infer trust dynamics. For example, [4] have used vocal and non-vocal cues to predict relationship risk-taking in HRI. By using facial and voice features they could predict whether people followed a robot's advice in a card game with an accuracy of 87%. Other examples are the use of psychophysiological measures such as heart rate (e.g. [63, 140]), electrodermal response (e.g. [18]), eye-tracking [102, 127] and neural measures [24, 51, 61], which are also promising endeavors. In Table 2 we have summarized some example measures for each trust model component [83] to provide some practical guidance. Of course, this table is by no means exhaustive but should serve as an inspiration when selecting trust-related measures for HRI research. A much more detailed review can be found in [67]. To prevent further confusion about what measures actually relate to which model-derived constructs, they have provided a reference guide with a list of available trust measures, ranging from self-report to physiological indices. Whereas this guide mainly refers to trust in automation, the suggested measures are easily transferable to HRI.

Furthermore, literature reviews [15, 74] and meta-analyses [100] on trust measurements in HRI have been published recently. However, for selecting appropriate trust measures these should be considered with caution. Understandably these works rely on the labels that the original authors assigned to their constructs, and the same applies to a recently compiled database [112, 113] of available HRI questionnaires. While these resources are highly valuable for providing an overview, it is crucial to ensure that the authors' conceptualization of the trust construct aligns with the theoretical foundation when selecting a questionnaire.

Another criterion that determines the appropriateness of trust measure is the context of HRI and the targeted group. Not all indices might be comparably applicable in social and industrial HRI, nor in scenarios with people of different age or cognitive skills. When trust is assessed in child-robot interaction, providing lengthy questionnaires

Table 2. Example measures for the different components of Mayer et al.'s trust model [83] for HRI research

Trust Model Component	Measure Type	Example Measure
Perceived Trustworthiness	Self-Report	Multi Dimensional Trust Questionnaire, MDTM [81] Trust Questionnaire [119]
Trust Propensity	Self-Report	Propensity to Trust [47, 119] Complacency Potential Rating Scale [122]
Trust	Self-Report	Trust Questionnaire [119] HRI-Specific Trust Questionnaire [40, 42]
Perceived Risk	Self-Report / Physiological	Custom Scales / Single Item [58, 78] Perceived Danger Scale [88] Self Assessment Manikin (Dimension Valence) [12, 58] Heart Rate Indices [133]
Risk-Taking in Relationship	Physiological / Behavioral	Event-Related Potential [23] Vocal & Non-Vocal Cues [4] Heart Rate Indices [63, 133, 139] Task Delegation [136]
Outcomes	Behavioral	Combined Team Performance [138] Cooperation & Collaboration [114]

that are rather abstract might not be the best option. The same is true for other vulnerable groups such as dementia patients. Therefore, studies often use behavioral markers as trust indices (e.g. choosing the same object / label for an object as the robot [34, 45]; choosing to follow one robot over another robot [49]). However, strictly speaking and with regard to trust models, these indices do not necessarily represent trust but risk-taking in relationship (see section 2.5 and 4.3). Nevertheless, we acknowledge the challenges that arise from such constraints by context and sample.

4.2 Measure Each Element Appropriately And Relate Them Back To The Model

As we have outlined, there is often confusion about what is measured and why. To be clear about concepts and why we measure them we should explicitly state which element of the model was targeted and by which measures. This was already called for by [67] but has apparently not received enough attention yet. We are therefore taking up this claim once again (see Table 2). Identifying concepts and related measures in our research and publications clearly will help us understand what part of trust is being investigated (and whether the used measures are appropriate).

4.3 Simulate Risk and Measure Risk-Taking

The relevance of situational risk has been numerously postulated for trust [104, 105]. However, this has rarely led to concrete implementation in study designs which might be due to the lack of feasibility to ethically implement risk. If at all, risk is often realized as payoff matrices, in which participants may lose a monetary reward based on their performance [135]. But whether this really poses a risk to participants is questionable. We should therefore explore more creative cover stories and experimental paradigms. Promising examples include the use of VR to simulate altitude as risk [57, 58], the framing as an emergency situation [104, 105, 109] or the use of real self-driving vehicles [31, 90, 128]. As already mentioned, when studying HRI with embodied robots, other intuitive ways of implementing risk are the physical proximity to the robot as well as speed or trajectory manipulations. Comparing online with in-person studies may further be beneficial [109] to understand whether such risk forms can be successfully implemented online and whether results are robust. Other forms of risk might be easily adapted to online vignette scenarios like privacy and security risks [84].

If risk is appropriately implemented in HRI, there is a great opportunity to further investigate risk behaviors associated with trust. These are often context-specific because trust expresses itself differently in each task and environment. For example, in a performance context trust-associated risk behaviors could include task intervention, delegating to a robot, and complying with a robot's advice [67]. In a social context with a robot, sharing a secret [7] and self-disclosure [134] or giving a hug [8, 9] are possible risk-taking behaviors associated with trust. More work is needed to understand these behavioral consequences of trust in environments including education, healthcare, and the home. Another emerging area for measures of trust and associated risk-taking are multi-modal physiological measurement approaches that include neural, hormonal and even genetic measures [61, 66, 72, 140].

4.4 Investigate, Not Assume Relationships Between Model Elements

Trust is a dynamic process, incorporating multiple elements that contribute to trust formation and behavior in HRI. To further our understanding of trust and its importance to HRI, we should not only measure each element but also their relationships. For example, the relationship between trustworthiness and trust is still not very well understood, neither whether contextual or individual factors might change this relationship. The same accounts for the link between trust and behavior, which is often assumed but rarely tested [99]. This calls for appropriate statistical analyses like mediation / moderation analyses, and structural equation models that account for the complexity and the process character of trust [64].

5 Not Done Yet - Further Blindspots in Trust Research

Our suggestions on how to improve our trust research were based on the identified confusions and are closely related to Mayer et al.'s [83] trust model. There are three other aspects that we should pay attention to when conducting trust research, the first two being also part of the trust model: individual differences (trustor's propensity), and feedback dynamics [83].

The first aspect deals with interindividual differences that modulate the trustor's propensity to trust another agent. Reliable individual differences that have been shown in research on trust in automation focused on traits such as extraversion, neuroticism, and self-esteem [70], and demographics with regard to age and ethnicity [116]. Similar differences have emerged in research on HRI [54]. In some cases, individual differences have explained why some trust repair strategies are effective for some individuals, but not others [40]. These first studies reveal the importance of including individual differences in our trust research in general and trust repair specifically, as it may explain some of our findings [98].

The second aspect is the importance of creating a feedback loop in an experiment to really get at trust dynamics. The feedback is the mechanism that links the outcomes back to trustworthiness assessments. Many models

of trust describe such a loop and this mechanism is responsible for the shift from dispositional trust based on a person's initial belief to experience or history-based trust that is based on the interaction with the agent [59, 86]. Importantly, initial beliefs can change the interaction and therefore the feedback that the trustor receives from an interaction. For example, when people initially do not trust a robot, they might be very attentive to the robot's performance (no risk-taking in relationship), which provides ongoing new information about the robot's trustworthiness. Consequently, we should observe an adaptation of trust over time. In contrast, when trust is initially very high, people might not pay much attention to the proper functioning of the robot and therefore, the trust dynamics might not come into play easily. These examples show that a small difference at the beginning might change the entire trust dynamics by enabling (positive and negative) feedback loops. [79] have described such trust divergences with bifurcation as an outcome of a dynamic system. Moreover, focusing on trust dynamics is also essential when investigating trust repair strategies. First, it has to be known what should be repaired, i.e., what was the trust level before the breakdown. Second, trust should be measured repeatedly also after the breakdown to know how much repair is needed and how long a "repair" takes. Third, such repair strategies should involve information that is related to an update of the perceived trustworthiness of a robot as a prerequisite for trust recovery [41, 43, 69, 109]. Lastly, recent work has focused on creating trust-aware or adaptive calibration robotic systems [19, 25, 30, 53, 95, 137]. These systems incorporate trust modeling and measurement into a closed-loop adaptive system [52]. While these systems measure different aspects of the trust process, these are promising approaches for incorporating trust directly into human-robot interaction using the feedback loop

As becomes clear from these considerations, focusing on the dynamic evolution of trust also implies that we need dynamic trust measures. This might be realized by presenting trust questionnaires repeatedly or by using indirect measures of trust such as psychophysiological measures which could even reveal a higher trust resolution [63, 140]. These two aspects have just recently gained attention in HRI research [2, 29, 80, 140] and therefore have not been subject to substantial discussion (and confusion) so far. As both aspects, individual differences and feedback play a crucial role in the trust model, these still seem to be a blindspot that we should actively address in our future research.

The same is true for the third aspect: dependence, which highlights the importance of creating dependent or interdependent relationships with robots in our studies. The Mayer et al. [83] and Lee and See [75] definitions of trust mention three conditions that need to be met for trust to be relevant and this includes 1) dependence on another agent, 2) in the pursuit of a common goal 3) in a situation with uncertainty and vulnerability [38]. We have already discussed the importance of risk and vulnerability, but the dependence component has not been addressed explicitly so far, neither in our paper nor in most scientific inquiry in HRI. Dependence is introduced when the human needs the robot to accomplish something that the human cannot do themselves, at least not easily or without any extra effort. For example, work that focuses on coordination of turn-taking between robots and humans to reach a shared goal does include high degrees of dependence [123].

Without dependence, investigation is limited to perceived trustworthiness because the human does not have to actually rely on the robot for anything which prevents the human from being vulnerable in relation to the robot. To gain insights into trust in HRI, we therefore need scenarios which incorporate at least a certain amount of dependence between human and robot. Full dependence might make the strongest case for trust becoming a relevant construct, which would imply that the robot would be somehow superior to the human (more informed, stronger, more persistent, ...),. But also weaker dependence scenarios can still create a trust-relevant situation, especially in combination with uncertainty and risk. That means, irrespective of strong or weak dependability, the task at hand should not be able to be easily performed by the human alone (nor by the robot).

6 Conclusion

In the previous pages, we acknowledged our own confusions as well as unfavorable trends in our community and hope that this serves as motivation and call to action to improve the field in the future. We hope that HRI researchers will differentiate between trustworthiness and trust, develop questionnaires for trust as an attitude, incorporate risk, and measure risk-taking behavior. However, we understand that the implementation of these changes is contingent upon the alignment of perspectives, which might be the consequential challenge. We are not pretending that we have solved it all. We are looking for solutions. We therefore conclude with four final thoughts on an outlook for trust in HRI research.

First, we hope that this article encourages the community to re-think trust research approaches. We believe there is a lot of great work out there. Nonetheless, we believe there is much to be gained to challenge the status quo, to change the way we design our studies, to improve our measurement and to ultimately test, break and reform our theories.

Second, we need to figure out a way to coordinate better, to leverage innovations in the field, and to integrate our research findings so we advance our field. This may require sharing data, organizing more trust specific workshops (as many have done before) and use common and standard test-bed approaches.

Third, we should re-articulate our goals for trust research as there might be several objectives. Goals could include to discover whether robots are trusted at all, to foster calibrated trust, to increase trust in robots generally, or to repair trust after robots make errors.

Lastly, we believe we are barely scratching the surface of what is possible in trust research. Beyond the dyadic human-robot interaction, trust impacts how groups and society at large work together [73]. Trust is often referenced in conversation when people express fear of robots that take over jobs, destroy the world or general uncertainty and dread about the technology. We therefore have a responsibility to not only conduct insightful research but to educate the public about our findings and to find solutions.

In conclusion, we are optimistic that the field will progress well. This work presents a small step to improve our work. We hope you will join us in these efforts.

References

[1] Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes* 50, 2 (1991), 179–211.

[2] Gene M Alarcon, August Capiola, and Marc D Pfahler. 2021. The role of human personality on trust in human-robot interaction. In *Trust in human-robot interaction*. Elsevier, 159–178.

[3] Basel Alhaji, Sebastian Büttner, Shushanth Sanjay Kumar, and Michael Prilla. 2025. Trust dynamics in human interaction with an industrial robot. *Behaviour & Information Technology* 44, 2 (2025), 266–288.

[4] Abdullah Alzahrani, Jauwairia Nasir, Ahmad J. Tayeb, Elisabeth André, and Muneeb I. Ahmad. 2025. What do the Face and Voice Reveal? Investigating Trust Dynamics During Human-Robot Interaction. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 400–409. doi:10.1109/HRI61500.2025.10974183

[5] Alexander Mois Aroyo, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. 2018. Trust and social engineering in human robot interaction: Will a robot make you disclose sensitive information, conform to its recommendations or gamble? *IEEE Robotics and Automation Letters* 3, 4 (2018), 3701–3708.

[6] Andrew Atchley, Hannah M Barr, Emily O'Hear, Kristin Weger, Bryan Mesmer, Sampson Gholston, and Nathan Tenhundfeld. 2024. Trust in systems: Identification of 17 unresolved research questions and the highlighting of inconsistencies. *Theoretical Issues in Ergonomics Science* 25, 4 (2024), 391–415.

[7] Cindy L Bethel, Matthew R Stevenson, and Brian Scassellati. 2011. Secret-sharing: Interactions between a child, robot, and adult. In *2011 IEEE International Conference on systems, man, and cybernetics*. IEEE, 2489–2494.

[8] Alexis E Block, Sammy Christen, Roger Gassert, Otmar Hilliges, and Katherine J Kuchenbecker. 2021. The six hug commandments: Design and evaluation of a human-sized hugging robot with visual and haptic perception. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*. 380–388.

[9] Alexis E Block and Katherine J Kuchenbecker. 2019. Softness, warmth, and responsiveness improve robot hugs. *International Journal of Social Robotics* 11, 1 (2019), 49–64.

[10] Matthew L Bolton. 2022. Trust is not a virtue: Why we should not trust trust. *Ergonomics in Design* (2022), 10648046221130171.

[11] Matthew L Bolton, Peter A Hancock, John D Lee, Enid Montague, and X Jessie Yang. 2023. Does Trust Have Value? A Discussion About the Importance of Trust to Human Factors and Engineering. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 67. SAGE Publications Sage CA: Los Angeles, CA, 137–138.
[12] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.
[13] Michael Bratman. 1987. *Intention, Plans, and Practical Reason.* Cambridge, MA: Harvard University Press, Cambridge.
[14] David Cameron, Stevienna de Saille, Emily C Collins, Jonathan M Aitken, Hugo Cheung, Adriel Chua, Ee Jing Loh, and James Law. 2021. The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. *Computers in human behavior* 114 (2021), 106561.
[15] Giulio Campagna and Matthias Rehm. 2025. A Systematic Review of Trust Assessments in Human–Robot Interaction. *ACM Transactions on Human-Robot Interaction* 14, 2 (2025), 1–35.
[16] Christiano Castelfranchi and Rino Falcone. 2010. *Trust theory: A socio-cognitive and computational model.* John Wiley & Sons.
[17] George Charalambous, Sarah Fletcher, and Philip Webb. 2016. The development of a scale to evaluate trust in industrial human-robot collaboration. *International Journal of Social Robotics* 8 (2016), 193–209.
[18] Hardik Chauhan, Youjin Jang, and Inbae Jeong. 2024. Predicting human trust in human-robot collaborations using machine learning and psychophysiological responses. *Advanced Engineering Informatics* 62 (2024), 102720.
[19] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 307–315.
[20] Meia Chita-Tegmark, Theresa Law, Nicholas Rabb, and Matthias Scheutz. 2021. Can you trust your trust measure?. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*. 92–100.
[21] Jason A Colquitt, Brent A Scott, and Jeffery A LePine. 2007. Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *Journal of applied psychology* 92, 4 (2007), 909.
[22] F David Schoorman, Roger C Mayer, and James H Davis. 2016. Empowerment in veterinary clinics: The role of trust in delegation. *Journal of Trust Research* 6, 1 (2016), 76–90.
[23] Ewart De Visser and Raja Parasuraman. 2011. Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making* 5, 2 (2011), 209–231.
[24] Ewart J De Visser, Paul J Beatty, Justin R Estepp, Spencer Kohn, Abdulaziz Abubshait, John R Fedota, and Craig G McDonald. 2018. Learning from the slips of others: Neural correlates of trust in automated agents. *Frontiers in human neuroscience* 12 (2018), 309.
[25] Ewart J de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. 2014. A design methodology for trust cue calibration in cognitive agents. In *International conference on virtual, augmented and mixed reality*. Springer, 251–262.
[26] Ewart J de Visser, Frank Krueger, Patrick McKnight, Steven Scheid, Melissa Smith, Stephanie Chalk, and Raja Parasuraman. 2012. The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 56. Sage Publications Sage CA: Los Angeles, CA, 263–267.
[27] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331.
[28] Ewart J De Visser, Richard Pak, and Tyler H Shaw. 2018. From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics* 61, 10 (2018), 1409–1427.
[29] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics* 12, 2 (2020), 459–478.
[30] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics* 12, 2 (2020), 459–478.
[31] Ewart J de Visser, Elizabeth Phillips, Nathan Tenhundfeld, Bianca Donadio, Christian Barentine, Boyoung Kim, Anna Madison, Anthony Ries, and Chad C Tossell. 2023. Trust in automated parking systems: A mixed methods evaluation. *Transportation Research Part F: Traffic Psychology and Behaviour* 96 (2023), 185–199.
[32] Finnur Dellsén and Maria Baghramian. 2021. Disagreement in science: Introduction to the special issue. *Synthese* 198, Suppl 25 (2021), 6011–6021.
[33] Munjal Desai, Mikhail Medvedev, Marynel Vázquez, Sean McSheehy, Sofia Gadea-Omelchenko, Christian Bruggeman, Aaron Steinfeld, and Holly Yanco. 2012. Effects of changing reliability on trust of robot systems. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 73–80.
[34] Cinzia Di Dio, Federico Manzi, Giulia Peretti, Angelo Cangelosi, Paul L Harris, Davide Massaro, and Antonella Marchetti. 2020. Shall I trust you? From child–robot interaction to trusting relationships. *Frontiers in psychology* 11 (2020), 469.
[35] Jiayuan Dong, Connor Esterwood, Xin Ye, Jennifer J Mitchell, Wonse Jo, Lionel P Robert, Chung Hyuk Park, and Myounghoon Jeon. 2024. Taking a Closer Look: Refining Trust and its Impact in HRI. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 1317–1319.

[36] Rodney B Douglass. 1977. Belief, attitude, intention, and behavior: An introduction to theory and research.
[37] Alice H Eagly and Shelly Chaiken. 1993. *The psychology of attitudes.* Harcourt brace Jovanovich college publishers.
[38] Simone Erchov. 2017. *Reconceptualizing trust: Defining, modeling, and measuring trust.* Ph. D. Dissertation. George Mason University.
[39] Raffaella Esposito, Alessandra Rossi, and Silvia Rossi. 2025. Deception in HRI and its Implications: a Systematic Review. *ACM Transactions on Human-Robot Interaction* 14, 3 (2025), 1–26.
[40] Connor Esterwood and Lionel P Robert. 2022. 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)*. IEEE, 332–341.
[41] Connor Esterwood and Lionel P Robert. 2022. A literature review of trust repair in HRI. In *2022 31st IEEE international conference on robot and human interactive communication (ro-man)*. IEEE, 1641–1646.
[42] Connor Esterwood and Lionel P Robert. 2023. The theory of mind and human–robot trust repair. *Scientific Reports* 13, 1 (2023), 9877.
[43] Connor Esterwood and Lionel P Robert. 2025. Repairing Trust in Robots?: A Meta-analysis of HRI Trust Repair Studies with A No-Repair Condition. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 410–419.
[44] David Feil-Seifer, Kerstin S Haring, Silvia Rossi, Alan R Wagner, and Tom Williams. 2020. Where to next? The impact of COVID-19 on human-robot interaction research. 7 pages.
[45] Teresa Flanagan, Nicholas C Georgiou, Brian Scassellati, and Tamar Kushnir. 2024. School-age children are more skeptical of inaccurate robots than adults. *Cognition* 249 (2024), 105814.
[46] Chelsea R Frazier, J Malcolm McCurry, Kevin Zish, and J Gregory Trafton. 2025. The Relationship Between Perceived Agency and Trust in Robots. *International Journal of Social Robotics* (2025), 1–16.
[47] M Lance Frazier, Paul D Johnson, and Stav Fainshmidt. 2013. Development and validation of a propensity to trust scale. *Journal of Trust Research* 3, 2 (2013), 76–97.
[48] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In *2007 International symposium on collaborative technologies and systems*. Ieee, 106–114.
[49] Denise Y. Geiskkovitch, Raquel Thiessen, James E. Young, and Melanie R. Glenwright. 2019. What? That's Not a Chair!: How Robot Informational Errors Affect Children's Trust Towards Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 48–56. doi:10.1109/HRI.2019.8673024
[50] Ioanna Giorgi, Aniello Minutolo, Francesca Tirotto, Oksana Hagen, Massimo Esposito, Mario Gianni, Marco Palomino, and Giovanni L Masala. 2023. I am robot, your health adviser for older adults: Do you trust my advice? *International Journal of Social Robotics* (2023), 1–20.
[51] Kimberly Goodyear, Raja Parasuraman, Sergey Chernyak, Ewart de Visser, Poornima Madhavan, Gopikrishna Deshpande, and Frank Krueger. 2017. An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents. *Social neuroscience* 12, 5 (2017), 570–581.
[52] Yaohui Guo and X Jessie Yang. 2021. Modeling and predicting trust dynamics in human–robot teaming: A Bayesian inference approach. *International Journal of Social Robotics* 13, 8 (2021), 1899–1909.
[53] Yaohui Guo, X Jessie Yang, and Cong Shi. 2024. TIP: A trust inference and propagation model in multi-human multi-robot teams. *Autonomous Robots* 48, 7 (2024), 20.
[54] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.
[55] Yaniv Hanoch, Francesco Arvizzigno, Daniel Hernandez García, Sue Denham, Tony Belpaeme, and Michaela Gummerum. 2021. The robot made me do it: Human–robot interaction and risk-taking behavior. *Cyberpsychology, Behavior, and Social Networking* 24, 5 (2021), 337–342.
[56] D Harrison McKnight and Norman L Chervany. 2001. Trust and distrust definitions: One bite at a time. In *Trust in cyber-societies: Integrating the human and artificial perspectives*. Springer, 27–54.
[57] Steffen Hoesterey and Linda Onnasch. 2023. The effect of risk on trust attitude and trust behavior in interaction with information and decision automation. *Cognition, Technology & Work* 25, 1 (2023), 15–29.
[58] Steffen Hoesterey and Linda Onnasch. 2024. A new experimental paradigm to manipulate risk in human-automation research. *Human factors* 66, 4 (2024), 1170–1185.
[59] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.
[60] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. 2013. Trust in Automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88. doi:10.1109/MIS.2013.24
[61] Sarah K Hopko and Ranjana K Mehta. 2024. Trust in shared-space collaborative robots: Shedding light on the human brain. *Human Factors* 66, 2 (2024), 490–509.
[62] Y-TC Hung, Alan R Dennis, and Lionel Robert. 2004. Trust in virtual teams: Towards an integrative model of trust formation. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*. IEEE, 11–pp.

[63] Halimahtun M Khalid, Liew Wei Shiung, Parham Nooralishahi, Zeeshan Rasool, Martin G Helander, Loo Chu Kiong, and Chin Ai-vyrn. 2016. Exploring psycho-physiological correlates to trust: Implications for human-robot-human interaction. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 697–701.
[64] Wonjoon Kim, Nayoung Kim, Joseph B Lyons, and Chang S Nam. 2020. Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach. *Applied ergonomics* 85 (2020), 103056.
[65] Kim Klüber and Linda Onnasch. 2025. Beyond the monotonic: Enhancing human-robot interaction through affective communication. *Computers in Human Behavior: Artificial Humans* 3 (2025), 100131.
[66] Stella Klumpe, Kelsey C Mitchell, Emma Cox, Jeffrey S Katz, Lucia Lazarowski, Gopikrishna Deshpande, Jonathan Gratch, Ewart J de Visser, Hasan Ayaz, Xingnan Li, et al. 2025. Social bonding between humans, animals, and robots: Dogs outperform AIBOs, their robotic replicas, as social companions. *PloS one* 20, 6 (2025), e0324312.
[67] Spencer C Kohn, Ewart J De Visser, Eva Wiese, Yi-Ching Lee, and Tyler H Shaw. 2021. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology* 12 (2021), 604977.
[68] Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports* 1, 4 (2020), 297–309.
[69] Esther S Kox, Milou Hennekens, Jason S Metcalfe, and José H Kerstholt. 2025. Trust Violations Due To Error or Choice: The Differential Effects on Trust Repair in Human-Human and Human-Robot Interaction. *ACM Transactions on Human-Robot Interaction* (2025).
[70] Johannes Kraus, David Scholz, and Martin Baumann. 2021. What's driving me? Exploration and validation of a hierarchical personality model for trust in automated driving. *Human factors* 63, 6 (2021), 1076–1105.
[71] Johannes Maria Kraus. 2020. *Psychological processes in the formation and calibration of trust in automation*. Ph. D. Dissertation. Universität Ulm.
[72] Frank Krueger and Andreas Meyer-Lindenberg. 2019. Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends in neurosciences* 42, 2 (2019), 92–101.
[73] Frank Krueger, René Riedl, Jennifer A Bartz, Karen S Cook, David Gefen, Peter A Hancock, Sirkka L Jarvenpaa, Lydia Krabbendam, Mary R Lee, Roger C Mayer, et al. 2025. A call for transdisciplinary trust research in the artificial intelligence era. *Humanities and Social Sciences Communications* 12, 1 (2025), 1–10.
[74] Theresa Law and Matthias Scheutz. 2021. Trust: Recent concepts and evaluations in human-robot interaction. *Trust in human-robot interaction* (2021), 27–57.
[75] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[76] Kurt Lewin. 1935. Intention, Will, and Need. In *A Dynamic Theory of Personality: Selected Papers*, D. Cartwright (Ed.). McGraw-Hill, New York, 95–153. Originally published in German in 1926 as "Vorsatz, Wille und Bedürfnis".
[77] Michael Lewis, Katia Sycara, and Phillip Walker. 2018. The role of trust in human-robot interaction. *Foundations of trusted autonomy* (2018), 135–159.
[78] Mengyao Li, Brittany E Holthausen, Rachel E Stuck, and Bruce N Walker. 2019. No risk no trust: Investigating perceived risk in highly automated driving. In *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications*. 177–185.
[79] Mengyao Li, Sofia I Noejovich, Ernest V Cross, and John D Lee. 2023. Explaining trust divergence: Bifurcations in a dynamic system. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 67. SAGE Publications Sage CA: Los Angeles, CA, 139–144.
[80] Joseph B Lyons, Chang S Nam, Sarah A Jessup, Thy Q Vo, and Kevin T Wynne. 2020. The role of individual differences as predictors of trust in autonomous security robots. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE, 1–5.
[81] Bertram F Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction*. Elsevier, 3–25.
[82] Dietrich Manzey, Juliane Reichenbach, and Linda Onnasch. 2012. Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making* 6, 1 (2012), 57–87.
[83] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.
[84] Peter E McKenna, Muneeb I Ahmad, Tafadzwa Maisva, Birthe Nesset, Katrin Lohan, and Helen Hastie. 2024. A Meta-analysis of Vulnerability and Trust in Human-Robot Interaction. *ACM Transactions on Human-Robot Interaction* (2024).
[85] Ross Mead, Amin Atrash, and Maja J Matarić. 2013. Automated proxemic feature extraction and behavior recognition: Applications in human-robot interaction. *International Journal of Social Robotics* 5 (2013), 367–378.
[86] Stephanie M Merritt and Daniel R Ilgen. 2008. Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors* 50, 2 (2008), 194–210.
[87] Joachim Meyer and John D Lee. 2013. Trust, reliance, and compliance. *The Oxford handbook of cognitive engineering* (2013), 109–124.
[88] Jaclyn Molan, Laura Saad, Eileen Roesler, J Malcolm McCurry, Nathaniel Gyory, and J Gregory Trafton. 2025. The Perceived Danger (PD) Scale: Development and Validation. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 420–428.

[89] Ali Momen, Ewart De Visser, Kyle Wolsten, Katrina Cooley, James Walliser, and Chad C Tossell. 2023. Trusting the moral judgments of a robot: perceived moral competence and Humanlikeness of a GPT-3 enabled AI. (2023).
[90] Ali Momen, Ewart J de Visser, Marlena R Fraune, Anna Madison, Matthew Rueben, Katrina Cooley, and Chad C Tossell. 2023. Group trust dynamics during a risky driving experience in a Tesla Model X. *Frontiers in psychology* 14 (2023), 1129369.
[91] Ali Momen, Chad C. Tossell, Richard E. Niemeyer, James Walliser, Michael Tolston, Gregory Funke, and Ewart J. de Visser. 2025. Perceived Trustworthiness and Moral Competence of a GenAI-Enabled Ethical Robot Advisor. *Interactions*.
[92] Jonathan Mumm and Bilge Mutlu. 2011. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*. 331–338.
[93] Chang S Nam and Joseph B Lyons. 2020. *Trust in human-robot interaction*. Academic Press.
[94] Manisha Natarajan and Matthew Gombolay. 2020. Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*. 33–42.
[95] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *Plos one* 15, 2 (2020), e0229132.
[96] Linda Onnasch and Clara Laudine Hildebrandt. 2021. Impact of anthropomorphic robot design on trust and attention in industrial human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 1 (2021), 1–24.
[97] Linda Onnasch, Paul Schweidler, and Maximilian Wieser. 2023. Effects of predictive robot eyes on trust and task performance in an industrial cooperation task. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 442–446.
[98] Richard Pak and Ericka Rovira. 2024. A theoretical model to explain mixed effects of trust repair strategies in autonomous systems. *Theoretical Issues in Ergonomics Science* 25, 4 (2024), 453–473.
[99] Colleen E Patton and Christopher D Wickens. 2024. The relationship of trust and dependence. *Ergonomics* (2024), 1–17.
[100] Yosef S. Razin and Karen M. Feigh. 2024. Converging Measures and an Emergent Model: A Meta-Analysis of Human-Machine Trust Questionnaires. *J. Hum.-Robot Interact.* 13, 4, Article 58 (Nov. 2024), 41 pages. doi:10.1145/3677614
[101] Tobias Rieger, Luisa Kugler, Dietrich Manzey, and Eileen Roesler. 2024. The (Im) perfect automation schema: Who is trusted more, automated or human decision support? *Human Factors* 66, 8 (2024), 1995–2007.
[102] Anthony J Ries, Stéphane Aroca-Ouellette, Alessandro Roncone, and Ewart J de Visser. 2025. Gaze-informed Signatures of Trust and Collaboration in Human-Autonomy Teams. *Computers in Human Behavior: Artificial Humans* (2025), 100171.
[103] Lionel P Robert, Alan R Denis, and Yu-Ting Caisy Hung. 2009. Individual swift trust and knowledge-based trust in face-to-face and virtual team members. *Journal of management information systems* 26, 2 (2009), 241–279.
[104] Paul Robinette, Ayanna M Howard, and Alan R Wagner. 2017. Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 425–436.
[105] Paul Robinette, Michael Novitzky, Brittany Duncan, Myounghoon Jeon, Alan Wagner, and Chung Hyuk Park. 2019. Dangerous HRI: testing real-world robots has real-world consequences. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 687–688.
[106] Eileen Roesler, Linda Onnasch, and Julia I Majer. 2020. The effect of anthropomorphism and failure comprehensibility on human-robot trust. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 64. SAGE Publications Sage CA: Los Angeles, CA, 107–111.
[107] Eileen Roesler*, Tobias Rieger*, and Dietrich Manzey. 2022. Trust towards human vs. automated agents: Using a multidimensional trust questionnaire to assess the role of performance, utility, purpose, and transparency. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 66. SAGE Publications Sage CA: Los Angeles, CA, 2047–2051.
[108] Eileen Roesler, Meret Vollmann, Dietrich Manzey, and Linda Onnasch. 2024. The dynamics of human–robot trust attitude and behavior—Exploring the effects of anthropomorphism and type of failure. *Computers in Human Behavior* 150 (2024), 108008.
[109] Kantwon Rogers, Reiden John Allen Webber, Jinhee Chang, Geronimo Gorostiaga Zubizarreta, and Ayanna Howard. 2024. Lie, Repent, Repeat: Exploring Apologies after Repeated Robot Deception. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 602–610.
[110] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. 2020. How social robots influence people's trust in critical situations. In *2020 29th IEEE International conference on robot and human interactive communication (RO-MAN)*. IEEE, 1020–1025.
[111] Alessandra Rossi, Silvia Rossi, Antonio Andriella, and Anouk van Maris. 2022. The Road to a Successful HRI: AI, Trust and ethicS (TRAITS) Workshop. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1284–1286.
[112] L. Saad, E. Roesler, B. K. Phillips, and J. G. Trafton. 2025. Choosing the "perfect" scale: A primer to evaluate existing scales in HRI. https://hriscaledatabase.psychology.gmu.edu.
[113] Laura Saad, Eileen Roesler, Elizabeth K. Phillips, and J. Gregory Trafton. 2025. Choosing the "perfect" scale: a primer to evaluate existing scales in HRI. *J. Hum.-Robot Interact.* (Oct. 2025). doi:10.1145/3772066 Just Accepted.
[114] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 141–148.

[115] Kristin E Schaefer. 2016. Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". In *Robust intelligence and trust in autonomous systems*. Springer, 191–218.
[116] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* 58, 3 (2016), 377–400.
[117] Nadine Schlicker, Kevin Baum, Alarith Uhde, Sarah Sterz, Martin C Hirsch, and Markus Langer. 2025. How Do We Assess the Trustworthiness of AI? Introducing the Trustworthiness Assessment Model (TrAM). *Computers in Human Behavior* (2025), 108671.
[118] Nadine Schlicker and Markus Langer. 2021. Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In *Proceedings of mensch und computer 2021*. 325–329.
[119] F. D. Schoorman, R. C. Mayer, and J. H. Davis. 1996. Empowerment in veterinary clinics: The role of trust in delegation. In *Presented at the 11th Annual Meeting of the Society for Industrial and Organizational Psychology*. San Diego, CA. Paper presented.
[120] Mariah Schrum, Muyleng Ghuy, Erin Hedlund-Botti, Manisha Natarajan, Michael Johnson, and Matthew Gombolay. 2023. Concerning trends in likert scale usage in human-robot interaction: Towards improving best practices. *ACM Transactions on Human-Robot Interaction* 12, 3 (2023), 1–32.
[121] Harjit Sekhon, Christine Ennew, Husni Kharouf, and James Devlin. 2014. Trustworthiness and trust: influences and implications. *Journal of marketing management* 30, 3-4 (2014), 409–430.
[122] Indramani L Singh, Robert Molloy, and Raja Parasuraman. 1993. Automation-induced" complacency": Development of the complacency-potential rating scale. *The International Journal of Aviation Psychology* 3, 2 (1993), 111–122.
[123] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* 67 (2021), 101178.
[124] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction - HRI '06*. ACM Press. doi:10.1145/1121241.1121249
[125] Rachel E Stuck, Brittany E Holthausen, and Bruce N Walker. 2021. The role of risk in human-robot trust. In *Trust in human-robot interaction*. Elsevier, 179–194.
[126] Nathan Tenhundfeld, Mustafa Demir, and Ewart de Visser. 2022. Assessment of trust in automation in the "real world": Requirements for new trust in automation measurement techniques for use by practitioners. *Journal of Cognitive Engineering and Decision Making* 16, 2 (2022), 101–118.
[127] Nathan L Tenhundfeld, Ewart J De Visser, Kerstin S Haring, Anthony J Ries, Victor S Finomore, and Chad C Tossell. 2019. Calibrating trust in automation through familiarity with the autoparking feature of a Tesla Model X. *Journal of cognitive engineering and decision making* 13, 4 (2019), 279–294.
[128] Nathan L Tenhundfeld, Ewart J de Visser, Anthony J Ries, Victor S Finomore, and Chad C Tossell. 2020. Trust and distrust of automated parking in a Tesla Model X. *Human factors* 62, 2 (2020), 194–210.
[129] Chad C Tossell, Nathan L Tenhundfeld, Ali Momen, Katrina Cooley, and Ewart J De Visser. 2024. Student perceptions of ChatGPT use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence. *IEEE Transactions on Learning Technologies* 17 (2024), 1069–1081.
[130] Daniel Ullman and Bertram F Malle. 2019. Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In *2019 14th ACM/IEEE international Conference on human-robot interaction (HRI)*. IEEE, 618–619.
[131] Irene Valori, Yichen Fan, Merel M Jung, and Merle T Fairhurst. 2024. Propensity to trust shapes perceptions of comforting touch between trustworthy human and robot partners. *Scientific Reports* 14, 1 (2024), 6747.
[132] James C Walliser, Ewart J de Visser, and Tyler H Shaw. 2023. Exploring system wide trust prevalence and mitigation strategies with multiple autonomous agents. *Computers in Human Behavior* 143 (2023), 107671.
[133] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology* 52 (2014), 113–117.
[134] Jacqueline M Kory Westlund, Hae Won Park, Randi Williams, and Cynthia Breazeal. 2018. Measuring young children's long-term relationships with social robots. In *Proceedings of the 17th ACM conference on interaction design and children*. 207–218.
[135] Rebecca Wiczorek and Dietrich Manzey. 2014. Supporting attention allocation in multitask environments: Effects of likelihood alarm systems on trust, behavior, and performance. *Human factors* 56, 7 (2014), 1209–1221.
[136] Yaqi Xie, Indu P Bodala, Desmond C Ong, David Hsu, and Harold Soh. 2019. Robot capability and intention in trust-based decisions across tasks. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 39–47.
[137] Anqi Xu and Gregory Dudek. 2015. Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 221–228.
[138] Sangseok You and Lionel Robert. 2018. Trusting robots in teams: Examining the impacts of trusting robots on team performance and satisfaction. In *You, S. and Robert, LP (2019). Trusting robots in teams: Examining the impacts of trusting robots on team performance and satisfaction, proceedings of the 52th hawaii international conference on system sciences, Jan.* 8–11.

[139] Sangseok You and Lionel P. Robert. 2023. Trusting and Working with Robots: A Relational Demography Theory of Preference for Robotic over Human Co-Workers. *MIS Quarterly* (2023). doi:10.25300/MISQ/2023/17403

[140] Yinsu Zhang, Aakash Yadav, Sarah K Hopko, and Ranjana K Mehta. 2024. In Gaze We Trust: Comparing Eye Tracking, Self-report, and Physiological Indicators of Dynamic Trust during HRI. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 1188–1193.