To Trust or Distrust AI: A Questionnaire Validation Study

Nicolas Scharowski[*] [†]
University of Basel
Basel, Switzerland
nicolas.scharowski@unibas.ch

Sebastian A. C. Perrig[*]
University of Basel
Basel, Switzerland
sebastian.perrig@unibas.ch

Nick von Felten
University of St. Gallen
St. Gallen, Switzerland
nick.vonfelten@unisg.ch

Lena Fanya Aeschbach
University of Basel
Basel, Switzerland
lena.aeschbach@unibas.ch

Klaus Opwis
University of Basel
Basel, Switzerland
klaus.opwis@unibas.ch

Philipp Wintersberger
University of Applied Sciences Hagenberg
Hagensberg, Austria
philipp.wintersberger@fh-hagenberg.at

Florian Brühlmann
University of Basel
Basel, Switzerland
florian.bruehlmann@unibas.ch

Abstract

Despite the importance of trust in human-AI interactions, researchers must often rely on questionnaires adapted from other fields, which lack validation in the AI context. Motivated by the need for reliable and valid measures, we investigated the psychometric quality of the most commonly used trust questionnaire in the context of AI by Jian, Bisantz, and Drury (2000). In a pre-registered online experiment (N = 1485), participants observed interactions with both trustworthy and untrustworthy AI and rated their trust. Our results did not support the originally proposed single-factor structure for the questionnaire, but instead suggested a two-factor solution that distinguishes between trust and distrust. Based on our findings, we provide recommendations for future studies on how to use the questionnaire. Finally, we present arguments for considering trust and distrust as two distinct constructs, emphasizing the opportunities of considering and measuring both in human-AI interactions.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**; **User studies**.

Keywords

AI, XAI, Trust, Distrust, Measurement, Questionnaires, Survey scale, Validation, Psychometrics, human-AI interaction

ACM Reference Format:
Nicolas Scharowski, Sebastian A. C. Perrig, Nick von Felten, Lena Fanya Aeschbach, Klaus Opwis, Philipp Wintersberger, and Florian Brühlmann. 2025. To Trust or Distrust AI: A Questionnaire Validation Study. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3715275.3732025

[*]Both authors contributed equally to this research.
[†]Corresponding Author.

FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732025

1 Introduction

With artificial intelligence (AI) becoming increasingly integrated into people's daily lives, there is a growing need for a comprehensive understanding and appropriate measurement of human trust in AI. Trust is not only an essential element in human-AI interactions as it shapes how people use and rely on AI [29, 48], but also a key motivation for research into explainable AI (XAI) to create more transparent AI systems [54]. Despite this importance, the operationalization and measurement of trust faces various challenges.

For one thing, a multitude of different definitions and conceptualizations of trust exist [3, 67, 94, 95] that are often not clearly distinguished from related terms (e.g., "reliance" [74], "situational trust" [29], "perceived trustworthiness" [97], "calibrated trust" [47] or "warranted trust" [30, 35]). Not clearly distinguishing between these terms can lead to theoretical entanglements and divergent operationalizations of trust [42]. For example, trust, viewed as an attitude [48], is a subjective psychological construct, typically measured via questionnaires, also called survey scales [84]. Meanwhile, reliance, as a behavior [48], can be assessed using more objective observational methods such as analyzing changes in an individual's behavior (e.g., switch ratios [55, 100]). Conceptualizations such as "calibrated" or "warranted" trust also emphasize that the motivation of XAI should not be to merely increase trust arbitrarily and unjustifiably. Instead, trust should be aligned and calibrated to the AI's trustworthiness [48, 99]. In this regard, trust is warranted when the AI is trustworthy and unwarranted when it is untrustworthy [35]. Although the importance of calibrated trust has been recognized by the community [99], the corresponding perspective - that distrust in untrustworthy AI is also warranted – remains relatively underemphasized, despite being an integral factor motivating XAI

[35]. Indeed, distrust seems a comparatively overlooked construct in current human-AI research [94].

Beyond these theoretical challenges, empirical studies measuring trust often use single-items [e.g., 101] or develop their own questionnaires [e.g., 65, 100]. However, self-developed questionnaires and single-items usually lack a rigid construction and quality assurance process and are often only used in an individual study, complicating comparing different study results [25]. Thus, it has been recommended to use validated trust questionnaires [99] whose psychometric quality (i.e., objectivity, reliability, and validity) has been scrutinized. But even if researchers address these challenges and use standardized questionnaires for measuring trust, they have to resort to and adapt scales from other disciplines, as there is no validated questionnaire for trust in AI. For example, it is common practice among researchers to use the *Trust between People and Automation* scale (TPA) by Jian et al. [37] and rephrase the questionnaires' items to fit the study context [95]. However, such practices raise concerns about whether the modified scale still measures what it was initially intended to measure [25, 38]. In fact, most studies measuring trust in AI do not report the psychometric quality of the questionnaires they used [95] and only recently, Lai et al. [45] pointed out that the research community lacks practices to validate and reuse standardized measurements. At best, this makes it challenging for other researchers to replicate or build upon existing work. At worst, using non-validated trust questionnaires in the context of AI can lead to ambiguous or inconsistent findings, impeding progress in human-AI research. Despite this need for standardized measures, the psychometric quality of the TPA remains to be thoroughly investigated in an AI context. Our research aims to fill this research gap by validating the questionnaire in a pre-registered online experiment, following current best practices for investigating scale quality.

The contribution of this paper is threefold. First, we present the first comprehensive psychometric evaluation of Jian et al. [37]'s TPA scale in an AI setting, comparing different theoretical models suggested for the TPA by past work. Second, we offer recommendations and guidance for researchers and practitioners who want to utilize the questionnaire in the context of AI. Third, we emphasize the added value and the opportunities for human-AI research to consider trust and distrust as two individual constructs to be measured independently. Results from the online experiment ($N = 1485$) show that the TPA with its originally proposed single-factor structure performs poorly and that acceptable quality was only achieved when considering a two-factor model differentiating between trust and distrust. Other disciplines have long been in a critical discourse on whether trust and distrust constitute the same construct at opposite ends of a continuum or should be treated as separate constructs on two distinct dimensions. However, this discourse has yet to find any real resonance in the (X)AI community, which could be an underappreciated opportunity for a more inclusive understanding of trust *and* distrust. Such a distinction could account for both warranted trust for trustworthy AI and warranted distrust for untrustworthy AI, which aligns more closely with the objectives of XAI [35]. Ultimately, our work provides future research with more reliable and valid tools for measuring trust in AI and extends the current understanding of trust for a more comprehensive and holistic understanding of human trust *and* distrust in AI.

2 Related work

2.1 Trust in AI

Trust has been studied extensively across various disciplines for decades, including philosophy [24], social sciences [26], and economics [4]. This comprehensive exploration has contributed to a multifaceted perspective on trust and, at times, divergent conceptualizations across different academic domains. Researchers have introduced accounts of interpersonal trust [61] that apply to human-machine interaction [48] and which more recently have been extended to trust in human-AI interaction [35]. There are several definitions [3, 94, 95] and models [e.g., 15, 29, 48, 53, 61, 63, 93] of trust in AI currently in circulation. However, the most commonly used definition in human-AI trust literature [94, 95] is attributed to Lee and See's definition of trust in automation as *"the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability"* [48, p. 6]. Most definitions describe trust either explicitly or implicitly as an attitude [10, 95] and necessitate the presence of risk, uncertainty, and vulnerability for trust to exist [8, 10, 29, 78, 95]. The definition of trust in automation by Lee and See [48] was adopted for trust in AI throughout this work for three reasons: (I) because it is the most widespread definition for trust in AI; (II) because of its emphasis on vulnerability for trust to be a meaningful concept; and (III) because of its broad applicability, as the definition does not specifically require the trusted party to be a robot, chatbot, or automated vehicle (AV).

Trust does not form on its own accord but has its foundation in the attributes, characteristics, or actions of the trustee [29, 61]. Mayer et al. [61], which Lee and See [48]'s definition of trust is based on, referred to these qualities as "factors of trustworthiness" and suggested that "ability," "benevolence," and "integrity" provide the foundation for the development of trust. Lee and See [48] extended the factors contributing to trustworthiness to the context of automation and included performance (i.e., *what* the automation does), process (i.e., *how* the automation works) and purpose (i.e., *why* the automation was developed) as a basis of trust. It is crucial to note the distinction between the *perceived trustworthiness* of the trustor and the *actual trustworthiness* of the trustee [85]. While the actual trustworthiness is a property of the trustee, the perceived trustworthiness is an assessment of these properties on the side of the trustor [61, 85]. More recent research has focused on trustworthiness factors specific to AI systems [40, 53, 92, 93]. For example, Liao and Sundar [53] highlighted the concept of trustworthiness cues. Trustworthiness cues are any information that can contribute to a person's trust assessment [21, 53].

By introducing trustworthiness as a property of the trustee, it is emphasized that trust should not exist for its own sake but requires justification. In light of this, Lee and See [48] have coined the term "trust calibration." Calibration refers to the correspondence between an individual's trust in a system and the system's trustworthiness [48]. Within this framework, two types of mismatches can occur: either an individual's trust exceeds the system's trustworthiness, leading to misuse of the system (i.e., over-reliance [69]), or the individual's trust falls short of the system's trustworthiness, leading to disuse (i.e., under-reliance [69]). Ideally, individuals should exhibit *calibrated trust*, where the level of trust matches the trustworthiness of the system. Wischnewski et al. [99] have encouraged the research community to more explicitly focus on and increase calibrated trust. Further, Jacovi et al. [35] introduced the notion of *warranted* and *unwarranted* trust in the context of AI. They refer to warranted trust as trust calibrated with trustworthiness [35]. Otherwise, if not calibrated with trustworthiness, trust is unwarranted.

2.2 Distrust in AI

The notion of warranted and unwarranted trust brings about an interesting distinction – presuming an AI system is untrustworthy (e.g., has poor performance), not only is trust unwarranted, but conversely distrust is warranted [35]. Jacovi et al. [35] argued that while the key motivation of XAI is commonly framed as increasing trust in AI systems, a more precise motivation should be to either increase trust in trustworthy AI *or* to increase distrust in untrustworthy AI. However, the research community seems to have mainly focused on trust [83], and while this has provided important insights into how trust in AI can be developed and maintained, distrust has been relatively understudied, with only 6% of papers on human-AI interaction measuring and reporting distrust [94].

This unilateral perspective on trust ignores decades of research that has extensively examined the coexistence and independence of trust *and* distrust. According to Lewicki et al. [50] there generally are two conceptualizations of trust:

- **I** The uni-dimensional conceptualization, which treats trust and distrust as bipolar opposites on a single continuum, ranging from distrust to trust [e.g., 37, 77, 86].
- **II** The two-dimensional conceptualization, which views trust and distrust as two distinct dimensions that can vary independently, each ranging from low to high [e.g., 50, 56, 63, 68, 79, 89].

The underlying question that these two conceptualizations raise is whether trust and distrust are independent constructs that can coexist simultaneously, or polar opposites. The uni-dimensional perspective suggests that high trust equates to low distrust, and low trust equals high distrust, implying that trust is always inversely related to distrust and vice versa [11]. In contrast, within the two-dimensional perspective, distrust is more than the absence of trust [44] and thus, high trust must not necessarily imply low distrust.

One of the main proponents of the two-dimensional conceptualization, Luhmann [56], argued that distrust is associated with stronger, more negatively charged emotional reactions, while trust tends to be calmer and more composed. Lewicki et al. [49] expanded on this idea and claimed that trust is grounded in positive emotions (e.g., hope, faith, confidence) and distrust in negative emotions (e.g., fear, skepticism, cynicism) with these emotions not only being mere opposites of one another (e.g., hope vs. no hope) but entirely distinct. This emotional distinctiveness suggests that trust and distrust may not occupy different ends of the same continuum but instead are orthogonal to one another [63].

For Luhmann [56], trust and distrust are both important to manage uncertainty and complexity, but in different ways. Trust reduces complexity by encouraging risk-taking (i.e., undesirable outcomes are removed from consideration to form positive expectations [2]), while distrust reduces complexity by prompting protective action to mitigate risk (i.e., undesirable outcomes are accentuated in consideration to form negative expectations [2]).

There are thus theoretical reasons for a potential distinction between trust and distrust as independent constructs rather than polar opposites. Such a distinction between trust and distrust could have profound implications that could inform future human-AI research, provided that these two constructs could be measured appropriately and separately.

2.3 Measuring trust and distrust in AI

Trust in AI is measured in various ways [31, 95] by both objective and subjective means [66]. Defining trust as an attitude implies that it should be viewed as a subjective psychological construct distinct from objective behavioral manifestations of trust, such as reliance [48, 84]. Therefore, studies that only measure trust-related behavior, such as reliance, do not genuinely measure trust.

Conceptualizing trust as subjective leads to multiple methods to measure trust, including interviews, think-aloud protocols, and open-ended questions [66, 95]. Nevertheless, questionnaires are the primary source for the measurement of subjective trust [94, 95], with Ueno et al. [94] estimating that 89% of publications measure trust in AI via questionnaires. Questionnaires, sometimes referred to as (survey) scales, are a series of questions (i.e., items) designed to measure an unobservable psychological construct of interest [16, 33]. Crucially, there are a multitude of questionnaires that are used to measure trust in AI [e.g., 31, 37, 43, 58, 60, 65, 81], originating from different disciplines such as human-human trust, human-agent trust, human-automation trust, and human-robot trust [95]. This abundance of questionnaires makes it hard for researchers to arrive at an informed decision about which scale to use [95].

Questionnaires should be distinguished from single-item questions and self-developed questionnaires that are also frequently employed to measure trust [42]. However, these often lack a thorough design and validation process [25] and are often employed in a single study only, which usually does not allow comparing results across different studies [22]. For this reason, Wischnewski et al. [99] recommended using validated and standardized trust questionnaires that have undergone scrutiny to ensure their psychometric quality. However, this recommendation poses challenges for researchers who want to measure trust in AI, as no validated trust questionnaire in the context of AI exists.

2.3.1 *The Trust between People and Automation scale.*

Among trust questionnaires, the TPA scale by Jian et al. [37] is by far the most frequently used in human-AI research [29, 42, 94, 95, 99]. The TPA was developed over 20 years ago and has been validated for the context of automation [90]. Researchers adopting the TPA to measure trust in AI thus need to modify the questionnaire items to fit them to the AI context. Vereschak et al. [95] estimated that more than half of all publications introduce such modifications to the original, validated questionnaires (e.g., changing "the system is dependable" to "the artificial intelligence is dependable"). However, terminological differences affect people's perceptions and evaluations of technology [46], and any modification of a questionnaire can undermine its psychometric quality and raise the question of whether an adapted scale measures the intended construct. Consequently, after any modification, the psychometric quality of a questionnaire should be reassessed [25], which is rarely done [95].

The TPA consists of 12 items, with seven positively formulated items for trust and five items being negatively formulated, capturing distrust. However, because of the strong negative correlations between ratings of trust and distrust, the original authors concluded that trust is a single-dimensional construct. Irrespective of this, past research has followed one of two approaches when using the TPA: either re-code the five negatively worded items to form a single trust score, or not re-code items and create two separate scores for trust and distrust [94]. This reflects the existing uncertainty of whether the TPA measures trust as a single construct or trust and distrust as two distinct constructs.

A first effort to validate the TPA was the work by Spain et al. [90], who independently validated the scale in the context of automation. Their results challenged the single-factor structure of the TPA originally proposed, showing that trust and distrust formed two independent factors. More recently, the preliminary work by Perrig et al. [71] further supported the two-factor structure of trust and distrust for the TPA in an AI context. However, their study was not a dedicated validation study and limited to one specific low-risk AI application (i.e., real estate valuation domain). Additionally, the AI system used in their study only exhibited trustworthy behavior. Hence, they could only investigate the psychometric quality of the TPA in a setting where participants interacted with trustworthy AI. The present study aims to expand on their work and seeks to overcome its limitations.

3 Study objectives and hypotheses

Motivated by the need for adequately validated and standardized measures for trust in the context of AI, we set out to validate the TPA in a pre-registered, high-powered online experiment. We formulated the following objective:

Research objective: Conducting a psychometric evaluation of the TPA scale by Jian et al. [37] in the context of AI.[1]

In order to meet this objective, the following methods of psychometric evaluation were used: For the quality of the TPA's individual items, several metrics were considered, namely item descriptive statistics, item difficulty and variance, discriminatory power, and inter-item correlations. Concerning construct validity and investigation of the scale's theoretical model, confirmatory factor analysis (CFA) was used to compare the initially proposed single-factor model for the TPA to an alternative two-factor model. For convergent and divergent validity, we considered correlations with a set of additional measures. Here, we were interested in the relationship of trust and distrust – if support for a two-factor solution to the TPA was found – to the related constructs of positive affect, negative affect, and situational trust, which is similar but distinct from general trust. For reliability, we calculated indicators of internal consistency, namely coefficients α [13] and ω [62]. Concerning scale ratings, taken as indicators of the TPA's criterion validity and manipulation checks (MC1 and MC2) for our stimuli, we formulated the following pre-registered hypotheses:

[1]Note that the pre-registration also contains plans to validate a second trust scale, the Trust Scale for the AI Context (TAI) by Hoffman et al. [31], as part of an overarching research project. To provide a clearer focus, the current manuscript focuses on the validation efforts for the TPA, the more popular of the two scales, although interested readers are referred to the supplementary materials for results on the psychometric quality of the TAI.

H1a: Ratings for the TPA overall score will be significantly higher for the trustworthy condition than the untrustworthy condition.[2]
H1b: Ratings for the TPA trust score will be significantly higher for the trustworthy condition than the untrustworthy condition.
H1c: Ratings for the TPA distrust score will be significantly higher for the untrustworthy condition than the trustworthy condition.
MC1: Ratings of risk will be significantly higher for the automated vehicle scenario compared to the chatbot scenario.
MC2: Ratings of risk will be significantly higher for the untrustworthy condition compared to the trustworthy condition.

4 Methods

A 2x2 Greco-Latin square design online experiment was conducted to validate the TPA. In order to reach the number of participants necessary for a high-impact validation study, we used a scenario-based approach, following prior work on trust [e.g., 5, 32, 36, 39, 81, 82]. Participants were presented with two pre-recorded videos, each accompanied by a brief description of what they were about to see. The experimental manipulation consisted of two independent variables. The first independent variable was the type of AI system presented, with the videos either showing an interaction with an AV or a chatbot (i.e., scenario). The second independent variable was whether the video displayed a trustworthy or an untrustworthy AI (i.e., condition). The order of all four videos was randomized. Thus, all participants were in the trustworthy condition for one scenario while being in the untrustworthy condition for the other scenario, forming a crossover design with four groups (see Figure 1 for a visualization of the stimuli). After each video, participants filled out the TPA, alongside other related survey scales crucial for its psychometric evaluation. The study was approved by the ethics committee of the corresponding author's university and pre-registered on OSF (https://osf.io/3eu4v/).

4.1 Stimuli

Participants were asked to watch two out of the four videos depicting an interaction with AI, one each showing an automated vehicle and a chatbot, displaying either trustworthy or untrustworthy behavior. A brief description of the scenario accompanied these videos. In the trustworthy condition, one video showed an AV without any automation failure, driving safely through an urban environment. The other video featured a chatbot, providing truthful answers to basic knowledge questions (e.g., "a mouse is smaller than an elephant"). In contrast, the videos in the untrustworthy condition showed the following failures: firstly, a staged video of an AV that approaches a crosswalk and seemingly not slowing down for a pedestrian attempting to cross the road [material taken from 32]. Secondly, a chatbot interaction, where the chatbot gives incorrect answers to basic knowledge questions such as "the number 50 is bigger than 5000." Based on the potential consequences of these two AI interactions, we defined the AV scenario as high-risk and the chatbot scenario as low-risk.

[2]Note that in the pre-registration, we referred to the two conditions as "trust" and "distrust." In writing this manuscript, however, we have decided that it is more appropriate to refer to the condition eliciting trust as "trustworthy" and distrust as "untrustworthy", which is more consistent with related work.

Figure 1: An illustration of the 2x2 online experiment stimuli by condition (trustworthy vs. untrustworthy) x scenario (chatbot vs. automated vehicle), constituting four groups in total.

4.2 Participants

We recruited 1500 participants over Prolific, a crowd-sourcing platform recently demonstrated to deliver high data quality [18, 70]. To be eligible for the study, participants had to be current residents of the United States of America (USA) and over 18 years of age. Those who completed the study were compensated £1.50 for their efforts, considered "fair pay" according to the Prolific guidelines [75]. Structural equation modeling, including the CFA approach employed in the present study, is generally considered to be a technique requiring large samples [41]. Thus, using rules of thumb for sample sizes in structural equation modeling, recommending ideally twenty observations per estimated parameter [41], the goal was to recruit at least 700 responses for each group (condition x scenario). Recruiting 1500 participants gave us additional leverage if participants were excluded from data analysis and further allowed us to explore more complex models should they become necessary.

We cleaned the data in line with recommendations by Brühlmann et al. [7], removing participants with incorrect responses to two instructed response items [14] or with negative responses to a self-reported data quality item [64]. After data cleaning, 1485 participants remained, with 2970 complete responses to the measures. Of the participants, 726 were women, 726 were men, and 25 were non-binary people. Two participants preferred to self-describe and six chose not to specify their gender. The mean participant age was 42.98 years ($SD = 13.95, min = 18, max = 82$). Participants were spread evenly across the four groups: 738 responses for the trustworthy chatbot video, 747 for untrustworthy chatbot, 747 for trustworthy AV, and 738 for untrustworthy AV.

4.3 Procedure

On the first page of the survey, participants provided their informed consent. Next, they were given instructions for the task to be completed. Participants were randomly assigned to one of the four videos and asked to watch the video at least once, which was verified by the survey tool. After watching the first video, participants filled out the TPA, followed by the additional measures. Participants were then shown the second video, this time for the other condition and scenario, before responding again to all measures. Finally, participants provided demographic information (age, gender, country of residence) before having the opportunity to give general open feedback and being redirected to Prolific for compensation. After the survey, participants were debriefed that all videos were staged and that no individual was ever in real danger or at risk. Completing the study took participants an average of 11.38 minutes ($SD = 6.07, min = 3.68, max = 49.03$).

4.4 Measures

Participants responded twice to all items of the TPA and additional scales to measure convergent and divergent constructs, once for each group they were assigned to. The TPA was always presented first, followed by the other scales in randomized order. The supplementary materials on OSF contain the exact wording of all items used. The internal consistency for all measures was examined using coefficients α [13] and ω [62], yielding good results for all scales (see subsection 5.4 for results on the TPA, and OSF).

Participants responded to all 12 items of the TPA [37]. Answers were collected on the proposed seven-point Likert-type response scale ranging from 1 ("not at all") to 7 ("extremely"). Responses to the five negatively formulated items of the scale were re-coded prior to data analysis, as theoretically implied by the original authors [37] and in line with prior work [90, 94]. All items were used in their original form, except for replacing the word "system" with the word "AI" (e.g., "I am confident in the AI").

Depending on the scenario (i.e., chatbot or AV), participants either responded to the STS [17] or the STS-AD [32]. The STS-AD is a six-item scale measuring situational trust in an automated driving context. In contrast, the STS is a generalized eight-item version of the STS-AD, assessing situational trust in AI systems in general. Responses to both scales were collected on the same seven-point Likert-type response scale ranging from 1 ("Fully disagree") to 7 ("Fully agree"), and mean values across all items of the respective scale were formed for the analysis. We chose the STS-AD because the scale demonstrated that it measures a "situational trust" factor that is related to but distinct from "general trust" measured with the TPA. The STS was chosen as an alternative to the STS-AD in the chatbot scenario to measure situational trust. We thus expected strong positive correlations of situational trust measured with the STS/STS-AD to trust measured with the TPA, and weaker or negative correlations with distrust.

To measure people's positive and negative affect experienced while seeing the AI interaction, we used the PANAS [96]. The PANAS consists of 20 items, ten each for positive affect and negative affect. Responses were collected on a five-point Likert-type response scale raining from 1 ("Very slightly or not at all") to 5 ("Extremely"), and mean values were formed across positive and negative items respectively to form scores for "positive affect" and "negative affect." The PANAS was chosen because trust and distrust are assumed to cause different emotional responses. According to Luhmann [56], distrust is associated with more negatively charged emotions, while trust is related to more positive emotional reactions. Consequently, we expected positive correlations of positive affect with trust and weaker or negative correlations between positive affect and distrust. For distrust, we expected a mirrored pattern, with strong positive correlations to negative affect and weaker or negative correlations to positive affect.

Finally, we employed a single-item for risk ("How risky did you consider the scenario in the video to be?") to which participants responded on a slider response scale from 0 ("Not at all risky") to 100 ("Extremely risky"). We used this single-item to measure risk because it is a key element [10, 29, 78, 95] and prerequisite for trust to exist [35]. Although we generally advise against single-items, we used one to assess risk, as it served solely as a manipulation check and because there was no appropriate risk questionnaire to use for the contexts under investigation.

5 Results

The analysis focused on different procedures to assess the psychometric quality of the TPA. Results were obtained using the statistical software R [76, version 4.4.1]. The complete analysis can be found in the supplementary materials on OSF.

5.1 Manipulation check

To verify the experimental manipulation, we performed a two-way analysis of variance (ANOVA) for the risk ratings with the factors scenario (AV vs. chatbot) and condition (trustworthy vs. untrustworthy). Results showed that the scenario had a statistically significant effect on the risk rating (manipulation check 1: $F(1, 2967) = 1426$, $p < .001$, $\eta^2 = .22$), with a higher risk rating for the AV ($M = 64.09, SD = 34.49$) compared to the chatbot scenario ($M = 27.22, SD = 34.72$). Concerning condition, there also was a significant difference (manipulation check 2: $F(1, 2967) = 1963$, $p < .001$, $\eta^2 = .31$), with a higher risk rating for the untrustworthy ($M = 67.44, SD = 36.49$) compared to the trustworthy condition ($M = 23.87, SD = 28.17$). We further calculated two Wilcoxon rank sum tests because assumptions for ANOVA were not met (normality, homogeneity of variance). Results were in line with those of the ANOVA, showing a significant difference in risk between the conditions and the scenarios ($p < .001$ for both tests). We thus concluded that the manipulation was successful. Separated by the four groups, mean risk ratings were as follows: 39.27 ($SD = 28.83$) for the trustworthy AV, 89.20 ($SD = 17.26$) for the untrustworthy AV, 8.28 ($SD = 16.52$) for the trustworthy chatbot, and 45.94 ($SD = 37.72$) for the untrustworthy chatbot.

5.2 Item analysis

We started with psychometric analysis of the individual items' quality, calculating descriptive statistics, item difficulty and variance, discriminatory power, and inter-item correlations for the 12 TPA items. Item analysis was performed across the four groups (condition x scenario), as well as for the aggregated overall data. Results indicated no major issues with any of the TPA items and no substantial differences in results between the groups (see OSF for the complete item analysis). Consequently, we decided to work with the overall data across all groups for the subsequent analyses. The complete factor loadings for each item are provided in Appendix A.

5.3 Confirmatory factor analysis

Concerning construct validity, we used CFA to investigate the originally proposed single-factor model of the TPA. In addition, we also tested an alternative two-factor model for the TPA, based on previous work [71, 90]. Based on Hu and Bentler [34], model fit was judged using the following criteria: Low χ^2 value and $p > .05$ for the χ^2 test, $RMSEA < .06$, $SRMR \leq .08$, and $.95 \leq CFI \leq 1$. Multivariate normality of the TPA data was not given, shown by Henze-Zirkler tests [28] and Mardia's tests [59]. Thus, we used a robust maximum likelihood estimator with a Yuan-Bentler scaling correction for all CFAs, which is recommended for non-normal data and reduces the risk of Type I error [6]. The χ^2 test was significant for both CFAs, which was to be expected given that the test is influenced by larger sample sizes (> 200) and departures from multivariate normality [98]. We thus focused on the other indicators to judge model fit.

CFA results showed that the originally proposed single-factor model did not fit the data well, with all indices outside of the desired values [$\chi^2(54) = 2857.47$, $p < .001$, $RMSEA = .157$, $SRMR = .085$, $CFI = .887$]. In contrast, the two-factor version resulted in an improved model fit [$\chi^2(53) = 1256.79$, $p < .001$, $RMSEA = .103$, $SRMR = .051$, $CFI = .952$], with the SRMR and CFI favoring the

Table 1: Descriptive statistics for all collected measures, separate per group (condition x scenario).

	Chatbot trustworthy		Chatbot untrustworthy		AV trustworthy		AV untrustworthy	
Construct	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
TPA trust	4.57	1.27	2.00	1.24	4.10	1.34	2.06	1.18
TPA distrust	2.31	1.25	4.81	1.55	2.90	1.25	4.95	1.26
STS situational trust	5.49	0.90	2.27	1.18	-	-	-	-
STS-AD situational trust	-	-	-	-	4.58	1.20	1.52	0.96
PANAS positive affect	2.65	0.97	2.36	0.87	2.73	0.95	2.41	0.78
PANAS negative affect	1.17	0.42	1.58	0.79	1.44	0.63	2.25	0.98

Note: Responses could range from 1 to 7 for all measures, except for the PANAS, which ranges from 1 to 5.

model and a substantially lower χ^2-value and lower RMSEA compared to the single-factor model. Based on these results, we concluded that the TPA should be used with a two-factor model, distinguishing between trust and distrust. For all subsequent analyses, we thus worked with this version of the TPA.

5.4 Reliability

Following recommendations by Dunn et al. [19], we calculated both coefficients α [13] and ω [62], including 95% confidence intervals, to assess the TPA's reliability. Results showed that the two-factor solution for the TPA was of good to excellent internal consistency [$>$ .80, 27], both for the trust items (α = .94, 95% CI[.94, .95]; ω = .95, 95% CI[.95, .95]) and the distrust items (α = .88, 95% CI[.88, .89]; ω = .89, 95% CI[.88, .89]).

5.5 Convergent and divergent validity

To assess the TPA's convergent and divergent validity, we calculated Pearson's product-moment correlations, reflecting the relationship between the TPA and the related measures. Given the results thus far, we refrained from forming a single trust score for the TPA across all items, as such an approach is not supported by the two-factor model. Rather, we formed two distinct scores for trust and distrust separately, reflecting the two-factor model. In particular, we calculated a "TPA trust" score based on the mean of items 6 to 12 and a "TPA distrust" score based on the mean across the non-reversed values of items 1 through 5. Based on past work [50, 71] and results of the pilot study, we expected the following pattern of correlations among the measured variables: weaker or negative correlations of the mean across the non-reversed TPA distrust items with TPA trust. Positive correlations of situational trust with TPA trust, and weaker or negative correlations with TPA distrust. Positive correlations of positive affect with TPA trust, and weaker or negative correlations with TPA distrust. For negative affect, we expected a mirrored pattern.

All correlations are presented in Appendix A. Results were as anticipated in the pre-registration, supporting the convergent and divergent validity of the TPA. Trust and distrust measured with the TPA correlated negatively, but to a lesser extent than what would be expected from a perfect correlation like those observed by Jian et al. [37]. The situational trust score correlated strongly with the trust measures from the TPA but to a lesser extent than what would be expected of a perfect correlation, suggesting that the ratings of the STS/STS-AD were different from the TPA's trust rating. Distrust, on the other hand, correlated negatively with situational trust. In addition, the pattern of correlations between the PANAS and the TPA showed a moderate positive correlation between positive affect and trust and a weak negative relationship of positive affect to distrust. Negative affect, on the other hand, was positively related to distrust and negatively related to trust. Notably, the magnitudes of these correlations between the PANAS and the TPA trust/distrust scores differed beyond a mere change in positive or negative sign, especially for positive affect, supporting the expected differences between trust, distrust, positive affect, and negative affect.

5.6 Criterion validity

Next, we investigated how the scores of the TPA differed between the four groups, addressing the pre-registered hypotheses. For this, we used two-way ANOVAs to test if the mean ratings for the scale differed significantly depending on the condition (trustworthy vs. untrustworthy) or the scenario (AV vs. chatbot). Given the large sample size of the present study, we used effect sizes (η^2) for a more nuanced interpretation of our findings beyond judging the significance of the statistical tests [91]. We used the following common rule of thumb for interpreting effects: small effect if $\eta^2 = .01$; medium effect if $\eta^2 = .06$; large effect if $\eta^2 \geq .14$ [[12] in 1]. Descriptive statistics separated by the four groups are presented in Table 1. Because we did not calculate an overall trust score for the TPA, as this was not supported by the two-factor model identified to fit the data best, we chose not to calculate results concerning hypothesis H1a.

H1b; higher TPA trust score for the trustworthy condition than untrustworthy condition. A first two-way ANOVA investigating the effect of the condition and scenario on the TPA trust score revealed statistically significant effects for condition ($F(1, 2967) = 2476.37$, $p < .001$, $\eta^2 = .45$) and for scenario ($F(1, 2967) = 18.89$, $p < .001$, $\eta^2 < .01$). Results thus supported H1b with a large effect of the condition on the TPA trust score but no substantial effect for the scenario.

H1c; higher TPA distrust score for the untrustworthy condition than trustworthy condition. Concerning the TPA distrust ratings, a second two-way ANOVA revealed a significant effect for the condition ($F(1, 2967) = 2131.60$, $p < .001$, $\eta^2 = .41$) and for the scenario

($F(1, 2967) = 54.23$, $p < .001$, $\eta^2 = .01$). Results thus favored H1c, suggesting a large effect of the condition on the TPA distrust score and a small effect for the scenario.

Furthermore, we calculated a set of Wilcoxon rank sum tests because the normality and homogeneity of variance assumptions for the ANOVAs were not met. Results were comparable to those of the ANOVAs, with significant effects of both condition and scenario on the TPA trust score and TPA distrust (p from $< .001$ to .030).

6 Discussion

Motivated by the need for standardized and validated scales to measure trust in AI, the present work investigated the psychometric quality of the most commonly used questionnaire to measure trust in AI, the TPA by Jian et al. [37]. In a pre-registered online experiment, 1485 participants observed two videos showing interactions with AI, either displaying trustworthy or untrustworthy behavior.

Results from the CFAs did not support the originally proposed single-factor structure for the TPA, while a substantial improvement in model fit was achieved when considering a two-factor solution. Thus, the TPA's construct validity was supported by the present results, as long as a two-factor structure was employed, distinguishing between a factor for trust (items 6 to 12) and a second factor for distrust (items 1 through 5). This two-factor solution is in line with previous work on the TPA in the context of automation [90] and preliminary work related to AI [71].

As hypothesized and pre-registered, results further indicated that the TPA could differentiate between the two experimental conditions (trustworthy vs. untrustworthy AI). Specifically, in the condition where participants were presented with trustworthy AI, we observed significantly higher trust scores (supporting H1b) and significantly lower distrust scores (supporting H1c) with large effect sizes, compared to the condition with untrustworthy AI. We took these results not only as an indication of the scale's criterion validity but, together with the significantly higher risk ratings in both the AV application area (manipulation check 1) and untrustworthy AI condition (manipulation check 2), as additional evidence of a successful experimental manipulation. Results also showed good to excellent reliability for both the TPA trust and distrust items, as indicated by internal consistency coefficients α and ω.

Regarding convergent and divergent validity, the relationships between the ratings of the TPA and related measures were consistent with our pre-registered expectations. Namely, the TPA's trust score correlated positively with situational trust and positive affect while correlating negatively with negative affect. In contrast, the pattern was reversed for the TPA distrust score, but the correlations also differed in their magnitude. Hence, results demonstrate that distrust and trust are associated with different affects and to a varying extent, as proposed by Lewicki et al. [49]. These results further suggest that trust and distrust may indeed represent two distinct constructs, given that the correlations between the two varied beyond a mere difference in sign (positive vs. negative). Our findings, furthermore, are in line with past research from the context of automation and information technology, proposing that distrust is negatively correlated to trust but not entirely so [57]. The present results thus stand in contrast to the original and almost perfect negative correlations between trust and distrust of $r = -.95$ to $r = -.96$ reported by Jian et al. [37], which led the original authors of the TPA to assume that trust and distrust are opposite ends on a one-dimensional continuum.

In summary, our findings do not suggest using the TPA as a single-factor questionnaire to measure trust in AI but rather support the reliability and validity of a two-factor version of the TPA that can effectively distinguish between trust and distrust. Our findings further indicate that trust and distrust may indeed be distinct constructs with different relations to other measured variables. In the following, we provide recommendations for researchers and practitioners who use the TPA in the context of AI, before elaborating on more general ramifications, emphasizing the opportunities and added value of measuring both trust *and* distrust in the human-AI interaction.

6.1 Recommended use for the TPA

In light of our findings, we strongly recommend that researchers and practitioners refrain from using the single-factor model originally proposed by Jian et al. [37] when applying the TPA in the context of AI. Instead, we suggest using a two-factor structure for the TPA that accounts for both trust *and* distrust. Accordingly, researchers should calculate a composite distrust score by averaging the distrust items (1 to 5) without reversing them, while the remaining items (6 to 12) should be averaged to a composite trust score. Researchers can then analyze and report their data separately for each of the two constructs. We strongly advise against aggregating all items into an overall score or re-coding the negatively formulated items, as our psychometric evaluation does not support such procedures. In addition, we urge researchers to investigate the quality of the TPA before interpreting their data, to ensure that the scale performs as expected in their particular use case.

Following these recommendations allows for the independent measurement of trust and distrust and reduces the uncertainty around TPA usage as highlighted by Ueno et al. [94]. We think that the TPA's ability to differentiate between trust and distrust represents a key strength of the questionnaire, which we will now further outline.

6.2 Implications of a two-dimensional understanding of trust and distrust in AI

The presented results supporting a two-factor solution for the TPA align well with prior research on the scale [71, 90] and theoretical work from interpersonal trust, emphasizing the importance of distrust [e.g., 50, 56, 63, 68, 79, 89]. Beyond the factor analytical findings, our results provide additional support that distrust and trust are associated with different affects, as proposed by Lewicki et al. [49], challenging the unilateral perspective on trust.

Drawing upon their proposed affectual and emotional differentiation between trust and distrust, Lewicki et al. [49] developed a 2x2 framework with trust on one axis and distrust on the other. This framework provides an explanatory approach for the simultaneous coexistence of trust and distrust. Especially for more generally applicable AI systems such as generative AI and large language models, a two-dimensional conceptualization of trust and distrust seems more appropriate, as these systems can perform different

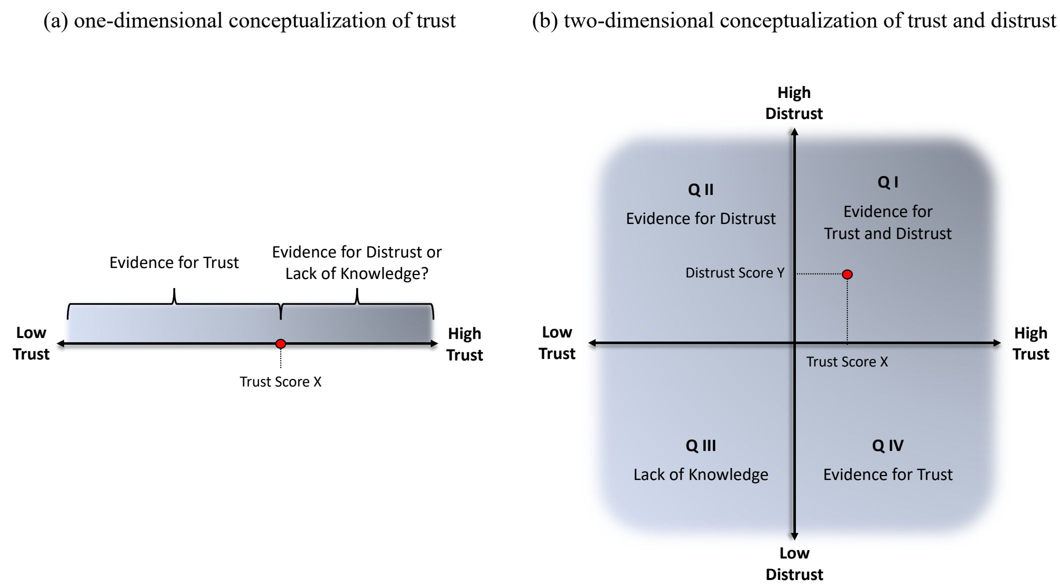

Figure 2: Conceptual frameworks of trust and distrust. (a) the one-dimensional conceptualization places trust on a single continuum ranging from low to high trust (adapted from Castelfranchi and Falcone [10]). (b) the two-dimensional conceptualization of trust and distrust separates trust and distrust scores into two distinct dimensions.

tasks with varying degrees of trustworthiness. Figure 2 shows an adapted version of this two-dimensional framework by Lewicki et al. [49], alongside a one-dimensional conceptualization of trust adapted from Castelfranchi and Falcone [10].

In Quadrant III (low trust, low distrust), trustworthiness judgments are still forming [49]. For example, a person encountering a chatbot for the first time lacks prior experience and knowledge of its capabilities and has thus no basis to either trust or distrust. In Quadrant IV (high trust, low distrust), positive experiences dominate, leading to trust while contradictory evidence is often disregarded [49]. For our example, the person might have frequently observed the chatbot's high capabilities in generating poetry. While trust is warranted [35] and calibrated [48] for this task, the person might over-rely [69] on the chatbot for other untrustworthy tasks, where distrusting the chatbot would be warranted (e.g., providing accurate scientific literature). Quadrant II (low trust, high distrust) arises from predominantly negative experiences, reinforcing distrust that necessitates monitoring [49]. Following our example, the person could be disappointed by the chatbot's inability to provide accurate scientific literature and may double-check responses, resulting in warranted distrust calibrated with untrustworthiness. Conversely, this could lead to under-reliance for trustworthy tasks when trusting the chatbot would be warranted. Finally, in Quadrant I (high trust, high distrust), judgments are more balanced, having experienced both trustworthy and untrustworthy behavior. The person has thus evidence to trust the chatbot for trustworthy tasks (i.e., capability to generate poems) and evidence to distrust the chatbot for untrustworthy tasks (i.e., incapability to provide accurate scientific literature). In this last case, trust and distrust are both *warranted* and *calibrated* with the chatbot's trustworthiness or untrustworthiness.

Figure 2 also highlights the limitations of a one-dimensional trust conceptualization. Aggregating the items for trust and distrust into a single overall score can obscure the underlying reasons for a specific trust score "X". It remains unclear if this score "X" is caused by either genuine distrust or merely from a lack of knowledge regarding the AI's trustworthiness [10]. While a one-dimensional conceptualization of trust cannot meaningfully distinguish between such cases, a two-dimensional conceptualization with distinct scores for trust (X) and distrust (Y) not only addresses this limitation but can also provide additional insights. For example, individuals could be categorized based on their respective levels of trust and distrust. This could allow identifying different user groups within the quadrants I - IV. Some users may exhibit high trust and low distrust, others low trust and high distrust, and yet others might exhibit low levels of both trust and distrust, indicating a lack of knowledge (see Figure 2). Such categorization could deepen our understanding of the specific concerns and needs of these user groups because the factors that contribute to trust differ from those that drive distrust [49]. This has been empirically demonstrated in other areas of human-computer interaction, where varying website characteristics distinctly contributed to trust and distrust [88]. Similarly, within the realm of (X)AI, certain cues may signal trustworthiness to enhance trust (e.g., certification labels), while others could indicate untrustworthiness that fosters distrust (e.g., low accuracy measures). A two-dimensional conceptualization would thus provide practitioners and researchers with a decision basis to either (I) increase the trustworthiness of their AI systems in the case of low levels of trust or (II) decrease the AI's untrustworthiness in the case of high levels of distrust, aligning more closely with the extended goals of XAI as envisioned by Jacovi et al. [35].

7 Limitations and future work

First, the present work utilized crowd-sourcing with exclusively US participants. While crowd-sourced data have been shown to be at least as reliable as other [9], future work using other means of recruitment could explore how the TPA performs in different settings and across varying populations to ensure the cross-cultural applicability of the TPA. Second, we used a scenario-based approach in which participants observed AI interactions. While this is a common approach [e.g., 32, 81] that had the advantage of reaching the necessary number of participants for a high-powered validation study, future work should investigate alternative approaches, using other forms of interaction with AI. Third, the present findings are limited to two application areas: automated vehicles and chatbots. While these are timely and crucial AI application areas, future work should consider additional AI contexts, such as medical diagnosis or content suggestions. Fourth, a general limitation of factor analysis is that the item wording with both positively and negatively formulated items in a questionnaire could potentially influence participant responses [72, 80, 87]. This may lead to distorted factor structures, which have been shown for scales of usability [51, 52, 87] and website aesthetics [72]. Thus, the simultaneous usage of negatively and positively worded items could be an alternative explanation for the two-factor structure of the TPA. However, while in these other cases the authors argued against distinguishing factors based on item wording due to a lack of theoretical grounding, we pointed out that a distinction between trust and distrust is theoretically justified and has merits that go beyond positive or negative item formulation [73, 83]. Ultimately, the underlying structure of psychological constructs, such as trust, is not rooted in statistical but rather in theoretical considerations [23]. Despite this limitation, the psychometric validation of the TPA and our recommendations for its use remain robust. While our work thus contributes to more reliable and valid tools for measuring trust, future research should further explore the dimensionality of trust and distrust. For instance, researchers could investigate how varying levels of trust and distrust differently affect behavioral measures such as reliance and whether high levels of distrust are more predictive of reliance than low levels of trust. Additionally, future research could use different methods to assess convergent and divergent validity of the TPA beyond the approach based on correlations employed in the present work, such as CFA-based techniques [20]. Such approaches could further inform a distinction between trust and distrust, and we encourage future work to investigate potential cues and factors that could influence the two constructs.

8 Conclusion

Trust is a central and frequently measured construct in studying human-AI interactions. However, because no trust in AI questionnaire exists to date, researchers rely on scales developed for other research areas that have not been validated in the AI context. Motivated by the need for validated and standardized questionnaires, the present work reported on the first comprehensive validation of the TPA [37], as the most commonly used trust questionnaire in the context of AI. In a 2x2 online study (N = 1485), using two conditions (trustworthy vs. untrustworthy) and two AI scenarios (AV vs. chatbot), 2970 complete responses to the TPA and related measures were collected. Results from the psychometric evaluation supported the scale's quality regarding reliability and various forms of validity. However, this was not the case for the originally proposed single-factor model of the scale. Consequently, we investigated an alternative two-factor solution based on previous work on the scale and theoretical considerations. A two-factor solution, distinguishing between trust and distrust, clearly enhanced the scale's psychometric quality. From our findings, we derived recommendations for researchers and practitioners who want to use the TPA in the context of AI. We also emphasized the practical and theoretical implications of accounting for both trust and distrust, underscoring the added value of this distinction beyond a theoretical discussion to actual measurement practice. Such a distinction could contribute to a more nuanced and holistic understanding of trust *and* distrust within human-AI interactions in a world where AI increasingly has the potential for both benefit and harm.

9 Data availability statement

The pre-registration (https://osf.io/3eu4v/) and supplementary materials (https://osf.io/7cdne/) for this study are available on OSF.

10 Funding and declaration of conflicting interests

This work is financed entirely by the corresponding author's research group, as we received no additional funding. The authors have no commercial or financial relationships to declare that could be construed as a potential conflict of interest.

11 CRediT author Statement

Nicolas Scharowski: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration **Sebastian A. C. Perrig:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration **Nick von Felten:** Writing - Review & Editing, Formal analysis, Validation **Lena Fanya Aeschbach:** Writing - Review & Editing **Klaus Opwis:** Writing - Review & Editing, Funding acquisition **Philipp Wintersberger:** Conceptualization, Resources, Writing - Review & Editing **Florian Brühlmann:** Conceptualization, Writing - Review & Editing, Supervision

References

[1] Marc A. Adams and Terry L. Conway. 2014. Eta Squared. In *Encyclopedia of Quality of Life and Well-Being Research*, Alex C. Michalos (Ed.). Springer, Dordrecht, Netherlands, 1965–1966. https://doi.org/10.1007/978-94-007-0753-5_918

[2] John Benamati, Mark A. Serva, and Mark A. Fuller. 2006. Are trust and distrust distinct constructs? An empirical study of the effects of trust and distrust among online banking users. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Vol. 6. IEEE, Piscataway, NJ, USA, 121b–121b. https://doi.org/10.1109/HICSS.2006.63

[3] Michaela Benk, Suzanne Tolmeijer, Florian von Wangenheim, and Andrea Ferrario. 2022. The Value of Measuring Trust in AI-A Socio-Technical System Perspective. *arXiv* (2022), 12 pages. arXiv:2204.13480 Retrieved from https://arxiv.org/abs/2204.13480.

[4] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and economic behavior* 10, 1 (1995), 122–142. https://doi.org/10.1006/game.1995.1027

[5] Reuben Binns, Max van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on*

Human Factors in Computing Systems (Montreal QC, Canada) *(CHI '18)*, Regan Mandryk, Mark Hancock, Mark Perry, and Anna Cox (Eds.). Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[6] Timothy A Brown. 2015. *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press, New York, NY, USA. 462 pages.

[7] Florian Brühlmann, Serge Petralito, Lena Aeschbach, and Klaus Opwis. 2020. The Quality of Data Collected Online: An Investigation of Careless Responding in a Crowdsourced Sample. *Methods in Psychology* 2 (2020), 100022. https://doi.org/10.1016/j.metip.2020.100022

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[9] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5. https://doi.org/10.1177/1745691610393980

[10] Christiano Castelfranchi and Rino Falcone. 2010. *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons, Hoboken, NJ, USA. https://doi.org/10.1002/9780470519851

[11] Yong-Sheng Chang and Shyh-Rong Fang. 2013. Antecedents and distinctions between online trust and distrust: Predicting high-and low-risk internet behaviors. *Journal of Electronic Commerce Research* 14, 2 (2013), 149.

[12] Jacob Cohen and Patricia Cohen. 1983. *Applied multiple regression/correlation analysis for the behavioral sciences* (2 ed.). Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

[13] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (1951), 297–334. https://doi.org/10.1007/BF02310555

[14] Paul G. Curran. 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* 66 (2016), 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

[15] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (1989), 319–340. https://doi.org/10.2307/249008

[16] Robert F. DeVellis. 2017. *Scale Development: Theory and Applications* (4th ed.). Sage Publications, Thousand Oaks, CA, USA.

[17] Lena Dolinek and Philipp Wintersberger. 2022. Towards a Generalized Scale to Measure Situational Trust in AI Systems. In *CHI 2022 TRAIT Workshop on Trust and Reliance in AI-Human Teams* (Virtual/New Orleans, LA, USA). Association for Computing Machinery, New York, NY, USA, 9 pages.

[18] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE* 18, 3 (2023), e0279720. https://doi.org/10.1371/journal.pone.0279720

[19] Thomas J. Dunn, Thom Baguley, and Vivienne Brunsden. 2014. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology* 105, 3 (2014), 399–412. https://doi.org/10.1111/bjop.12046

[20] Michael Eid, Mario Gollwitzer, and Manfred Schmitt. 2017. *Statistik und Forschungsmethoden [Statistics and research methods]* (5th ed.). Beltz, Weinheim; Basel.

[21] Fredrick Ekman, Mikael Johansson, and Jana Sochor. 2017. Creating appropriate trust in automated vehicle systems: A framework for HMI design. *IEEE Transactions on Human-Machine Systems* 48, 1 (2017), 95–101. https://doi.org/10.1109/THMS.2017.2776209

[22] Jessica K. Flake and Eiko I. Fried. 2020. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science* 3, 4 (2020), 456–465. https://doi.org/10.1177/2515245920952393

[23] Eiko I. Fried. 2020. Theories and Models: What They Are, What They Are for, and What They Are About. *Psychological Inquiry* 31, 4 (2020), 336–344. https://doi.org/10.1080/1047840X.2020.1854011

[24] Francis Fukuyama. 1996. *Trust: The social virtues and the creation of prosperity*. Free Press, New York, NY, USA.

[25] Mike Furr. 2011. *Scale Construction and Psychometrics for Social and Personality Psychology*. Sage Publications, London, UK.

[26] Diego Gambetta. 2000. Can We Trust Trust? In *Trust: Making and Breaking Cooperative Relations, electronic edition*, Diego Gambetta (Ed.). Department of Sociology, University of Oxford, Oxford, UK, Chapter 13, 213–237.

[27] Darren George and Paul Mallery. 2019. *IBM SPSS statistics 26 step by step: A simple guide and reference* (16 ed.). Routledge, New York, NY, USA. https://doi.org/10.4324/9780429056765

[28] N. Henze and B. Zirkler. 1990. A Class of Invariant Consistent Tests for Multivariate Normality. *Communications in Statistics-Theory and Methods* 19, 10 (1990), 3595–3617. https://doi.org/10.1080/03610929008830400

[29] Kevin A. Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570

[30] Robert R. Hoffman, John D. Lee, David D. Woods, Nigel Shadbolt, Janet Miller, and Jeffrey M. Bradshaw. 2009. The dynamics of trust in cyberdomains. *IEEE Intelligent Systems* 24, 6 (2009), 5–11. https://doi.org/10.1109/MIS.2009.124

[31] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023), 15 pages. https://doi.org/10.3389/fcomp.2023.1096257

[32] Brittany E. Holthausen, Philipp Wintersberger, Bruce N. Walker, and Andreas Riener. 2020. Situational Trust Scale for Automated Driving (STS-AD): Development and Initial Validation. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Virtual Event, DC, USA) *(AutomotiveUI '20)*. Association for Computing Machinery, New York, NY, USA, 40–47. https://doi.org/10.1145/3409120.3410637

[33] Kenneth D. Hopkins. 1998. *Educational and Psychological Measurement and Evaluation*. Pearson, London, UK.

[34] Li-tze Hu and Peter M. Bentler. 1999. Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus new Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6, 1 (1999), 1–55. https://doi.org/10.1080/10705519909540118

[35] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[36] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 310–323. https://doi.org/10.1145/3531146.3533097

[37] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[38] Elizabeth F. Juniper. 2009. Validated Questionnaires Should not be Modified. *European Respiratory Journal* 34, 5 (2009), 1015–1017. https://doi.org/10.1183/09031936.00110209

[39] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% Right and Safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. https://doi.org/10.1145/3491102.3517533

[40] Alexandra D. Kaplan, Theresa T. Kessler, J. Christopher Brill, and P. A. Hancock. 2023. Trust in artificial intelligence: Meta-analytic findings. *Human factors* 65, 2 (2023), 337–359. https://doi.org/10.1177/00187208211013988

[41] Rex B. Kline. 2016. *Principles and practice of structural equation modeling* (4 ed.). The Guilford Press, New York, NY, USA.

[42] 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. https://doi.org/10.3389/fpsyg.2021.604977

[43] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.). Springer International Publishing, Cham, Switzerland, 13–30. https://doi.org/10.1007/978-3-319-96074-6_2

[44] Frens Kroeger. 2019. Unlocking the treasure trove: How can Luhmann's theory of trust enrich trust research? *Journal of Trust Research* 9, 1 (2019), 110–124. https://doi.org/10.1080/21515581.2018.1552592

[45] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[46] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2022. "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 581, 28 pages. https://doi.org/10.1145/3491102.3517527

[47] Markus Langer, Daniel Oster, Timo Speith, Lena Kästner, Holger Hermanns, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What Do We Want From Explainable Artificial Intelligence (XAI)? A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. *Artificial Intelligence* 296 (2021), 103473. https://doi.org/10.1016/j.artint.2021.103473
[48] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
[49] Roy J. Lewicki, Daniel J. McAllister, and Robert J. Bies. 1998. Trust and Distrust: New Relationships and Realities. *The Academy of Management Review* 23, 3 (1998), 438–458. https://doi.org/10.2307/259288
[50] Roy J. Lewicki, Edward C. Tomlinson, and Nicole Gillespie. 2006. Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management* 32, 6 (2006), 991–1022. https://doi.org/10.1177/0149206306294405
[51] James R. Lewis and Jeff Sauro. 2017. Revisiting the Factor Structure of the System Usability Scale. *Journal of Usability Studies* 12, 4 (2017), 183–192.
[52] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2013. UMUX-LITE: When There's No Time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2099–2102. https://doi.org/10.1145/2470654.2481287
[53] Q. Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1257–1268. https://doi.org/10.1145/3531146.3533182
[54] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* 16, 3 (2018), 31–57. https://doi.org/10.1145/3236386.3241340
[55] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 78, 16 pages. https://doi.org/10.1145/3411764.3445562
[56] Niklas Luhmann. 1979. *Trust and Power.* Wiley, Hoboken, NJ, USA.
[57] Joseph B. Lyons, Charlene K. Stokes, Kevin J. Eschleman, Gene M. Alarcon, and Alex J. Barelka. 2011. Trustworthiness and IT suspicion: An evaluation of the nomological network. *Human Factors* 53, 3 (2011), 219–229. https://doi.org/10.1177/0018720811406726
[58] Maria Madsen and Shirley Gregor. 2000. Measuring Human-Computer Trust. In *11th Australasian conference on information systems*, Vol. 53. Citeseer, Australia, 6–8.
[59] Kantilal Vardichand Mardia. 1970. Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika* 57, 3 (1970), 519–530. https://doi.org/10.1093/biomet/57.3.519
[60] Roger C. Mayer and James H. Davis. 1999. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of applied psychology* 84, 1 (1999), 123. https://doi.org/10.1037/0021-9010.84.1.123
[61] 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. https://doi.org/10.2307/258792
[62] Roderick P. McDonald. 1999. *Test theory: A unified treatment* (1 ed.). Psychology Press, New York, NY, USA. https://doi.org/10.4324/9781410601087
[63] D. Harrison McKnight and Norman L. Chervany. 2001. What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *International journal of electronic commerce* 6, 2 (2001), 35–59. https://doi.org/10.1080/10864415.2001.11044235
[64] Adam W. Meade and S. Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological methods* 17, 3 (2012), 437–455. https://doi.org/10.1037/a0028085
[65] Stephanie M. Merritt. 2011. Affective Processes in Human–Automation Interactions. *Human Factors* 53, 4 (2011), 356–370. https://doi.org/10.1177/0018720811411912
[66] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv: Human-Computer Interaction* (2020), 46 pages. arXiv:1811.11839 Retrieved from https://arxiv.org/abs/1811.11839.
[67] Bonnie M. Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 11 (1994), 1905–1922. https://doi.org/10.1080/00140139408964957
[68] Carol Xiaojuan Ou and Choon Ling Sia. 2009. To Trust or to Distrust, That is the Question: Investigating the Trust-Distrust Paradox. *Commun. ACM* 52, 5 (2009), 135–139. https://doi.org/10.1145/1506409.1506442
[69] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2 (1997), 230–253. https://doi.org/10.1518/001872097778543886
[70] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54 (2022), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3
[71] Sebastian A. C. Perrig, Nicolas Scharowski, and Florian Brühlmann. 2023. Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3585808
[72] Sebastian A. C. Perrig, Nick von Felten, Marimo Honda, Klaus Opwis, and Florian Brühlmann. 2023. Development and Validation of a Positive-Item Version of the Visual Aesthetics of Websites Inventory: The VisAWI-Pos. *International Journal of Human–Computer Interaction* 40, 20 (2023), 1–25. https://doi.org/10.1080/10447318.2023.2258634
[73] Tobias M. Peters and Roel W. Visser. 2023. The Importance of Distrust in AI. In *Explainable Artificial Intelligence*, Luca Longo (Ed.). Springer Nature, Cham, Switzerland, 301–317. https://doi.org/10.1007/978-3-031-44070-0_15
[74] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315
[75] Prolific. 2023. https://researcher-help.prolific.com/hc/en-gb/articles/4407695146002-Prolific-s-payment-principles
[76] R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[77] Julian B. Rotter. 1980. Interpersonal trust, trustworthiness, and gullibility. *American psychologist* 35, 1 (1980), 1–7. https://doi.org/10.1037/0003-066X.35.1.1
[78] Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. 1998. Not So Different After All: A Cross-Discipline View Of Trust. *Academy of Management Review* 23, 3 (1998), 393–404. https://doi.org/10.5465/AMR.1998.926617
[79] Mark NK Saunders, Graham Dietz, and Adrian Thornhill. 2014. Trust and distrust: Polar opposites, or independent but co-existing? *Human Relations* 67, 6 (2014), 639–665. https://doi.org/10.1177/0018726713500831
[80] Jeff Sauro and James R. Lewis. 2011. When Designing Usability Questionnaires, Does It Hurt to Be Positive?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2215–2224. https://doi.org/10.1145/1978942.1979266
[81] 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*, Ranjeev Mittu, Donald Sofge, Alan Wagner, and W.F. Lawless (Eds.). Springer US, Boston, MA, USA, 191–218. https://doi.org/10.1007/978-1-4899-7668-0_10
[82] Nicolas Scharowski, Michaela Benk, Swen J. Kühne, Léane Wettstein, and Florian Brühlmann. 2023. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 248–260. https://doi.org/10.1145/3593013.3593994
[83] Nicolas Scharowski and Sebastian A. C. Perrig. 2023. Distrust in (X)AI – Measurement Artifact or Distinct Construct? *CHI 2023 TRAIT Workshop on Trust and Reliance in AI-Human Teams* (2023), 8 pages. https://doi.org/10.48550/arXiv.2303.16495
[84] Nicolas Scharowski, Sebastian A. C. Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and Reliance in XAI–Distinguishing Between Attitudinal and Behavioral Measures. *CHI TRAIT Workshop* (2022), 6 pages. https://doi.org/10.48550/arXiv.2203.12318
[85] Nadine Schlicker, Alarith Uhde, Kevin Baum, Martin C. Hirsch, and Markus Langer. 2022. A Micro and Macro Perspective on Trustworthiness: Theoretical Underpinnings of the Trustworthiness Assessment Model (TrAM). *PsyArXiv* (2022), 60 pages. https://doi.org/10.31234/osf.io/qhwvx
[86] F. David Schoorman, Roger C. Mayer, and James H. Davis. 2007. An integrative model of organizational trust: Past, present, and future. *Academy of Management review* 32, 2 (2007), 344–354. https://doi.org/10.5465/amr.2007.24348410
[87] Beat Vollenwyder Sebastian A. C. Perrig, Nick von Felten and Klaus Opwis. 2024. Development and Psychometric Validation of a Positively Worded German Version of the System Usability Scale (SUS). *International Journal of Human–Computer Interaction* 0, 0 (2024), 1–21. https://doi.org/10.1080/10447318.2024.2434720
[88] Mirjam Seckler, Silvia Heinz, Seamus Forde, Alexandre N. Tuch, and Klaus Opwis. 2015. Trust and distrust on the web: User experiences and website characteristics. *Computers in human behavior* 45 (2015), 39–50. https://doi.org/10.1016/j.chb.2014.11.064

[89] Sim B. Sitkin and Nancy L. Roth. 1993. Explaining the limited effectiveness of legalistic "remedies" for trust/distrust. *Organization science* 4, 3 (1993), 367–392. https://doi.org/10.1287/orsc.4.3.367

[90] Randall D. Spain, Ernesto A. Bustamante, and James P. Bliss. 2008. Towards an Empirically Developed Scale for System Trust: Take Two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52, 19 (2008), 1335–1339. https://doi.org/10.1177/154193120805201907

[91] Barbara G. Tabachnick and Linda S. Fidell. 2007. *Experimental designs using ANOVA*. Duxbury, Belmont, CA, USA.

[92] Lauren Thornton, Bran Knowles, and Gordon Blair. 2021. Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 64–76. https://doi.org/10.1145/3442188.3445871

[93] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The Relationship Between Trust in AI and Trustworthy Machine Learning Technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 272–283. https://doi.org/10.1145/3351095.3372834

[94] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. 2022. Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 254, 7 pages. https://doi.org/10.1145/3491101.3519772

[95] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (2021), 39 pages. https://doi.org/10.1145/3476068

[96] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

[97] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2021. "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces* 15, 2 (2021), 87–98. https://doi.org/10.1007/s12193-020-00332-0

[98] Tiffany A. Whittaker and Randall E. Schumacker. 2022. *A beginner's guide to structural equation modeling* (5 ed.). Routledge, New York, NY, USA. https://doi.org/10.4324/9781003044017

[99] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 755, 16 pages. https://doi.org/10.1145/3544548.3581197

[100] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding The Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[101] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. In *Proceedings of the 22nd international conference on intelligent user interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3025171.3025219

A Appendix

Table 2: Correlations between the TPA trust and distrust scores and the other measures, including 95% confidence intervals.

	TPA trust	TPA distrust
TPA distrust	-.70 [-.72, -.68]	-
STS/STS-AD situational trust	.85 [.84, 86]	-.78 [-.79, -.77]
PANAS positive affect	.41 [.38, .44]	-.17 [-.20, -.13]
PANAS negative affect	-.32 [-.35, -.29]	.46 [.43, .49]

Note: Mean scores for the STS and STS-AD were combined into one variable. All correlations were significant at $p < .001$.

Table 3: Factor loadings and standard errors (SE) for single-factor and two-factor models.

	Single-Factor Model Trust		Two-Factor Model Trust		Two-Factor Model Distrust	
Item	Factor Loading (Standardized)	SE	Factor Loading (Standardized)	SE	Factor Loading (Standardized)	SE
Item 1	1.000 (0.592)	-	-	-	1.000 (0.743)	-
Item 2	0.759 (0.477)	0.023	-	-	0.862 (0.679)	0.017
Item 3	1.178 (0.655)	0.028	-	-	1.180 (0.822)	0.022
Item 4	1.294 (0.746)	0.034	-	-	1.131 (0.818)	0.030
Item 5	1.229 (0.693)	0.034	-	-	1.129 (0.798)	0.029
Item 6	1.577 (0.952)	0.042	1.000 (0.953)	-	-	-
Item 7	1.216 (0.826)	0.037	0.775 (0.831)	0.011	-	-
Item 8	1.094 (0.740)	0.035	0.697 (0.745)	0.013	-	-
Item 9	1.572 (0.944)	0.041	0.997 (0.945)	0.008	-	-
Item 10	1.605 (0.960)	0.041	1.017 (0.960)	0.007	-	-
Item 11	1.532 (0.946)	0.041	0.971 (0.947)	0.007	-	-
Item 12	0.753 (0.487)	0.033	0.481 (0.491)	0.016	-	-