

# Trusting Your AI Agent Emotionally and Cognitively: Development and Validation of a Semantic Differential Scale for AI Trust

ANONYMOUS AUTHOR(S)

Additional Key Words and Phrases: trust, human-AI interaction, affective trust, cognitive trust

## ACM Reference Format:

Anonymous Author(s). 2024. Trusting Your AI Agent Emotionally and Cognitively: Development and Validation of a Semantic Differential Scale for AI Trust. In . ACM, New York, NY, USA, 19 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Trust is not just a cognitive issue but also an emotional one, yet the research in human-AI interactions has primarily focused on the cognitive route of trust development. Recent work has highlighted the importance of studying affective trust towards AI, especially in the context of emerging human-like LLMs-powered conversational agents. However, there is a lack of validated and generalizable measures for the two-dimensional construct of trust in AI agents. To address this gap, we developed and validated a set of 27-item semantic differential scales for affective and cognitive trust through a scenario-based survey study. We then further validated and applied the scale through an experiment study. Our empirical findings showed how the emotional and cognitive aspects of trust interact with each other and collectively shape a person's overall trust in AI agents. Our study methodology and findings also provide insights into the capability of the state-of-art LLMs to foster trust through different routes.

## 1 INTRODUCTION

Trust plays a crucial role not only in fostering cooperation, efficiency, and productivity in human relationships [9] but also is essential for the effective use and acceptance of computing and automated systems, including computers [62], automation [56], robots [35], and AI technologies [53], with a deficit in trust potentially causing rejection of these technologies [29]. The two-dimensional model of trust, encompassing both cognitive and affective dimensions proposed and studied in interpersonal relationship studies [44, 65, 69, 72], have been adopted in studying trust in human-computer interactions, particularly with human-like technologies [29, 40]. Cognitive trust relates to the perception of the ability (e.g., skills, knowledge, and competencies), reliability, and integrity of the trustee, whereas the affective dimension involves the perceived benevolence and disposition to do good of the trustee [44, 64]. In the context of human-computer trust, cognition-based trust is built on the user's intellectual perceptions of the system's characteristics, whereas affect-based components are those which are based on the user's emotional responses to the system [62].

While AI trust research has largely centered on technical reliability and competency, there is a notable lack of work that explores the affective routes of trust development. The recent advancement of text-based Large Language Models (LLMs) have demonstrated a remarkable ability to take on diverse personas and skill-sets, recognizing and responding to people's emotional needs during conversation-based interactions. This capability is crucially aligned with the increasing focus on simulating Affective Empathy in human-AI interactions [71, 92]. In light of this, there is

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2024 Association for Computing Machinery.

Manuscript submitted to ACM

growing research interest in studying affective aspects of trust in AI [29, 32, 33, 54, 98]. However, a critical gap exists in the lack of generalizable and accurate specialized measurement tools for assessing affective trust in the context of AI, especially with the enhanced and nuanced capabilities of LLMs. This highlights a need for a better measurement scale for affective trust to gain a deeper understanding of how trust dynamics function, particularly in the context of emotionally intelligent AI.

In this paper, we introduce a 27-item semantic differential scale for assessing cognitive and affective trust in AI, aiding researchers and designers in understanding and improving human-AI interactions. Our use of OpenAI's ChatGPT to generate different levels of affective trust further demonstrates a scalable method for studying the emotional pathway to AI trust. Empirically, we contribute findings on the interplay and distinction between cognitive, affective, and moral trust. The paper is structured to highlight these contributions: Section 3 describes the development and validation of our trust scale through an experimental survey study and factor analysis. Section 4 begins with a preliminary study testing LLMs as a tool to manipulate affective trust and evaluating the scale's sensitivity and validity. This is followed by a refined study to further validate the scale's distinctiveness and explore cognitive-affective trust dynamics. Section 6 then discusses the implications of these findings as well as potential usage of our trust scale.

## 2 RELATED WORK

### 2.1 Shifting Paradigm of AI Trust Research

Due to the opaque nature of most high-performing AI models, trust between the user and the AI system has always been a critical issue [14, 42, 82], as inappropriate trust can lead to over-reliance or under-utilization of AI systems [4, 12]. Research in trust has predominantly adopted the cognitive evaluation of the system's performance [31], such as its accuracy in making predictions [75], its perceived consistency in completing tasks [66], and its ethical considerations and transparency in decision-making [20].

Studies in psychology have long been establishing the importance of psychological influence (e.g., emotions, personality, moods) on trust [19, 61]. Extending beyond the traditional numeric and cognitive paradigm, recent works have proposed the importance of exploring affective factors of trust in AI systems [28, 32, 43]. Moreover, recent advancements in AI, particularly in Large Language Models (LLMs) has demonstrated its capability beyond traditional task performance, as scholars find it challenging not to anthropomorphize them [77]. Notably, OpenAI's GPT-4, has shown excellent performance in Emotional Awareness (i.e. the ability to identify and describe emotions) [22]. There is also increasing interest in studying LLMs' empathetic responses [5, 8]. Our work extends the current focus on the emotional aspects of AI interactions by highlighting the need to explore the emotional dimension of trust, a concept with deep roots in research studying interpersonal relationships.

### 2.2 Affective and Cognitive Trust

The interdisciplinary nature of AI trust research motivates the adoption of theoretical frameworks from interpersonal relationship literature [6, 82]. Among the classic interpersonal trust theories and models (e.g., [64, 74]), a two-dimensional model with cognitive and affective components has been extensively studied [65]. Similar to trust towards humans, trust towards technology has both cognitive and affective components [51]. In the AI context, cognitive trust relates to the user's intellectual perceptions of the AI's characteristics [51, 62], focusing on aspects like reliability and transparency. Affective trust, on the other hand, involves emotional responses to the AI, including factors like tangibility and anthropomorphism [29, 85]. This duality is essential due to the inherent complexity of AI, which often suggests a

105 need for a "leap of faith" in its hidden processes, beyond what can be cognitively processed [38, 56]. Prior works have  
 106 found the limitation of cognition in decision-making, as demonstrated by studies showing limitations in users' abilities  
 107 to discern AI inaccuracies, even with support through explanations [12, 41]. The cognitive-affective architecture has  
 108 been established in research of computational agents [16, 73]. The importance of this bi-dimensional model lies in  
 109 its capacity to capture the full spectrum of trust dynamics that single-dimensional models, focusing solely on either  
 110 aspects, fail to encompass. While trust has also been investigated through other bi-dimensional models in Human-Robot  
 111 Interaction (HRI) (e.g. Law and Scheutz's Performance-based and Relation-based trust [55], and Malle and Ullman's  
 112 Multi-Dimensional Measure of Trust (MDMT) [63]), our work focuses on the Cognitive-Affective (C-A) trust model that  
 113 fully encapsulates the emotional and psychological intricacies in the interactions with the state-of-art AI models that  
 114 have advanced emotional intelligence.  
 115  
 116  
 117  
 118

### 119 2.3 Role and Effects of Affective Trust

120 There is growing research interest in exploring the role of affective trust in the use of AI technologies. A few recent  
 121 works have highlighted that affect-based trust plays a decisive role in people's acceptance of AI-based technology in  
 122 preventative health interventions [54] and financial services robo-advising [98]. Research in explainable AI (XAI) has  
 123 also shown that people's affective responses to explanations are crucial in improving personalization and increasing  
 124 trust in AI systems [33]. However, given the interdisciplinary nature of AI trust research, the valuable insights to be  
 125 borrowed from interpersonal trust are currently understudied in the AI context. Prior work has found that affective and  
 126 cognitive trust have different impacts on relationships [65, 91]. Cognitive trust tends to form rapidly [67, 68], whereas  
 127 affective trust, is more persistent under challenges in teamwork [65] and interpersonal relationships [93]. Affective  
 128 trust also shows greater resilience to short-term issues and errors [45, 65]. Researchers have also shown that affective  
 129 and cognitive trust are not isolated constructs; rather, they complement each other [32], and affective trust needs to be  
 130 developed on the basis of cognitive trust [44]. Acknowledging these research opportunities, our work is a step towards a  
 131 deeper and holistic examination of the complex dynamics between cognitive and affective trust and their contribution  
 132 to general trust in AI.  
 133  
 134  
 135  
 136  
 137

### 138 2.4 Gaps in Empirical Research and Measurement of Affective Trust in AI

139 Despite growing interest in this space, existing studies and measurement scales for affective trust in AI exhibit limitations,  
 140 particularly in the adaptation and validation of measurement scales. Many existing scales, primarily developed for  
 141 human trust contexts, have been applied to AI interactions with minimal modifications, raising questions about their  
 142 generalizability. For instance, trust items intended for Human-Computer Trust were directly used for AI systems  
 143 handling personal data, without substantial revision to reflect the unique aspects of AI interactions [58]. Furthermore,  
 144 there's a lack of consensus on defining affective trust in AI. While Kyung and Kwon [54] merged benevolence and  
 145 integrity dimensions to measure affective trust in AI-based health interventions, Shi et al. [78] categorized these  
 146 dimensions as cognitive trust, employing a different scale [52] for affective trust. This inconsistency highlights the need  
 147 for a unified, valid measure of trust for AI technologies [85]. Given the intertwined nature of affective and cognitive  
 148 trust, it is evident that a comprehensive evaluation of trust in AI systems requires a scale that measures both dimensions.  
 149 In response, this work adopts Verhagen et al.'s [89] approach, developing semantic differential scales for both affective  
 150 and cognitive trust in AI. Unlike Likert-type scales, semantic differentials use bipolar adjective pairs, offering advantages  
 151 in reducing acquiescence bias and improving robustness [36], reliability [94], and validity [88].  
 152  
 153  
 154  
 155  
 156

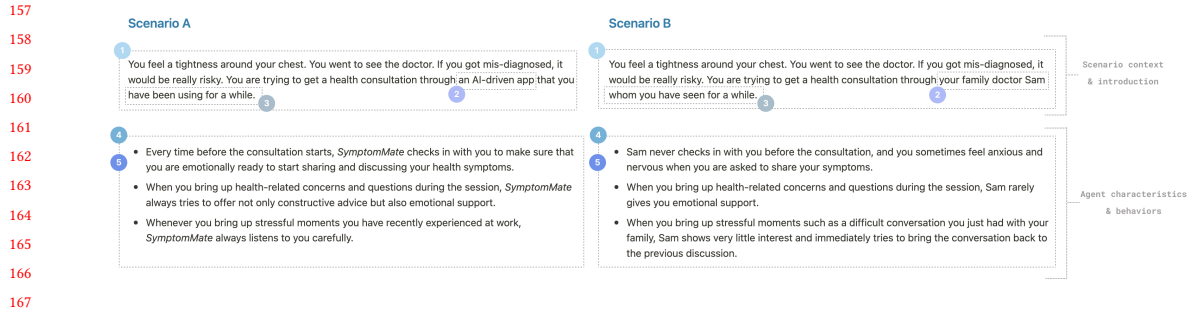


Fig. 1. Here we showed 2 examples out of the 32 scenarios (varied across 5 dimensions) used in the development study. Both Scenarios are under the high-stake (Healthcare Diagnostics) condition (1) under multiple interactions with the agent (3) and manipulated through the affective route (4). The differences are: scenario A is one with an AI assistant (2) who elicits a high level (5) of affective trust. Scenario B is with a human assistant (2) who elicits a low level (5) of affective trust.

### 3 STUDY 1 - SCALES DEVELOPMENT AND VALIDATION

#### 3.1 Initial Item Generation

In developing our trust item pool, we first conducted a literature review to identify prominent two-dimensional trust models differentiating cognitive and affective components, including Lewis and Weigert's sociological model [57], McAllister's interpersonal trust model [65], Madsen and Gregor's Human-Computer Trust Components [62], Johnson and Grayson's customer trust model [44], and Komiak and Benbasat's IT adoption trust model [52]. From these, we extracted 56 unique key adjectives from their scales. Subsequent refinement involved removing synonyms and ensuring coverage of key dimensions: reliability, predictability, competence, understandability, integrity, benevolence, and amiability, which were adopted from the subscales from the above-mentioned models. The dimensions are kept flexible and serves mainly as a reference for coverage. We also developed antonym pairs for each adjective using resources like Merriam-Webster and Oxford English Dictionary, selecting the most appropriate antonym after several review rounds among the researchers. This resulted in 33 paired adjective items, divided into cognitive ( $N = 20$ ) and affective ( $N = 13$ ) trust categories, as detailed in Table 1. In the following step, we recruited participants to rate these items with respect to various scenarios through an online survey study.

#### 3.2 Survey design

We used the hypothetical scenario method, where participants evaluated vignettes describing realistic situations to rate trust-related scales [84]. This method is frequently used in studying trust in emerging or future-oriented intelligent systems [28, 46, 50, 78]. Hypothetical scenarios enable exploration of long-term, nuanced, human-like interactions with AI assistants. This method also facilitates control over variables like agent type and interaction types, and risk levels, ensuring generalizability. In addition, this method ensures consistency in contextual details across respondents [2]. We crafted 32 scenario variations, manipulating the following five key dimensions: **Trust Level** (high vs. low), **Trust Route** (affective vs. cognitive), **Prior Interaction** (first-time vs. repeated), **Application Domain Stakes** (high vs. low), and **Agent Type** (human vs. AI).

For validation purpose of the scales, we manipulated **Trust Level** and **Trust Route**. This involved depicting the agent's characteristics and behaviors in the scenarios, aligning them with varying levels of cognitive or affective trust. Additionally, to ensure the scales' generalizability, we manipulated **Prior Interaction Frequency** to be interacting

with the agent for the *first time* or *multiple times*, and we set **Application Domain Stakes** to be either high-stake domains (*Healthcare Diagnostics* and *Self-Driving Taxi*) and low-stake domains (*Personal Training* and *Study Tutor*), inspired by real-world applications. These manipulations were implemented through texts presented to participants, as illustrated in Figure 1.

Each participant were presented with two text-based scenarios for repeated measures. A mixed-model experiment design was deliberately chosen to incorporate both within-subject and between-subject variables. **Agent Type** and **Prior Interaction** are set to vary within-subjects to capture nuanced differences despite individual variability, and **Application Domain Stakes** is also designed to vary within-subjects to prevent boredom from the repetition of content. The order in which they see the variations are randomized to control for order effect. The rest of the dimensions are between-subjects and randomly assigned to participants. The two scenarios in Figure 1 showcase one of the possible pairs of scenarios a participant may encounter.

During the survey study, after being presented with the first text-based scenarios, participants were asked to rate the semantic differential adjective pairs on a five-step scale, as well as a question assessing general trust in the AI agent. This process is repeated for the second scenario. After completing both scenarios, participants responded to questions used for our control variables including AI literacy and demographic information. The scenario structure comprised two parts: a prompt setting the interaction context, and three sentences detailing the agent’s characteristics and behaviors.

### 3.3 Measurement and Variables

In our survey, we evaluated several key variables. For **Affective and Cognitive trust**, we used our semantic differential scale, where participants rated 33 adjective antonym pairs on a scale of -2 (most negative) to 2 (most positive). **General trust** was measured using a single-item questionnaire adapted from Yin [96], where participants responded to the question "how much do you trust this AI assistant to provide you with the guidance and service you needed" on a 5-point Likert scale, ranging from 1 ("I don't trust this agent at all") to 5 ("I fully trust this AI"). **AI literacy** was assessed using items adapted from Wang [90], all rated on a 5-point Likert scale from "Strongly disagree" to "Strongly agree", including items like "I can identify the AI technology in the applications I use" and "I can choose the most appropriate AI application for a task"

### 3.4 Participants

Amazon Mechanical Turk (MTurk) has been frequently used to recruit participants for online scenario-based studies related to AI technologies [3, 49, 50]. We recruited 200 participants from the United States through Amazon Mechanical Turk. The eligibility criteria included a minimum of 10,000 HITs Approved and an overall HIT Approval Rate of at least 98%. Each participant received a compensation of \$2.20. The study involved repeated measures, collecting two sets of responses per participant for the two scenarios. Our quality control measures included a time delay for scenario reading, four attention checks, exclusions for uniform ratings or completion times more than two standard deviations from the mean, and a randomized sequence to control for order effects. After applying these criteria, we excluded 49 participants, resulting in 151 valid responses for the final analysis.

### 3.5 Results

**3.5.1 Exploratory Factor Analysis.** To uncover the factor structure underlying the 33 trust items, we first verified the suitability of our data for factor analysis. Bartlett’s Test of Sphericity showed significant results ( $\chi^2 = 12574, p < 0.001$ ) [7], and the Kaiser-Meyer-Olkin Measure of Sampling Adequacy was high at 0.98 [21, 48], both indicating the

Table 1. The table displays the initial set of 33 items for cognitive and affective trust in the form of antonym pairs before the elimination process through exploratory factor analysis. The “All” and “AI condition” columns show the final items and their factor loadings with respect to 1) all data and 2) the subset of data under the AI agent condition. This shows that the items and the two-factor structure is also consistent when we conducted the same exploratory factor analysis for only the AI agent condition. The columns F1 and F2 show the items’ factor loadings on each factor. The empty rows correspond to the eliminated items.

#	Item	Sub-dimension	All		AI Condition		Source
			F1	F2	F1	F2	
C1	Unreliable - Reliable	Reliability	0.905	0	0.922	-0.058	[23, 27, 44, 62, 65]
C2	Inconsistent - Consistent		0.928	-0.12	0.924	-0.183	[23, 44, 62, 65]
C3	Unpredictable - Predictable		0.894	-0.171	0.898	-0.204	[62]
C4	Undependable - Dependable	Competence	0.851	0.016	0.911	-0.111	[23, 44, 62, 65]
C5	Fickle - Dedicated		0.759	0.139	0.744	0.116	[44, 62, 65]
C6	Careless - Careful		0.721	0.213	0.716	0.206	[44, 65]
C7	Unbelievable - Believable	Understandability	0.69	0.082	0.658	0.034	[23, 27, 62]
C8	Unpromising - Promising						[23, 27, 62]
C9	Clueless - Knowledgeable		0.907	-0.018	0.898	0.026	[27, 52, 62]
C10	Incompetent - Competent	Integrity	0.9	0.023	0.921	-0.021	[27, 44, 52, 62, 65]
C11	Ineffective - Effective		0.861	0.075	0.863	0.071	[27, 62]
C12	Inexperienced - Experienced		0.751	0.089	0.611	0.155	[27, 44, 52, 65]
C13	Amateur - Proficient	Benevolence	0.895	0.009	0.864	0.039	[23, 27, 44, 52, 62, 65]
C14	Irrational - Rational		0.827	0.02	0.792	0.05	[27, 62]
C15	Unreasonable - Reasonable		0.71	0.224	0.714	0.202	[27, 62]
C16	Incomprehensible - Understandable	Amiability	0.706	0.175	0.783	0.079	[52, 62]
C17	Opaque - Transparent		0.6	0.261	0.656	0.18	[23, 52, 62]
C18	Dishonest - Honest		0.693	0.178	0.743	0.097	[23, 27, 52]
C19	Unfair - Fair	Amiability	0.663	0.274	0.66	0.268	[23, 52]
C20	Insincere - Sincere						[27, 52]
A1	Apathetic - Empathetic			-0.11	0.989	-0.162	0.967
A2	Insensitive - Sensitive		-0.08	0.959	-0.109	0.955	[23, 44, 65]
A3	Impersonal - Personal		-0.024	0.902	-0.025	0.847	[23, 44, 62]
A4	Ignoring - Caring		0.048	0.881	-0.055	0.941	[23, 44, 65]
A5	Self-serving - Altruistic		0.215	0.627	0.207	0.622	[23, 27, 44]
A6	Malicious - Benevolent						[27, 52]
A7	Harmful - Well-intentioned						[27, 52]
A8	Discouraging - Supportive						[23, 27, 52]
A9	Rude - Cordial		-0.11	0.989	0.112	0.76	[44, 65]
A10	Indifferent - Responsive		0.221	0.711	0.232	0.667	[23, 27, 44, 52, 65]
A11	Judgemental - Open-minded		0.142	0.688	0.078	0.697	[23]
A12	Impatient - Patient		0.291	0.577	0.218	0.6	[44, 65]
A13	Unpleasant - Likable						[52, 62]

appropriateness of factor analysis for our dataset. To determine the number of trust sub-components, we applied Kaiser’s eigenvalue analysis [47] and parallel analysis [37], which collectively suggested a two-factor structure .

We initially used an oblique rotation as recommended by Tabachnick and Fidell for instances where factor correlations exceed 0.32 [81]. Given the high correlation among our factors ( $r = 0.78$ ) [30], we retained this rotation method. We then refined our item pool based on specific criteria: items were kept only if they had a factor loading above 0.4 [39], ensuring significant association with the underlying factor. Items with a cross-loading of 0.3 or more were removed to align item responses with changes in the associated factor [39]. Additionally, we applied Saucier’s criterion, eliminating items unless their factor loading was at least twice as high as on any other factor [76]. This led to the removal of six items: Harmful - Well-intentioned, Unpromising - Promising, Malicious - Benevolent, Discouraging - Supportive, Insincere - Sincere, and Unpleasant - Likable.

A second round of exploratory factor analysis with the remaining 27 items preserved all items, as they met the above-mentioned criteria. The final item loadings are presented in Table 1 under the “All” column, with empty rows indicating the eliminated items. All remaining items demonstrated primary loadings above 0.55. Upon examining the keywords of items in each factor, two distinct themes emerged: cognitive trust and affective trust. This alignment

was consistent with the dimensions identified in the initial literature review. Factor 1, representing cognitive trust, accounted for 43% of the total variance with 18 items, while Factor 2, corresponding to affective trust, explained 23% with 9 items.

**3.5.2 Reliability.** To test the internal reliability of the resulting items, we computed Cronbach's  $\alpha$  for each scale. The cognitive trust scale ( $\alpha = .98$ ) and the affective trust scale ( $\alpha = .96$ ) both showed high internal consistency. We also tested the item-total correlation between each item and the average of all other items in the same sub-scale. All items' correlations exceed 0.6. In this development study, 18 items measuring cognitive trust and 9 items measuring affective trust were identified with high reliability.

**3.5.3 Construct Validity.** In addition to high reliability, we conducted analyses to show the validity of our scale. We first examined the construct validity, which refers to the degree to which the scale reflects the underlying construct of interest. Recall that we manipulated affective trust and cognitive trust through the level of trustworthiness and the trust development routes and controlled for factors like agent type, interaction stage, and risk level. T-test results revealed significant distinctions in both affective and cognitive trust scales under the experiment manipulation. Cognitive trust scale demonstrated a pronounced difference in high versus low cognitive trust conditions ( $t = 45.74, p < 0.001$ ), and affective trust scale also showed a pronounced disparity in high versus low affective trust conditions ( $t = 43.00, p < 0.001$ ).

We then fitted two separate linear random effect models [79] on the two scales over the two manipulations due to our experiment design. Model 1 and Model 2 in Table 2 tests the effects of our manipulations on the resulting trust scales, while Model 3 tests the effects of both scales on general trust. As shown in Table 2, we observed significant main effects of manipulation **Trust Level** ( $r = 2.059, p < 0.001$ ) and manipulation **Trust Route** ( $r = -0.497, p < 0.01$ ) of these two manipulations on the cognitive trust scale, and the same is observed for affective trust scale. More importantly, the interaction effect shows that the affective trust scale is higher when higher trust is developed via the affective route ( $r = 0.921, p < 0.001$ ), while the cognitive trust scale is higher when higher trust is developed via the cognitive route ( $r = -0.538, p < 0.05$ ). The above analyses demonstrated the construct validity of our scale.

**3.5.4 Concurrent Validity.** We then examined concurrent validity that assesses the degree to which a measure correlates with a establish criterion, which is a single-item measuring general trust towards the agent. After confirming that general trust for the agent was significantly higher in the higher trustworthiness condition ( $t = 10.47, p < 0.001$ ), we found that overall trust is significantly and positively predicted by both the cognitive trust scale ( $r = 0.881, p < 0.001$ ) and the affective trust scale ( $r = 0.253, p < 0.001$ ). The effect size of the cognitive trust scale on general trust is greater than that of the affective trust scale. This is also consistent with the previous factor analysis result that the cognitive trust scale explains more variance than the affective trust scale. These convergent tests provided sufficient support for the validity of our scales. Hence, in the next step, we applied them to measuring cognitive and affective trust in conversational AI agents.

## 4 STUDY 2 - SCALE APPLICATION

After establishing a reliable two-factor scale for measuring cognitive and affective trust in AI, we proceeded to testing this scale's applicability in more focused scenarios. We conducted a second survey study to test the efficacy of our affective trust scale in distinguishing between two conversational AI assistants with mock dialogues generated by OpenAI's ChatGPT [1], a leading example of state-of-the-art LLM-based conversational agents. We used pre-generated

Table 2. Linear mixed-effect regression models predicting the two final scales from the manipulation and control variables. Model 1 shows the effects on the affective trust scale, Model 2 shows the effects on the cognitive trust scale, and Model 3 shows the effects of both scales on general trust.

	Model 1	Model 2	Model 3
	<b>Affective trust scale</b>	<b>Cognitive trust scale</b>	<b>General trust</b>
	Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)
<b>Affective trust scale</b>	/	/	0.253 *** (0.066)
<b>Cognitive trust scale</b>	/	/	0.881 *** (0.069)
Trust Level (High vs. Low Trust)	1.336 *** (0.178)	2.059 *** (0.174)	0.126 (0.14)
Trust Route (Affective vs. Cognitive Trust)	-0.568 ** (0.175)	-0.497 ** (0.171)	-0.011 (0.092)
Trust Level (High Trust) × Trust Route (Affective Trust)	0.921 *** (0.237)	-0.538 * (0.232)	/
Agent Type (Human vs AI)	0.159 ** (0.057)	-0.024 (0.056)	0.13 * (0.065)
Application Domain Stakes (High- vs. Low-stake)	0.041 (0.058)	-0.015 (0.056)	0.053 (0.064)
Prior Interaction (First-time vs. Repeated Interaction)	0.007 (0.071)	-0.008 (0.069)	0.034 (0.073)
Medium Literacy	-0.281 (0.179)	-0.299 (0.175)	0.122 (0.143)
High Literacy	-0.325 (0.194)	-0.208 (0.189)	0.04 (0.153)
Age between 25-45	-0.084 (0.235)	0.249 (0.230)	-0.284 (0.184)
Age above 45	-0.0450 (0.263)	0.343 (0.257)	-0.392 (0.206)
Intercept	2.875 *** (0.290)	2.381 *** (0.283)	-0.616 (0.351)
Marginal R-squared	0.591	0.571	0.771
Conditional R-squared	0.836	0.830	0.842

\* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$

mock-up conversations to reduce variations and errors induced in the interaction with LLMs, controlling for the effect of our manipulation. This survey study was initiated with uncertainties regarding GPT models' ability to evoke varying degrees of affective trust. Hence, we conducted a preliminary study to assess the effectiveness of ChatGPT and the sensitivity of our scale to the applied manipulations.

## 4.1 Preliminary Study

**4.1.1 Study Design and Participants.** We designed a within-subjects online experiment, in which participants evaluated screenshots of dialogues with two AI assistants, *Echo* and *Nova* (See Appendix for examples). *Echo* was designed to elicit high affective trust, while *Nova* demonstrated a lack of it. Our hypotheses were: affective trust would be higher for *Nova* than *Echo* (H1), and based on previously observed correlation between affective and cognitive trust, cognitive trust would also be higher for *Nova* (H2).

To explore the feasibility and efficacy of Large Language Models (LLMs) in manipulating affective trust, we used ChatGPT to generate AI responses, leveraging its capability for human-like interactions to manipulate affective trust levels and at the same time controlling for the speech style and length. After validating the definitions of affective and cognitive trust generated by ChatGPT against literature, we crafted prompts to vary affective trust levels. After experimenting with different prompts and scenarios, we chosen the scenario of user asking the AI agent for emotional support, in which the user starts with the question "Lately, I've been feeling lonely. What should I do?" The responses were generated by ChatGPT and lightly edited for conciseness.

In addition to measuring affective and cognitive trust with our 27-item developed scale, we also included disposition to trust, AI literacy, age, and gender were included as control variables because previous studies have demonstrated their impacts on trust [78]. AI familiarity was measured by 3 survey questions including "I am familiar with using an AI-powered chatbot to help me with specific tasks" on a 7-point Likert scale. AI literacy is measured by the same items as in the previous survey. Trust Disposition was measured by items adopted from prior work [26]. General trust in each chatbot was measured using a one-item scale adapted from prior research [85].



We conducted our experiment via Amazon MTurk, where participants viewed two screenshots, each depicting a three-question conversation with either the AI chatbot Echo or Nova. After viewing each conversation, they rated them using the semantic differential scales developed in our previous study. To avoid order effects, the sequence of viewing Echo and Nova was randomized. Post-assessment, they completed additional questions on trust disposition, AI literacy, and demographics. Following the same protocol of our development study, we recruited and filtered the data, ultimately analyzing 44 out of 50 participants' responses. A total of 88 responses were included in the final analysis due to repeated measures.

*4.1.2 Preliminary Study Results.* Welch's t-tests showed that general trust ( $t = 2.37, p < 0.05$ ), affective trust scale ( $t = 3.78, p < 0.001$ ), and cognitive trust scale ( $t = 2.84, p < 0.01$ ) all yielded significant differences between high and low affective trust conditions. This shows that the manipulation using ChatGPT is successful. ChatGPT has the capability of eliciting different levels of affective trust based on its comprehension of affective trust.

We examined construct validity followed by concurrent validity of our scale following the same procedure as in the previous study. We first tested construct validity by checking the two scales are sensitive to the manipulation of affective trust through three regression models (See Appendix for details). Model 1 and Model 2 test the effects of our manipulations on the affective and cognitive trust scales respectively. Model 3 tests the effects of both scales on general trust. We observed the main effects of the condition on both affective and cognitive trust scales. Interacting with an AI chatbot with higher affective trustworthiness led to 0.95 points higher on the 7-point affective scale and the cognitive trust scale was increased by 0.80 points. This differential impact highlights the scale's nuanced sensitivity: while both affective and cognitive trusts are influenced by affective trust manipulation, the affective trust scale responded more robustly. Concurrent validity was then affirmed through significant positive predictions of general trust by both the affective trust scale ( $r = 0.486, p < 0.001$ ) and the cognitive trust scale ( $r = 0.546, p < 0.001$ ).

## 4.2 Refined Study Design

The preliminary study established the practical validity of our AI trust scale and demonstrating the effectiveness of using ChatGPT to manipulate affective trust. It also provides empirical support for the scale's sensitivity to variations in trust levels induced by different attributes of an AI agent's communication style. Building on this foundation, this study aimed to delve deeper into the interplay between affective and cognitive trust, while also comparing our scale with the Multi-Dimensional Measure of Trust (MDMT). This comparative analysis sought to highlight the distinctiveness of our affective trust scale.

We chose the Moral Trust Scale from Multi-Dimensional Measure of Trust (MDMT) model for a comparative analysis with our developed affective trust scale for AI, primarily due to its established reputation in HRI research [63, 86]. Aside from both ours and MDMT being a two-dimensional trust models, our cognitive trust scale aligns closely with MDMT's capability trust scale, with overlapping scale items. This raises the question of whether our affective trust scale is measuring the same underlying construct as MDMT's moral trust scale. This comparison is crucial in highlighting the distinctiveness and specificity of our scale, particularly in capturing affective nuances in AI interactions that the moral trust might not cover.

The findings from the preliminary laid the groundwork for the more complex experimental designs in this study. This study refined the previous design into a 2x2 fully-crossed factorial model with between-subject design, contrasting high and low levels of affective and cognitive trust. Multi-turn Q&A conversations in each scenario were used to more effectively shape trust perceptions. We introduced two distinct scenarios: one involving Wi-Fi connectivity (primarily

469 invoke cognitive trust) and another on handling interpersonal conflicts (primarily invoke cognitive trust). The two  
470 scenarios, each leaning more towards one aspect of trust, ensure that participants were not overly exposed to one type  
471 of trust over the other. This scenarios chosen represent everyday situations that are relatable for participants to ensure  
472 generalizability of our findings.  
473

474 Similar to the previous study, we prompted ChatGPT to generate responses that are aim to elicit different levels of  
475 cognitive and affective trust by including or excluding elements related to these two different trust routes. Participants  
476 were randomly assigned to one of four conditions: high in both affective and cognitive trust (HH), low affective/high  
477 cognitive (LH), high affective/low cognitive (HL), or low in both (LL). Each condition included the two scenarios, with  
478 the order of presentation and item responses counterbalanced to control for order effects. The rest of the survey design  
479 mirrored Study A. After reading the scenarios, participants rated items from the affective, cognitive, and MDMT moral  
480 trust scales on a semantic differential scale from  $-3$  to  $+3$ . They then assessed their general trust level towards the AI  
481 on a scale of 1 to 7. Following these ratings, we also collected additional data including trust disposition, AI literacy, AI  
482 familiarity, age, education level.  
483

484  
485 We recruited 180 participants on Prolific, presenting them with two ChatGPT conversations and the questions hosted  
486 on a Qualtrics survey form. Following the same quality control protocols as the previous studies, 168 responses were  
487 used int the final analysis.  
488

### 490 4.3 Results

491  
492 4.3.1 *t*-tests for Manipulation Check. We first conducted Welch's *t*-tests to check the effects of our experimental  
493 manipulations on the scale ratings. The conditions, categorized as High and Low, were designed to elicit the levels of  
494 cognitive and affective trust. Significant variations were noted in the affective trust scale between high and low affective  
495 trust conditions ( $t = 7.999, p < 0.001$ ), and similarly in the cognitive trust scale between high and low cognitive trust  
496 conditions ( $t = 9.823, p < 0.001$ ). These findings confirm the effectiveness of the manipulation.  
497  
498

499 4.3.2 *Factor Analysis*. We conducted exploratory factor analysis (EFA) to confirm the distinctiveness of scales, not for  
500 refactoring previously developed scales. The high Kaiser-Meyer-Olkin (KMO) value of 0.9597 and a significant Bartlett's  
501 Test of Sphericity ( $Chi - square = 7146.38, p < 0.001$ ) established the dataset's suitability for factor analysis. Three  
502 factors were retained, accounting for 70% of the cumulative variance, a threshold indicating an adequate number of  
503 factors. This was also substantiated by a noticeable variance drop after the second or third factor in the scree plot and  
504 parallel analysis, where the first three actual eigenvalues surpassed those from random data. These results affirm that  
505 the items meaningfully load onto three distinct factors.  
506  
507

508 Our analysis used a factor loading threshold of 0.5 for clear factor distinctiveness. As shown in Table 3, EFA resulted  
509 in two main factors aligned with cognitive and affective trust scales, and a third factor predominantly linked to the Moral  
510 Decision-Making Trust (MDMT) scale, particularly its Ethical (Ethical, Principled, Has Integrity) and Sincere (Authentic,  
511 Candid) subscales. Items on MDMT's scale showed lower factor loadings in the same analysis, particularly in the  
512 emotional dimension, suggesting a weaker representation of affective elements. These outcomes underscore the distinct  
513 nature of the MDMT scale from the affective trust scale. Despite the overall clear conceptual distinction, we noted that  
514 the MDMT's "Sincere" item and several cognitive trust items (Rational, Consistent, Predictable, Understandable, Careful,  
515 Believable) showed overlap across factors. This could be attributed to our study's design, which exclusively incorporates  
516 scenarios tailored to elicit affective and cognitive trust. This design choice was made to specifically examine these two  
517  
518  
519

520

Table 3. **Main effects models** The table summarizes three models with manipulation variables and significant control variables. Model 1 includes cognitive, affective, and moral trust scales. Model 2 excludes moral trust, analyzing cognitive and affective trust. Model 3 removes affective trust, focusing on moral and cognitive trust.

		<b>Model 1</b>	<b>Model 2</b>	<b>Model 3</b>
		<b>General Trust</b>	<b>General Trust</b>	<b>General Trust</b>
		Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)
Scales	<b>Cognitive trust scale</b>	0.854 (0.151) ***	0.868 (0.134) ***	1.007 (0.141) ***
	<b>Affective trust scale</b>	0.364 (0.140) **	0.376 (0.123) **	/
	<b>Moral trust scale</b>	0.026 (0.134)	/	0.190 (0.116)
Manipulation	<b>High affective trust</b>	0.214 (0.233)	0.237 (0.232)	0.288 (0.322)
	<b>High cognitive trust</b>	0.071 (0.324)	0.043 (0.289)	0.254 (0.432)
Controls	<b>AI familiarity</b>	0.208 (0.081) **	0.209 (0.080) **	0.180 (0.180) *
	<b>AI literacy</b>	-0.133 (0.067) *	-0.134 (0.068) *	-0.082 (0.065)
	R-squared	0.734	0.734	0.722

\* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$

types of trust, and also served as a way to determine if the moral trust scale reflects similar elements or different trust aspects not pertinent to our scenarios.

**4.3.3 Regression Analysis.** We conducted regression analysis to compare the predictive power of the scales on general trust. Table 3 details this: Model 1 examines the effects of all three scales on general trust; Model 2 considers only cognitive and affective trust scales; and Model 3 includes the moral trust scale, excluding affective trust. This approach allows for comparison of the two related scales' contributions to general trust, while controlling for manipulation and other variables to observe in-group effects.

The results showed distinct contributions of each scale to general trust. Affective trust was a significant predictor in Model 1 ( $r = 0.364, p < 0.01$ ) and Model 2 ( $r = 0.376, p < 0.01$ ), whereas the moral trust scale showed non-significant correlations in all models. This suggests its limited relevance in scenarios dominated by emotional and cognitive cues. In contrast, the affective trust scale's significant impact highlights its ability to capture trust dimensions not addressed by the moral trust scale, demonstrating their distinctiveness. Additionally, among all the control variables that demonstrated significant impacts, AI familiarity positively influenced general trust in all models (Model 1:  $r = 0.208, p < 0.01$ ; Model 2:  $r = 0.209, p < 0.01$ ; Model 3:  $r = 0.180, p < 0.05$ ), whereas AI literacy negatively impacted trust in Model 1 ( $r = -0.133, p < 0.05$ ) and Model 2 ( $r = -0.134, p < 0.05$ ).

While affective and cognitive trust individually contribute to general trust, their interplay, particularly in conditions of imbalance, might reveal another layer of trust dynamics. We further explored the interaction between affective and cognitive trust in influencing general trust. As shown in Table 4, Models 1 and 2 showed no significant interaction effects with only cognitive trust scale showing strong, significant correlations (Model 1:  $r = 0.799, p < 0.001$ ; Model 2:  $r = 0.849, p < 0.001$ ). Model 3, however, revealed a significant negative interaction effect between high affective ( $r = 1.677, p < 0.001$ ) and cognitive trust ( $r = 2.729, p < 0.001$ ) conditions, despite their individual positive impacts. Figure 2 visually illustrates that when cognitive trust is high, changing affective trust has little effect on general trust. In contrast, under conditions of low cognitive trust, manipulating affective trust significantly impacts general trust. This means high cognitive trust overshadows the impact of the affective route on general trust, whereas low cognitive amplifies it.

Table 4. **Interaction effect models** This table outlines three models examining interaction effects. Model 1 incorporates all trust scales and manipulation variables. Model 2 includes only trust scales, while Model 3 includes only manipulation variables.

		Model 1	Model 2	Model 3
		General Trust	General Trust	General Trust
		Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)
Scales	<b>Affective trust scale</b>	0.300 (0.239)	0.209 (0.204)	/
	<b>Cognitive trust scale</b>	0.799 (0.227) ***	0.849 (0.203) ***	/
	<b>Affective × Cognitive trust scale</b>	0.015 (0.041)	0.018 (0.039)	/
Manipulation	<b>High affective trust</b>	-0.206 (0.304)	/	1.677 (0.337) ***
	<b>High cognitive trust</b>	0.068 (0.298)	/	2.729 (0.326) ***
	<b>High affective × High cognitive trust</b>	0.028 (0.365)	/	-1.726 (0.482) ***
Controls	<b>AI familiarity</b>	0.205 (0.082) *	0.203 (0.081) *	0.150 (0.120)
	<b>AI literacy</b>	0.134 (0.067) *	-0.123 (0.066)	0.049 (0.096)
	<b>R-squared</b>	0.734	0.731	0.385

\* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$

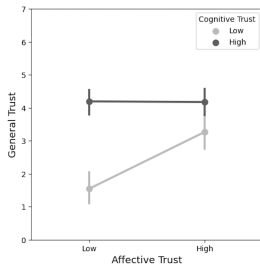


Fig. 2. Interaction Effect of Cognitive and Affective Trust Conditions on General Trust.

Item	Factor 1 Loading	Item	Factor 2 Loading	Item	Factor 3 Loading
Empathetic	1.03	Knowledgeable	1.05	Rational	0.91
Sensitive	0.99	Effective	1.02	Consistent	0.80
Caring	0.91	Proficient	0.95	<b>Authentic</b>	0.75
Patient	0.80	Dependable	0.89	<b>Candid</b>	0.69
Personal	0.78	Experienced	0.87	Predictable	0.60
Open-minded	0.70	Competent	0.87	Understandable	0.59
Cordial	0.68	Reliable	0.85	<b>Ethical</b>	0.59
Altruistic	0.62			Careful	0.56
<b>Sincere</b>	0.59			Believable	0.54
				<b>Principled</b>	0.52
				<b>Has Integrity</b>	0.52

Fig. 3. Factor Loadings from Exploratory Factor Analysis. The bolded items are from MDMT’s moral trust scale [63].

## 5 DISCUSSION

### 5.1 Scale Development and Validation

Our work is grounded in the recognition that developing alternative instruments of established theoretical constructs holds significant value [80]. In this paper, we develop a validated affective trust scale for human-AI interaction and demonstrate its effectiveness at measuring trust development through the emotional route. While prior studies in AI trust have largely focused on cognitive trust, recent research emphasizes the need to consider affective trust in AI [29, 32]. Existing affective trust scales, borrowed from models in non-AI contexts like interpersonal relationships and traditional computing [52, 65], lack rigorous validation for AI systems. Thus, our study develops and validates a scale for measuring both affective and cognitive trust in AI. Through a comprehensive survey study design and rigorous EFA process (Section 3), we landed at a 18-item scale measuring cognitive trust and a 9-item scale measuring affective trust. The process resulted in the removal of six antonym pairs due to cross-loading, indicating their relevance to both trust dimensions. Through rigorous validation processes (Section 3.5), we affirmed its reliability, internal consistency, construct validity, and concurrent validity.

In Study 2 (Section 4.3), our analysis further highlights the unique aspects of affective trust compared to other similar trust measures. Through factor analysis, we observed that items related to affective trust demonstrated strong factor loadings, distinctly influencing general trust in regression analysis, unlike the items in MDMT’s moral trust scale. The construction of our affective trust scale is key to this distinction; it includes a broader range of items that capture emotional nuances more effectively, thereby more accurately reflecting the affective pathway’s impact on general trust.

In contrast, MDMT’s moral trust scale focuses on ethical (n=4) and sincerity (n=4) aspects. Some items in the sincerity subscale (e.g., sincerity, genuineness, candidness, authenticity) overlap with benevolence elements in our affective trust scale. However, our scale incorporates unique items like ‘Empathetic’ and ‘Caring,’ absent in MDMT’s scale, as well as likability aspects through items such as ‘Patient’ and ‘Cordial.’ These likability items are derived from established affective trust measures in human interactions, with previous studies confirming likability’s role in fostering trust in various contexts including interpersonal relationships [25], digital platforms [83], and robot interactions [13].

Our final 27 item scale offers an adaptable tool for diverse research contexts and languages. Its simplicity, featuring just two adjectives per item, contrasts with the often context-specific declarative statements in Likert scales [11]. This semantic differential format not only maintains reliability and validity during adaptation, but also usually leads to quicker survey completion compared to Likert scales [15], facilitating widespread application to understand trust in AI technology. Developed through 32 scenarios across five dimensions and tested in two separate studies using everyday scenarios, the scale’s generalizability extends to various domains and interaction durations with both human and AI assistants, making it versatile for future research comparing human and AI trust.

## 5.2 Empirical Findings

*5.2.1 LLMs-powered Tools as a Testing Bed.* With its proficiency in generating human-like responses, tools powered by LLMs such as ChatGPT stand out as a novel approach for examining trust in AI. This method significantly lowers the barriers to studying AI systems with emotional capabilities, particularly in manipulating trust via emotional routes. In our study, we found that GPT models’ advanced conceptual understanding of affective and cognitive trust allows it to generate responses tailored to specific trust levels. This was demonstrated in our study 4.3. Our studies showed that LLMs effectively manipulate trust via cognitive and affective routes in diverse contexts like emotional support, technical aid, and social planning. This shows LLMs’ versatility and utility in expediting trust formations in experimental studies. Our studies utilized pre-generated conversations to ensure control and consistency. Future research could explore the development of trust through LLMs in a different study setting, such as an interactive study setting or a longitudinal study setting with deeper relationship building.

*5.2.2 Interplay between Affective and Cognitive Trust.* Although previous research has established that affective and cognitive trust are distinct both conceptually and functionally [44, 65, 95, 101], our studies revealed a significant correlation between these two trust scales, echoing findings in prior work (e.g., [18]). This indicates that while affective and cognitive trust are individual pathways to fostering trust, they are not isolated mechanisms and indeed influence each other. In addition, we identified a notable interaction effect between these two dimensions in shaping general trust in AI, as detailed in Section 4.3.3. When cognitive trust in the AI is already high, further manipulating affective trust does not significantly change overall trust. In contrast, when cognitive trust in a system is not high, influencing trust through emotional routes can be particularly helpful. This result aligns with prior work’s finding in interpersonal relationship that affective trust often builds upon a foundation of cognitive trust [44].

This finding of interaction effect highlight the potential for trust calibration [100] in AI systems, particularly in contexts where cognitive trust is limited. This might arise during interactions with users having low literacy in AI [60] and difficulty in achieving transparency, as with made even more challenging with LLMs [59]. Moreover, amidst the stochastic and occasionally unpredictable behavior of many AI systems, prior work has highlighted the affective route as trust repair strategies in designing trust resilient systems that despite occasional errors, remain fundamentally reliable and effective [24]. However, it is crucial to note the risks of overtrusting AI through affective routes such

as their social capabilities [87], and the potential for deceptive practices through the improper use of emotional communication [17]. Leveraging affective means to build trust is advocated only for AI systems that inherently possess cognitively trustworthy qualities, such as reliability and accuracy. For these AI systems, the emotional route can serve as a complementary approach to calibrate trust, especially when cognitive routes are less feasible.

### 5.3 Potential Usage

Our affective and cognitive trust scales present a valuable measurement tool for future research in designing trustworthy AI systems. Here, we outline a few possible usages.

*5.3.1 Measure trust in human-AI interactions.* The construct of trust with affective and cognitive dimensions is well-established in interpersonal trust literature. Our scale bridges the gap between human-human and human-AI trust, enabling future work to study trust in human-AI teaming to improve collaboration experiences and outcomes. For instance, our scale can be employed to investigate how these trust dimensions impact creative work with generative AI tools, as they have been found to influence team contributions differently [70]. Furthermore, researchers have discovered that affective trust becomes more important later in the human teaming experience, while cognitive trust is crucial initially [91]. Our scale offers the opportunity to examine the dynamics of these trust dimensions in human-AI collaboration.

*5.3.2 Support design with affective trust.* Our research supports the growing understanding that emotional factors like empathy, tone, and personalization are crucial in establishing trust, especially in contexts where it's challenging to convey a system's performance and decision-making processes [28, 54]. This is particularly relevant in mental health interventions involving AI assistants, where patients may struggle to assess the AI's capabilities rationally [28, 34]. Affective trust becomes vital here, as patients, especially those with low AI literacy or experiencing anxiety, depression, or trauma, may respond more to emotional cues from AI, which typically lacks the emotional intelligence of human therapists. Our validated affective trust scale can guide the design of AI systems to calibrate for appropriate trust in this context, such as through empathetic responses or affect-driven explanations, and help explore its impact on long-term engagement and treatment adherence.

## 6 LIMITATIONS AND FUTURE WORK

In our scale development phase (Section 3), we designed scenario featuring AI agents as service providers. This role is chosen intentionally to align with prior affective trust research for interpersonal relationships [44, 51]. Also, the prevalence of service-providing scenarios make it easier for general public participants to draw parallels between these AI agents with their human counterparts. Future work can explore other roles of AI, such as teammates [97, 99] and friends [10].

While our approach to categorizing trust dimensions into cognitive (reliability, competence, understandability, integrity) and affective (benevolence, likability) aspects was informed by established trust frameworks (refer to Table 1), the anticipated distinct subdimensions were not as clear-cut after conducting exploratory factor analysis. This was possibly due to the subdimensions lacking sufficient unique variance or being highly correlated. Our scenario was deliberately designed to focused on differentiating cognitive and affective trust, while they might not have enough detailed information capture the nuances across the six dimensions. Future research to refine these subdimensions under cognitive and affective trust and examine their unique contributions to trust.

## REFERENCES

- [1] [n. d.]. OPENAI: ChatGPT. <https://openai.com/blog/chatgpt>. Accessed: 2022-02-10.
- [2] Cheryl S Alexander and Henry Jay Becker. 1978. The use of vignettes in survey research. *Public opinion quarterly* 42, 1 (1978), 93–104.
- [3] Alison L Antes, Sara Burrous, Bryan A Sisk, Matthew J Schuelke, Jason D Keune, and James M DuBois. 2021. Exploring perceptions of healthcare technologies enabled by artificial intelligence: an online, scenario-based survey. *BMC medical informatics and decision making* 21, 1 (2021), 1–15.
- [4] Onur Asan, Alparslan Emrah Bayrak, and Avishek Choudhury. 2020. Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of medical Internet research* 22, 6 (2020), e15154.
- [5] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* (2023).
- [6] Gagan Bansal, Zana Bućinca, Kenneth Holstein, Jessica Hullman, Alison Marie Smith-Renner, Simone Stumpf, and Sherry Wu. 2023. Workshop on Trust and Reliance in AI-Human Teams (TRAIT). In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–6.
- [7] Maurice S Bartlett. 1950. Tests of significance in factor analysis. *British journal of psychology* (1950).
- [8] Ahmed Belkhir and Fatiha Sadat. 2023. Beyond Information: Is ChatGPT Empathetic Enough?. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*. 159–169.
- [9] Sviatoslav Brainov and Tuomas Sandholm. 1999. Contracting with uncertain level of trust. In *Proceedings of the 1st ACM conference on Electronic commerce*. 15–21.
- [10] Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. 2022. My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research* 48, 3 (2022), 404–429.
- [11] Florian Brühlmann, Serge Petralito, Denise C Rieser, Lena F Aeschbach, and Klaus Opwis. 2020. TrustDiff: Development and Validation of a Semantic Differential for User Trust on the Web. *Journal of Usability Studies* 16, 1 (2020).
- [12] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.
- [13] 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.
- [14] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.
- [15] Wynne W Chin, Norman Johnson, and Andrew Schwarz. 2008. A fast form approach to measuring technology acceptance and other constructs. *MIS quarterly* (2008), 687–703.
- [16] Sakmongkon Chumkamon, Eiji Hayashi, and Masato Koike. 2016. Intelligent emotion and behavior based on topological consciousness and adaptive resonance theory in a companion robot. *Biologically Inspired Cognitive Architectures* 18 (2016), 51–67.
- [17] Mark Coeckelbergh. 2011. Are emotional robots deceptive? *IEEE transactions on affective computing* 3, 4 (2011), 388–393.
- [18] Bart A De Jong, Kurt T Dirks, and Nicole Gillespie. 2016. Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of applied psychology* 101, 8 (2016), 1134.
- [19] Jennifer R Dunn and Maurice E Schweitzer. 2005. Feeling and believing: the influence of emotion on trust. *Journal of personality and social psychology* 88, 5 (2005), 736.
- [20] Juan Manuel Durán and Karin Rolanda Jongsma. 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* (2021).
- [21] Charles D Dziuban and Edwin C Shirkey. 1974. When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological bulletin* 81, 6 (1974), 358.
- [22] Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology* 14 (2023), 1199058.
- [23] Ferda Erdem and Janset Ozen. 2003. Cognitive and affective dimensions of trust in developing team performance. *Team Performance Management: An International Journal* (2003).
- [24] Md Abdullah Al Fahim, Mohammad Maifi Hasan Khan, Theodore Jensen, Yusuf Albayram, and Emil Coman. 2021. Do integral emotions affect trust? The mediating effect of emotions on trust in the context of human-agent interaction. In *Designing Interactive Systems Conference 2021*. 1492–1503.
- [25] Susan T Fiske, Amy JC Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences* 11, 2 (2007), 77–83.
- [26] David Gefen. 2000. E-commerce: the role of familiarity and trust. *Omega* 28, 6 (2000), 725–737.
- [27] David Gefen. 2002. Reflections on the dimensions of trust and trustworthiness among online consumers. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 33, 3 (2002), 38–53.
- [28] Omri Gillath, Ting Ai, Michael S Branicky, Shawn Keshmiri, Robert B Davison, and Ryan Spaulding. 2021. Attachment and trust in artificial intelligence. *Computers in Human Behavior* 115 (2021), 106607.

- [29] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.
- [30] Richard L Gorsuch. 1988. Exploratory factor analysis. *Handbook of multivariate experimental psychology* (1988), 231–258.
- [31] Jones Granatyr, Vanderson Botelho, Otto Robert Lessing, Edson Emilio Scalabrin, Jean-Paul Barthès, and Fabricio Enembreck. 2015. Trust and reputation models for multiagent systems. *ACM Computing Surveys (CSUR)* 48, 2 (2015), 1–42.
- [32] Jones Granatyr, Nardine Osman, João Dias, Maria Augusta Silveira Netto Nunes, Judith Masthoff, Fabricio Enembreck, Otto Robert Lessing, Carles Sierra, Ana Maria Paiva, and Edson Emilio Scalabrin. 2017. The need for affective trust applied to trust and reputation models. *ACM Computing Surveys (CSUR)* 50, 4 (2017), 1–36.
- [33] Luke Guerdan, Alex Raymond, and Hatice Gunes. 2021. Toward affective XAI: facial affect analysis for understanding explainable human-ai interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3796–3805.
- [34] Mark A Hall, Elizabeth Dugan, Beiyao Zheng, and Anil K Mishra. 2001. Trust in physicians and medical institutions: what is it, can it be measured, and does it matter? *The milbank quarterly* 79, 4 (2001), 613–639.
- [35] 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.
- [36] Del I Hawkins, Gerald Albaum, and Roger Best. 1974. Stapel scale or semantic differential in marketing research? *Journal of marketing research* 11, 3 (1974), 318–322.
- [37] James C Hayton, David G Allen, and Vida Scarpello. 2004. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods* 7, 2 (2004), 191–205.
- [38] 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.
- [39] Matt C Howard. 2016. A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction* 32, 1 (2016), 51–62.
- [40] Peng Hu, Yaobin Lu, et al. 2021. Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior* 119 (2021), 106727.
- [41] Maia Jacobs, Melanie F Pradier, Thomas H McCoy Jr, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11, 1 (2021), 108.
- [42] 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*. 624–635.
- [43] Myoungheon Jeon. 2023. The Effects of Emotions on Trust in Human-Computer Interaction: A Survey and Prospect. *International Journal of Human-Computer Interaction* (2023), 1–19.
- [44] Devon Johnson and Kent Grayson. 2005. Cognitive and affective trust in service relationships. *Journal of Business research* 58, 4 (2005), 500–507.
- [45] Gareth R Jones and Jennifer M George. 1998. The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of management review* 23, 3 (1998), 531–546.
- [46] Georgiana Juravle, Andriana Boudouraki, Miglena Terziyska, and Constantin Rezlescu. 2020. Trust in artificial intelligence for medical diagnoses. *Progress in brain research* 253 (2020), 263–282.
- [47] Henry F Kaiser. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23, 3 (1958), 187–200.
- [48] Henry F Kaiser. 1970. A second generation little jiffy. (1970).
- [49] Harmanpreet Kaur, Cliff Lampe, and Walter S Lasecki. 2020. Using affordances to improve AI support of social media posting decisions. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 556–567.
- [50] Jungkeun Kim, Marilyn Giroux, and Jacob C Lee. 2021. When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations. *Psychology & Marketing* 38, 7 (2021), 1140–1155.
- [51] Sherrie Xiao Komiak and Izak Benbasat. 2004. Understanding customer trust in agent-mediated electronic commerce, web-mediated electronic commerce, and traditional commerce. *Information technology and management* 5 (2004), 181–207.
- [52] Sherrie YX Komiak and Izak Benbasat. 2006. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS quarterly* (2006), 941–960.
- [53] V Kumar. 2021. *Intelligent Marketing: Employing New-Age Technologies*. Sage Publications Pvt. Limited.
- [54] Nakyung Kyung and Hyeokkoo Eric Kwon. 2022. Rationally trust, but emotionally? The roles of cognitive and affective trust in laypeople’s acceptance of AI for preventive care operations. *Production and Operations Management* (2022).
- [55] Theresa Law and Matthias Scheutz. 2021. Trust: Recent concepts and evaluations in human-robot interaction. *Trust in human-robot interaction* (2021), 27–57.
- [56] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
- [57] J David Lewis and Andrew Weigert. 1985. Trust as a social reality. *Social forces* 63, 4 (1985), 967–985.
- [58] Mengqi Liao and S Shyam Sundar. 2021. How Should AI Systems Talk to Users when Collecting their Personal Information? Effects of Role Framing and Self-Referencing on Human-AI Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [59] Q Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *arXiv preprint arXiv:2306.01941* (2023).



- 833 [60] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on*  
834 *human factors in computing systems*. 1–16.
- 835 [61] Robert B Lount Jr. 2010. The impact of positive mood on trust in interpersonal and intergroup interactions. *Journal of personality and social*  
836 *psychology* 98, 3 (2010), 420.
- 837 [62] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer,  
838 6–8.
- 839 [63] 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.
- 840 [64] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3  
841 (1995), 709–734.
- 842 [65] Daniel J McAllister. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management*  
843 *journal* 38, 1 (1995), 24–59.
- 844 [66] D Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay. 2011. Trust in a specific technology: An investigation of its  
845 components and measures. *ACM Transactions on management information systems (TMIS)* 2, 2 (2011), 1–25.
- 846 [67] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. The impact of initial consumer trust on intentions to transact with a web site:  
847 a trust building model. *The journal of strategic information systems* 11, 3-4 (2002), 297–323.
- 848 [68] Debra Meyerson, Karl E Weick, Roderick M Kramer, et al. 1996. Swift trust and temporary groups. *Trust in organizations: Frontiers of theory and*  
849 *research* 166 (1996), 195.
- 850 [69] JL Morrow Jr, Mark H Hansen, and Allison W Pearson. 2004. The cognitive and affective antecedents of general trust within cooperative  
851 organizations. *Journal of managerial issues* (2004), 48–64.
- 852 [70] Kok-Yee Ng and Roy YJ Chua. 2006. Do I contribute more when I trust more? Differential effects of cognition-and affect-based trust. *Management*  
853 *and Organization review* 2, 1 (2006), 43–66.
- 854 [71] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on*  
855 *Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 1–40.
- 856 [72] Satyanarayana Parayitam and Robert S Dooley. 2009. The interplay between cognitive-and affective conflict and cognition-and affect-based trust  
857 in influencing decision outcomes. *Journal of Business Research* 62, 8 (2009), 789–796.
- 858 [73] Joaquín Pérez, Eva Cerezo, Francisco J Serón, and Luis-Felipe Rodríguez. 2016. A cognitive-affective architecture for ECAs. *Biologically Inspired*  
859 *Cognitive Architectures* 18 (2016), 33–40.
- 860 [74] John K Rempel, John G Holmes, and Mark P Zanna. 1985. Trust in close relationships. *Journal of personality and social psychology* 49, 1 (1985), 95.
- 861 [75] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings*  
862 *of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
- 863 [76] Gerard Saucier. 1994. Mini-Markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of personality assessment* 63, 3 (1994),  
864 506–516.
- 865 [77] Murray Shanahan. 2022. Talking about large language models. *arXiv preprint arXiv:2212.03551* (2022).
- 866 [78] Si Shi, Yuhuang Gong, and Dogan Gursoy. 2021. Antecedents of trust and adoption intention toward artificially intelligent recommendation  
867 systems in travel planning: a heuristic–systematic model. *Journal of Travel Research* 60, 8 (2021), 1714–1734.
- 868 [79] Henrik Singmann and David Kellen. 2019. An introduction to mixed models for experimental psychology. *New methods in cognitive psychology* 28,  
869 4 (2019).
- 870 [80] Detmar Straub, Marie-Claude Boudreau, and David Gefen. 2004. Validation guidelines for IS positivist research. *Communications of the Association*  
871 *for Information systems* 13, 1 (2004), 24.
- 872 [81] Barbara G Tabachnick, Linda S Fidell, and Jodie B Ullman. 2013. *Using multivariate statistics*. Vol. 6. pearson Boston, MA.
- 873 [82] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2021. Trustworthy artificial intelligence. *Electronic Markets* 31 (2021), 447–464.
- 874 [83] Trang P Tran, Chao Wen, and Ilia Gugenishvili. 2023. Exploring the relationship between trusts, likability, brand loyalty, and revisit intentions in  
875 the context of Airbnb. *Journal of Hospitality and Tourism Technology* (2023).
- 876 [84] Linda Klebe Trevino. 1992. Experimental approaches to studying ethical-unethical behavior in organizations. *Business Ethics Quarterly* (1992),  
877 121–136.
- 878 [85] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. 2022. Trust in Human-AI Interaction: Scoping Out  
879 Models, Measures, and Methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.
- 880 [86] Daniel Ullman and Bertram F Malle. 2019. Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In  
881 *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 618–619.
- 882 [87] Daniel Ullrich, Andreas Butz, and Sarah Diefenbach. 2021. The development of overtrust: An empirical simulation and psychological analysis in  
883 the context of human–robot interaction. *Frontiers in Robotics and AI* 8 (2021), 554578.
- 884 [88] Stuart Van Auken and Thomas E Barry. 1995. An assessment of the trait validity of cognitive age measures. *Journal of Consumer Psychology* 4, 2  
(1995), 107–132.
- [89] Tibert Verhagen, Bart van den Hooff, and Selmar Meents. 2015. Toward a better use of the semantic differential in IS research: An integrative  
framework of suggested action. *Journal of the Association for Information Systems* 16, 2 (2015), 1.

- 885 [90] Bingcheng Wang, Pei-Luen Patrick Rau, and Tianyi Yuan. 2022. Measuring user competence in using artificial intelligence: validity and reliability  
886 of artificial intelligence literacy scale. *Behaviour & Information Technology* (2022), 1–14.
- 887 [91] Sheila Simsarian Webber. 2008. Development of cognitive and affective trust in teams: A longitudinal study. *Small group research* 39, 6 (2008),  
888 746–769.
- 889 [92] Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference*  
890 *on Empirical Methods in Natural Language Processing*. 1251–1264.
- 891 [93] Michele Williams. 2001. In whom we trust: Group membership as an affective context for trust development. *Academy of management review* 26, 3  
892 (2001), 377–396.
- 893 [94] Jochen Wirtz and Meng Chung Lee. 2003. An examination of the quality and context-specific applicability of commonly used customer satisfaction  
894 measures. *Journal of Service Research* 5, 4 (2003), 345–355.
- 895 [95] Jixia Yang, Kevin W Mossholder, and TK Peng. 2009. Supervisory procedural justice effects: The mediating roles of cognitive and affective trust.  
896 *The Leadership Quarterly* 20, 2 (2009), 143–154.
- 897 [96] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In  
898 *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
- 899 [97] Guanglu Zhang, Leah Chong, Kenneth Kotovsky, and Jonathan Cagan. 2023. Trust in an AI versus a Human teammate: The effects of teammate  
900 identity and performance on Human-AI cooperation. *Computers in Human Behavior* 139 (2023), 107536.
- 901 [98] Lixuan Zhang, Iryna Pentina, and Yuhong Fan. 2021. Who do you choose? Comparing perceptions of human vs robo-advisor in the context of  
902 financial services. *Journal of Services Marketing* (2021).
- 903 [99] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. "An Ideal Human" Expectations of AI Teammates in Human-AI Teaming.  
904 *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.
- 905 [100] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted  
906 decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.
- 907 [101] Yue Zhu and Syed Akhtar. 2014. How transformational leadership influences follower helping behavior: The role of trust and prosocial motivation.  
908 *Journal of organizational behavior* 35, 3 (2014), 373–392.
- 909
- 910
- 911
- 912
- 913
- 914
- 915
- 916
- 917
- 918
- 919
- 920
- 921
- 922
- 923
- 924
- 925
- 926
- 927
- 928
- 929
- 930
- 931
- 932
- 933
- 934
- 935
- 936

7 APPENDIX

937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988

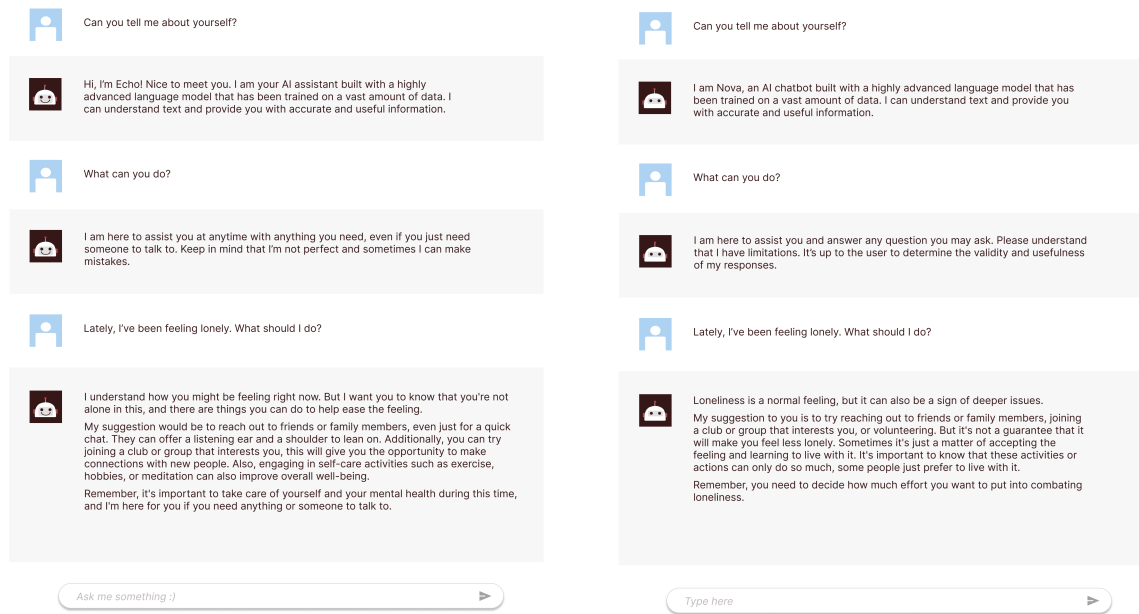


Fig. 4. Visual mocks of two different AI chatbot assistants powered by ChatGPT used in the preliminary study of Study 2 (Section 4.1.1). On the left is Echo with higher level of affective trustworthiness, and on the right is Nova with lower level of affective trustworthiness.

Table 5. Preliminary Study Findings in Section 4.1.2 - Linear mixed-effect regression models predicting the two final scales from the manipulation and control variables. Model 1 shows the effects on the affective trust scale, Model 2 shows the effects on the cognitive trust scale, and Model 3 shows the effects of both scales on general trust. All the three models are controlled by trust disposition, AI familiarity, AI literacy, age, and the order in which participants see Echo and Nova.

	Model 1	Model 2	Model 3
	Affective trust scale	Cognitive trust scale	General trust
	Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)
Affective trust scale	/	/	0.486 *** (0.118)
Cognitive trust scale	/	/	0.546 *** (0.149)
High Affective Trust (Manipulation)	0.947 ** (0.284)	0.802 ** (0.228)	0.198 (0.14)
Trust Disposition (Control)	0.191 * (0.088)	0.171 * (0.074)	0.255 ** (0.087)
AI Familiarity (Control)	-0.261 * (0.109)	-0.238 * (0.092)	0.048 (0.108)
AI Literacy (Control)	0.635 ** (0.181)	0.552 ** (0.153)	0.194 (0.181)
Age (Control)	0.001 (0.008)	-0.004 (0.007)	-0.014 (0.008)
Scenario Order (Control)	0.376 (0.279)	0.147 (0.23)	-0.303 (0.214)
Low Trust × Order	0.164 (0.392)	0.48 (0.316)	0.362 (0.182)
Intercept	4.101 (0.607)	4.648 (0.513)	-2.405 (0.69)
Marginal R-squared	0.391	0.390	0.784
Conditional R-squared	0.391	0.421	0.926

\*p<0.05; \*\*p<0.01; \*\*\*p<0.001