Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Construction of the Avatar Trust Scale (ATS): A First Step Toward Measuring Trust in Avatars

Jule Schilling
University of Würzburg
Johrine Cronjé
Maritime Research Institute Netherlands
Jinghuai Lin
University of Würzburg
Michael F. Clements
King's College London
Jessica Fahn
University of Würzburg
Giorgio Ballestin
Maritime Research Institute Netherlands
Ivo Käthner
HMU Health and Medical University
Marc Erich Latoschik
University of Würzburg
Johannes Rodrigues
johannes.rodrigues@uni-wuerzburg.de

University of Würzburg

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11301260/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Construction of the Avatar Trust Scale (ATS): A First Step Toward Measuring Trust in Avatars

Jule Schilling[1,*], Johrine Cronjé[2,5,*], Jinghuai Lin[3], Michael F. Clements[4], Jessica Fahn[2], Giorgio Ballestin[5], Ivo Käthner[6], Marc Erich Latoschik[3], and Johannes Rodrigues[1,†]

[1]University of Würzburg, Department of Psychology V: Differential Psychology, Personality Psychology and Psychological Diagnostics, Würzburg, 97070, Germany
[2]University of Würzburg, Department of Psychology I: Biological Psychology, Clinical Psychology and Psychotherapy, Würzburg, 97070, Germany
[3]University of Würzburg, Chair for Human-Computer Interaction (HCI), Würzburg, 97074, Germany
[4]Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, London, SE5 8AF, United Kingdom
[5]Maritime Operations Department of the Maritime Research Institute Netherlands (MARIN), Wageningen, 6708, Netherlands
[6]Health and Medical University, Department of Health and Sports: Medical Psychology, Potsdam, 14471, Germany
[*]These authors contributed equally to this work.
[†]johannes.rodrigues@uni-wuerzburg.de

ABSTRACT

Avatars serve as virtual representations of users in digital environments. Their appearance can be freely customized, often leading to discrepancies between visual representation and the user's real identity. Such appearances may unconsciously influence interpersonal trust, complicating the assessment of trustworthiness in virtual interactions. To investigate how avatar characteristics influence trust, we developed and preliminarily validated the Avatar Trust Scale (ATS), a tool designed to quantify trust specifically in the context of avatars. Based on data from a German sample of 840 participants, our findings suggest a two-factor structure: Specific Trust (*trustworthiness, security*, and *willingness to interact*) and General Trust (*propensity to trust avatars and humans*, and *avatar literacy*). The scale was validated using questionnaire data and behavioral tasks (i.e., trust game, dictator game, ultimatum game). Participants interacted with a virtual agent rather than a user-controlled avatar, providing initial validation in a controlled context and laying the groundwork for future research on avatar-mediated trust.

Introduction

Trust is an exceedingly important construct in human life. It has been approached from various scientific perspectives: behavioral economics, organization science, psychology, sociology, and neuroscience[1,2]. Due to technical innovations, global catastrophes like the COVID-19 pandemic and associated contact restrictions, the extensions of human interpersonal trust[3] to virtual representations have become more important. Different models of trust have been formulated, including but not limited to trust in humans[1,2,4–7]. However, a systematic method for assessing trust in avatars has not yet been developed. Our work aims to provide an initial questionnaire for measuring trust toward avatars controlled by humans.

One can distinguish between the conceptual definitions and technical implementations of *virtual humans*, (virtual) *avatars*, and *virtual agents*. *Virtual humans* are digital entities with human-like appearances in virtual worlds, which can be controlled by humans, pre-scripted systems, or AI algorithms[8]. *Avatars* are virtual representations of real users, often represented as virtual humans but not limited to that form[9,10]. *Agents* are autonomous digital entities that act within virtual environments, typically controlled by AI or scripted logic rather than a human operator[11]. For example, in a virtual meeting room, a virtual human might appear as a realistic, animated humanoid character. If it is operated in real time by a participant, it functions as an avatar. If it responds automatically using preprogrammed scripts or AI, it is a virtual agent. Alternatively, a user's avatar in the same environment might not resemble a virtual human at all: it could be a static photo, a stylized icon, or a non-human 3D model, depending on the platform and user preferences.

General trust refers to the tendency to trust over time, across different contexts, and toward various relevant entities[3,12–14]. High general trust indicates that an individual is likely to make positive assumptions about others and to expect goodwill and benign intent, regardless of familiarity. Often considered a trait, it reflects a general disposition to trust across contexts such as interactions with others, information sources, experts and expertise, institutions, and governments[3,12–14]. General trust plays a

crucial role in enabling trust in others when specific information about them is lacking[14].

For our purposes, we divided trust into the *propensity to trust humans*[3,15] and the *propensity to trust avatars*, both reflecting a general tendency to trust either humans or avatars. *Trusting avatars* refers to an individual's inclination to trust virtual representations of other people, such as those encountered in shared virtual environments[16]. *Trusting humans*, by contrast, refers to an individual's inclination to trust other human beings in face-to-face or traditional interpersonal interactions[3]. Both concepts are grounded in interpersonal trust theories[3,4,15,17,18]. However, for avatars, technical aspects, their implementation, and trust in these implementations are integral to the concept, whereas this is not the case for humans. This work focuses on trust toward avatars (i.e., virtual humans or other digital entities controlled by real users). While the term "avatar" is sometimes used more broadly in public discourse to include computer-controlled entities or static digital representations[19], our scale was not designed for these interpretations.

Besides recreational and entertainment purposes, Social Virtual Reality use cases of today also target various fields of application, e.g., therapeutic settings, medicine, politics, or the financial industry[20,21]. Interpersonal trust scales such as the Specific Interpersonal Trust Scale (SITS)[15] assess trust in known individuals; however, interacting with avatars presents additional challenges. The appearance of avatars can be arbitrary, and users can take on any digital body they wish[22]. Thus, users cannot directly verify the identity of the person behind the avatar, introducing uncertainty about authenticity and intentions. Trust in avatars therefore encompasses not only interpersonal expectations but also perceptions of the avatar's design, privacy and technical reliability. Existing instruments do not capture these dimensions, that is why a dedicated measure is needed.

Several studies have examined how trust forms in virtual environments. Pan and Steed showed that participants who interacted with self--avatars invested more credits in a repeated trust game than those without. In addition, participants indicated a higher self-reported trust in the conditions with a present avatar compared to those with no avatar. Interestingly, there was no difference between self-reported trust in the self-avatar condition compared to the face-to-face human interaction condition[23]. In another study, they compared avatar-, video- and robot--mediated advice and found that participants were less willing to seek advice from avatars that did not use a virtual human representation[24]. George et al. (2018) used a standard trust game to compare virtual human and robot-looking avatars and found similar amounts of credits sent to both, although feelings of social presence differed[25]. These investigations rely on economic games or advice--seeking tasks to infer trust, capturing only specific behaviors. Several studies highlighted the need for a standardized questionnaire applicable to broader contexts[26,27]. Our study takes a step toward addressing this gap by developing the Avatar Trust Scale (ATS).

A general trait that facilitates trust experiences in the context of avatar-related trust is *avatar literacy*. Avatar literacy refers to the knowledge, skill, and understanding individuals have about the creation, customization, and effective use of avatars in virtual environments[28]. In our context, it includes the expectation of one's own ability to assess the trustworthiness of avatars: avatar-literate individuals understand the nuances of avatar design and behavior and can evaluate the authenticity and credibility of virtual humans. By being aware of techniques used to create or customize avatar appearances and behaviors, individuals are better able to distinguish technical artifacts from intentional behavior and to judge whether an avatar is genuinely trustworthy or attempting to deceive. Additionally, avatar literacy involves an understanding of social norms and conventions within virtual environments. This enables individuals to evaluate whether virtual humans adhere to these norms, which can influence trust assessments[29]. Thus, avatar literacy is a skill that supports nuanced trust assessments of avatars. It helps individuals navigate and participate effectively in virtual environments, using avatars for communication, self-expression, and meaningful interaction within virtual communities. However, avatar literacy does not apply to trust towards humans. This makes it a unique factor for trust toward avatars and highlights the need for a deliberate questionnaire.

In addition to the general tendency to trust a human or virtual interaction partner, *specific trust* (i.e., trust toward a specific avatar or entity) was considered an important factor to assess. This concept was divided into three facets. The *perceived trustworthiness of the specific avatar* evaluates the avatar's qualities based on the broader concept of interpersonal trust[3,4,15]. The *feeling of being secure with the specific avatar*[15] addresses the perceived safety of the virtual representation. The *willingness to interact with the avatar*[7] reflects the user's readiness to engage in behavioral interaction in VR. As with general trust, specific trust in the avatar context is not entirely equivalent to specific trust in humans. The structure of the virtual world may influence user behavior, introducing distinct rules, boundaries, and limitations that shape how trust manifests in digital spaces[30]. In this context, specific trust refers to trust toward a particular avatar within a particular virtual environment.

The primary goal of this study was to construct and validate the ATS as a tool for assessing trust toward a specific virtual avatar. Based on the conceptual distinctions, we proposed two competing confirmatory factor models. *Model 1* ("Basic Model") treats all six subscales (*propensity to trust avatars, propensity to trust humans, avatar literacy, perceived trustworthiness, feeling secure,* and *willingness to interact*) as indicators of a single latent construct of virtual trust. *Model 2* ("Target Model") uses a hierarchical structure: the three general--trust subscales load onto a second--order *General Trust* factor, the three specific--trust subscales load onto a second--order *Specific Trust* factor, and these two factors load onto an overarching *Virtual Trust* factor (Fig. 1).

In addition to these two substantive models, we evaluated a baseline 0--model (also known as the null or independence

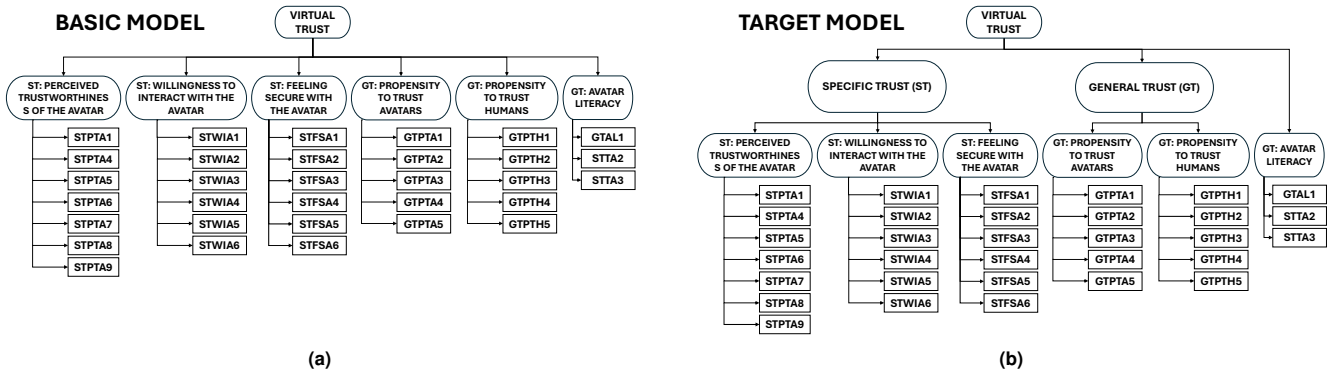


Figure 1. Structural equation models of the ATS. (a) Basic Model: all six subscales load onto a single latent Virtual Trust factor. (b) Target Model: the three specific-trust subscales (*perceived trustworthiness, feeling secure, willingness to interact*) load onto a higher-order *Specific Trust* factor; the two general-trust subscales (*propensity to trust avatars, propensity to trust humans*) load onto a higher-order *General Trust* factor. *Avatar literacy* loads directly onto the overall *Virtual Trust* factor, rather than being grouped with general trust. Both higher-order factors, along with avatar literacy, contribute to the latent Virtual Trust construct. A baseline 0--model, in which no latent factors are specified, was also tested for model comparison but is not depicted. Details of the confirmatory factor analysis are provided in the Results section.

model) in which no latent factors were specified and all inter-item covariances were fixed to zero. This model provides a baseline for model comparison; we hypothesized that both Model 1 and Model 2 would fit the data better than the 0--model. Furthermore, we expected that Model 2 would provide a superior fit because it reflects the theoretical distinction between general dispositions to trust and perceptions of a specific avatar. Table 1 lists all preregistered hypotheses relating to the factor structure and to convergent/divergent validity. Additional preregistered hypotheses and results are available in the Supplemental Material S1.

Throughout this paper we refer to *convergent validity*, the extent to which ATS subscales correlate positively with established measures of trust (e.g., the Interpersonal Trust Scale and the Specific Interpersonal Trust Scale)[15,31], and *divergent validity*, the extent to which ATS subscales show little or no correlation with theoretically unrelated constructs, such as vividness of visual imagery[32].

In summary, our aim is to develop and validate the ATS as a multidimensional instrument to measure trust toward avatars. Following the literature on interpersonal trust, we hypothesize that trust in avatars encompasses both *general* and *specific* aspects[2,33]. *General trust* reflects a person's *propensity to trust others* (humans or avatars) and their knowledge of avatar technology (*avatar literacy*), whereas *specific trust* captures perceptions of a particular avatar's *trustworthiness*, the *feeling of security when interacting with it*, and the *willingness to interact with it*. To test this structure, we formulated two competing confirmatory factor models: (i) a single--factor "Basic Model"(see Fig. 1a), and (ii) a hierarchical "Target Model" (see Fig. 1b).

Results

Structure of the ATS

Exploratory Factor Analysis

The results of the exploratory factor analysis showed that the indicator variables of all scales behaved as expected, with the exception of items two "*I believe this avatar would act in my best interest*" and three "*This avatar can be trusted in general*" from the specific trust scale (see Supplemental Material S2). These items showed either no relevant loadings or cross-loadings onto the general aspect of avatar trust. Once these items were excluded, the confirmatory factor analysis could be computed successfully (see also Preregistration and Changes). The raw dataset is available here: https://osf.io/tz693/ .

Confirmatory Factor Analysis

For the confirmatory factor analysis, the fit indices for the basic model were: $\chi^2(458) = 1642.48$, $p < .001$; RMSEA = .053 [90% CI: .050 –– .057]; CFI = .908; TLI = .900; SRMR = .062.

For the target model, the fit indices were: $\chi^2(457) = 1616.91$, $p < .001$; RMSEA = .053 [90% CI: .050 — .056]; CFI = .910; TLI = .902; SRMR = .061, indicating a slightly better fit compared to the basic model (supporting Hypothesis H1.2).

The 0-model demonstrated a significantly worse fit: $\chi^2(464) = 5735.65$, $p < .001$; RMSEA = .115 [90% CI: .112 –– .118]; CFI = .564; TLI = .533; SRMR = .110. Hence, the target model was deemed the best fitting model among the proposed options (supporting Hypotheses H1.1 and H1.2).

Variable	Hypotheses
	Subjective measurement instruments
ATS model structure, reliability and intercorrelation	**H1.1:** *Target model* and the *basic model* are better fits than the *0-model*. **H1.2:** *Target model* is a better fit than the *basic model*. **H1.3:** The subscales of the ATS show a decent homogeneity (Cronbach's α / McDonald's $\omega \geq 0.75$), except the *general trust avatar literacy* scale (Cronbach's α / McDonald's $\omega \geq 0.6$). **H1.4:** Moderate positive intercorrelation ($r \geq 0.3$) of the *specific trust* subscales and small positive ($r \geq 0.1$) correlation with the *general trust* subscales. **H1.5** Moderate positive intercorrelation ($r \geq 0.3$) of the *general trust* subscales and small positive ($r \geq 0.1$) correlation with the *specific trust* subscales.
ITS	**H2.1:** Strong positive correlation ($r \geq 0.5$) of *general trust* subscale: *propensity to trust humans* and the ITS overall score. **H2.2:** Strong positive correlation ($r \geq 0.5$) of *general trust* subscale: *propensity to trust avatars* and the ITS overall score.
SITS	**H3.1:** Strong positive correlation ($r \geq 0.5$) of *specific trust* subscale *feeling secure with the avatar* with SITS subscale *emotional trust*. **H3.2:** Small positive correlation ($r \geq 0.1$) of the other ATS subscales and the SITS: *overall trust and reliableness* subscales.
VVIQ	**H4.1:** *Good visualizers* will have scores on the specific subscales that are independent from their vivid imaginary ability. **H4.2:** *Poor visualizers* will have scores on the specific subscales that are dependent on their vivid imaginary ability.
	Behavioral measurement instruments
Trust Game	**H5:** All *general trust* scales predict trusting behavior as trustor and trustee.
Ultimatum Game	**H6.1:** Proposer offers correlate positively with *specific trust: trustworthiness* of avatar subscale. **H6.2:** Proposer offers correlate positively with *specific trust: willingness to interact with the avatar* subscale. **H6.3** Receiver rejection threshold correlates negatively with *specific trust: trustworthiness of avatar* subscale. **H6.4** Receiver rejection threshold correlates negatively with *specific trust: willingness to interact with the avatar* subscale. **H6.5:** Proposer offers are independent of the *general trust* subscales. **H6.6:** Receivers' rejection thresholds are independent of the *general trust* subscales.
Dictator Game	**H7.1** Moderate positive correlation ($r > .3$) between the *specific trust: willingness to interact with the avatar* and the credits given to the avatar. **H7.2:** Moderate positive correlation ($r > .3$) between the *specific trust: trustworthiness of avatar* and the credits given to the avatar. **H7.3:** No correlation between the *general trust: propensity to trust humans* and the credits given to the avatar. **H7.4:** No correlation between the *general trust: propensity to trust avatars* and the credits given to the avatar.

Table 1. Hypotheses of the Present Study.

Internal Consistencies

Scale homogeneity results are presented in Table 2. All scales exceeded the assumed homogeneity threshold of α, $\omega > .75$, except for the avatar literacy scale. However, this scale consists of only three items, and the adjusted threshold of $> .6$ was successfully met (supporting Hypothesis H1.3). Further details on item-total correlations and leave-one-out homogeneity analysis are available in the Supplementary Material S3.

Intercorrelations of the Scales

The intercorrelations of the scales can be seen in Table 2. All intercorrelations were significant (all $ps \leq 0.003$). The three specific trust subscales correlated moderately to strongly with each other. Only the interrelations between the *specific trust* subscales *willingness to interact with the avatar* and *feeling secure with the avatar* were slightly lower than expected ($r \geq .30$). Correlations between the *specific* and the *general trust* subscales were all $r \geq .10$. However, correlations with *propensity to trust avatars* were higher than expected. These findings show mixed support H1.4.

The two *general trust* subscales, *propensity to trust avatars* and *propensity to trust humans*, were moderately correlated, whereas correlations involving *avatar literacy* were small. As mentioned above, correlations with *specific trust* subscales were larger than expected. Overall, these patterns partially support H1.5.

	STPTA	STWIA	STFSA	GTPTA	GTPTH	GTAL
STPTA	.910 / .912					
STWIA	.553	.874 / .876				
STFSA	.412	.246	.819 / .829			
GTPTA	.651	.385	.245	.855 / .856		
GTPTH	.280	.132	.353	.325	.788 / .794	
GTAL	.287	.217	.135	.262	.102	.603 / .610

Table 2. Intercorrelation and Internal Consistencies (Cronbach's α left / McDonald's ω right) of the Scales of the ATS. *Specific Trust*: Perceived Trustworthiness of the Avatar (STPTA), Willingness to Interact With the Avatar (STWIA), Feeling Secure with the Avatar (STFSA); *General Trust*: Propensity to Trust Avatars (GTPTA), Propensity to Trust Humans (GTPTH), Avatar Literacy (GTAL).

Validating Constructs

Interpersonal Trust Scale (ITS)

The correlations of the ATS subscales with all questionnaires can be found in Table 3 (for correlations with the Interpersonal Reactivity Index, see Supplementary Material S1). In line with H2.1, the *propensity to trust humans* showed a strong positive correlation with the ITS overall score ($r \geq .5$).

The *propensity to trust avatars* only had a small correlation with the ITS, falling short of the pre-specified threshold for a strong correlation ($r \geq .5$), thus H2.2 was not supported. All other ATS subscales showed small positive correlations ($r \geq .10$), except for *willingness to interact* and *avatar literacy* ($r < .10$).

Specific Interpersonal Trust Scale (SITS)

We found significant small positive correlations between the specific trust subscale *feeling secure with the avatar* and the SITS subscales emotional trust for males and for females (see Table 3) . We expected a strong positive correlation ($r \geq .5$), so our expectations could not be met to the extent we were looking for (H3.1 rejected).

For the other correlations between ATS and SITS subscales, we found a mixed pattern. While our assumptions were met for the *feeling secure with the human* subscale ($r \geq .1$), the correlation pattern was more mixed for all other subscales. Hence our expectations (H3.2) were only partially met.

	ITS	SITS: Male - Overall	SITS: Male - Emotional Trust	SITS: Male - Reliableness	SITS: Female - Emotional Trust	SITS: Female - Reliableness	VVIQ: Good Visualizers	VVIQ: Bad Visualizers
STPTA								
r	.228	.141	.210	.174	.085	.137	-.079	.149
p-value	< .001	.019	< .001	.004	.045	.001	.404	.281
STWIA								
r	.069	.082	.151	.049	.034	.060	.048	.154
p-value	.045	.170	.011	.411	.425	.157	.608	.267
STFSA								
r	.284	.288	.246	.177	.235	.275	.172	-.103
p-value	< .001	< .001	< .001	.003	< .001	< .001	.066	.460
GTPTA								
r	.246	-.027	.056	.016	.039	.076	-.069	-.008
p-value	< .001	.657	.350	.796	.361	.074	.465	.955
GTPTH								
r	.628	.247	.256	.221	.210	.218	-.043	-.143
p-value	< .001	< .001	< .001	< .001	< .001	< .001	.648	.303
GTAL								
r	.048	-.072	.036	-.066	-.139	-.117	-.089	.156
p-value	.165	.232	.544	.271	< .001	.005	.345	.259

Table 3. Correlation between the Avatar Trust Scale (ATS) subscales and Interpersonal Trust Scale (ITS), Specific Interpersonal Trust (SITS) subscales and Vividness of Visual Imagery Questionnnaire (VVIQ).

Vividness of Visual Imagery Questionnaire (VVIQ)

Table 3 shows the correlations between the ATS and the VVIQ for good and poor visualizers. In line with the cutoffs of Marks[32] we divided the groups into "good" and "bad" visualizers (mean values below 1.53 for good, above 3.15 for bad visualizers). This resulted in a sample size of $N = 115$ for good visualizers and $N = 54$ for bad visualizers. None of the relationships for good and bad visualizers became significant. As H4.1 is a nullhypothesis, we applied two one-sided tests (TOST) within the borders of -.1 to .1. Even though, none of the correlations was significant (see Table 3), TOST results were also not significant (see Supplemental Material S4 for all TOST results). Thus, H4.1 was not supported. For bad visualizers, several positive and one negative correlation were found with substantial relationships ($r \geq .1$). However, due to a small sample size in the bad visualizers group, none of the effects were significant. Thus, H4.2, was not supported.

Trust Game

Table 4 shows the correlations between the ATS and the Trust Game. Trust-game behavior showed only small associations with ATS subscales ($r \leq .23$). In the first round (TG1), all subscales except *perceived trustworthiness of the avatar* and *avatar literacy* were associated with slightly higher investments. In the second round (TG2), the correlations were similar. However, in the second round, *perceived trustworthiness of the human* did not show any significant relationships. Thus, mostly lacking significant effects for general trust in the trust game do not support H5.

Ultimatum Game

Table 4 summarizes the correlation between the ATS and the offers of the Ultimatum Game. There were only few meaningful correlations between ATS subscales and behavioral measures in the Ultimatum Game ($|r| \leq .15$). We did not find a meaningful correlation between *perceived trustworthiness of the avatar* and the proposer's offer, thus, H6.1.was not supported. In addition, we also did not find any meaningful correlation between proposer's offers and *willingness to interact* ($|r| \leq .10$). Thus, H6.2 was also not supported. H6.3 and H6.4 assumed meaningful correlations with *perceived trustworthiness* and *willigness to interact* which we did not find in both cases. Thus, H6.3 and H6.4 were not supported. As H6.5 and H6.6 are nullhypotheses, we, again, applied TOST and found supporting results for H6.5 and mixed support for H6.6 (see Supplementary Material S5 for TOST results).

	UG: Proposer's money kept	UG: Acceptance threshold receiver	UG: Acceptance 1 cent offer receiver	UG: Acceptance 3 cent offer receiver	UG: Acceptance 5 cent offer receiver	UG: Acceptance 7 cent offer receiver	DG: Proposer's money kept	TG1: Participant as trustor	TG2: Participant as trustee
STPTA									
r	-.021	-.074	-.022	-.112	-.057	-.049	-.033	-.136	-.110
p-value	.535	.033	.527	.001	.100	.158	.336	<.001	.001
STWIA									
r	-.090	-.003	.058	-.113	.016	.008	-.146	-.203	-.225
p-value	.009	.925	.091	.001	.635	.816	< .001	< .001	< .001
STFSA									
r	-.043	-.121	-.063	-.149	-.080	-.123	-.018	-.180	-.160
p-value	.216	< .001	.069	< .001	.020	< .001	.598	< .001	< .001
GTPTA									
r	-.025	-.021	-.024	-.011	-.012	-.021	-.027	-.042	-.034
p-value	.470	.543	.487	.760	.720	.550	.426	.225	.327
GTPTH							,		
r	-.013	-.094	-.064	-.098	-.068	-.068	-.025	-.114	-.051
p-value	.701	.006	.065	.004	.049	.050	.467	< .001	.143
GTAL									
r	-.052	-.005	-.045	.046	.052	.044	-.049	-.056	.005
p-value	.129	.896	.191	.179	.135	.200	.152	.107	.878

Table 4. Correlation of the Avatar Trust Scale (ATS) and the behavioral measures: Ultimatum Game (UG), Dictator Game (DG) and Trust Game (TG). Coding: UG (accept (1), reject (2)), DG (amount kept (1,3,5,7)), TG (pass/share credits (1), keep credits (2))

Dictator Game

Table 4 summarizes the correlation between the ATS, and the participant's offer in the Dictator Game. We expected moderate positive correlations between the number of credits given and the specific trust subscales *willingness to interact* and *perceived trustworthiness* (H7.1 and H7.2). The correlation for *perceived trustworthiness* was very close to zero ($r \approx 0$). For *willingness to interact* there was a small correlation, however, the size of the correlation was smaller than expected. Thus, both hypotheses were not supported. For H7.3 and H7.4 we applied TOST again. TOST results supported the lacking correlation between the *propensity to trust humans* and *propensity to trust avatars* and the Dictator Game results. Thus, the results of H7.3 and H7.4 were in line with our expectations.

Discussion

Our goal was to develop and validate a questionnaire as a trust measurement instrument for virtual avatars, including general as well as specific trust aspects. Overall, the proposed hierarchical structure could be found for the questionnaire. However, other than expected, "avatar literacy" did not contribute to general trust, but contributed to Virtual Trust overall (see Figure 1). Avatar literacy seems not to be related to general trust, as it appears to represent more of a skill than a disposition for trust. In addition, the structure could not be computed with all originally chosen items. We had to exclude two items from the *specific trust: perceived trustworthiness of the avatar scale* (see Supplementary Material S2 as well as section Preregistration and Changes). Yet, the wording of the questions appeared to be misleading and led to cross- or no loadings with other factors, for crossloadings especially general trust (see Supplementary Material). Once these items had been removed, the proposed confirmatory structure could be identified, emphasizing the importance of the specific virtual trust factor and general trust factor.

Concerning the homogeneity of the scales, all scales are rather homogeneous, except the avatar literacy scale (see also Supplementary Material S3). This scale only got three items that account for very different aspects of the construct of avatar literacy, so the heterogeneity could have been anticipated. Yet, a lack of homogeneity is only a problem in narrow constructs, and the criterion of Cronbach´s $\alpha > .7$ is only given for scales with six or more items[34]. Hence, this lack of homogeneity may not be a flaw in assessment quality per se.

For intercorrelations of the scales, we can contextualize the strong correlation between the *specific trustworthiness of the avatar* and the *general propensity to trust avatars* along with the *specific willingness to interact*. General trust in avatars may enhance the trustworthiness expectation and this in turn may influence the willingness to interact and vice versa[2,4]. The moderate correlations of the subscales further support that the different aspects of trust depict subdomains that are not entirely the same but can be integrated into different domains of avatar trust (see Figure 1). To assess these different aspects of trust, we included two standardized subjective measurements originally developed to measure interpersonal trust in humans (ITS,[3], SITS[15]). We found a large correlation of the subscale *general trust: propensity to trust humans* with the ITS, indicating that a similar construct is measured. However, Table 3 shows that general trust in humans and general trust in avatars (r = .33) appear to be two different constructs with different applications. This supports the difference between human trust and machine-related trust[7,35]. There are differences in interaction patterns, as well as the respective trust concepts, thus, the need for a questionnaire that is specifically developed to measure trust in virtual situations was supported.

For a behavioral validation, we used the trust game[12,36] to study trust behavior toward the virtual avatar. Contrary to our hypothesis that the general trust subscales of the ATS would be related to higher trusting behavior, the specific trust subscales showed stronger relationships. Yet, despite contrary expectations, these findings are plausible given the framing of the avatar in the economic games. People were instructed that they would be playing against the same avatar in all economic games. Thus, the finding is plausible when people actually believed that they were playing against the same avatar the whole time. In line with this, the finding is consistent with previous findings on trust, aimed at specific individuals[37]. Our finding that higher scores on specific trust subscales were associated with greater investment behavior demonstrates that the perceptions of avatar trustworthiness, security, and willingness to interact of participants align with their risk--taking in an economic exchange. Importantly, however, we do not view the ATS as a substitute for behavioral games. Trust games capture a specific aspect of trust, e.g. the willingness to take a monetary risk. The ATS allows researchers to measure perceptions of trust in avatars beyond economic decisions.

To disentangle more general behavior in economic paradigms from trust, the ultimatum game[12,36] and the dictator game[13,14] were chosen as divergent validation. As expected, general trust was independent from ultimatum game offers as proposers. However, specific trust in avatars was not consistently associated with significant changes in proposer behavior. For the receiver positions, the proposed independence from general trust aspects was not confirmed. Importantly, the general trust in humans was slightly correlated with a lower acceptance threshold. Taken together, these findings possibly indicate a common third variable, such as prosocial behavior, that might drive this acceptance and general trust in humans[6,12,38].

For the dictator game, participants with higher scores in the *willingness to interact with the avatar* subscale offered slightly higher amounts to the avatar. However, the association was lower than expected. In addition, we did not find a significant association between dictator game offers and *perceived trustworthiness*. However, independence of dictator game offers from general trust could be shown by TOSTER results.

Unexpectedly, the vividness of imagery was not related to any subscale scores of the ATS. However, sample sizes within the group of good and especially within the group of bad visualizers were small. Thus, non-significant correlations should be interpreted with caution. Nevertheless, we find this unexpected result promising. It suggests that the ATS may be used even for very short interactions with avatars. On a critical note, the questionnaire aimed to assess trust for real interactions with human-controlled avatars. It would have been desirable to have an immersive interaction with the avatar (e.g. in VR). However, such a procedure would have been challenging with a large sample size like ours. Many participants may have had limited or no prior experience with avatars, so we relied on their ability to imagine what such interactions would be like. Although they were provided with a clear definition of an avatar and instructed to perceive the virtual human as if it were controlled by a real person, it remains difficult to ensure consistent understanding of the context across participants.

Furthermore, the interaction was rather brief. Richer interactions could provide additional insight. However, our goal was to capture initial trust impressions. Previous research shows that people form judgments of traits such as trustworthiness from faces (within a fraction of a second, and these judgments are robust even when time is extended[39,40]). Based on these findings, our 10-second video offered more than enough exposure to elicit an initial trust impression while ensuring a standardized procedure. Future work should explore how the ATS performs in longer, more interactive, or immersive settings.

A further limitation is the lack of validation and measurement invariance testing for the English version, which was developed in parallel with the German version. Although this was not preregistered, we aimed to examine the structure and validity in an English-speaking sample. Due to insufficient data, this analysis could not be conducted, but future studies using the English version (see Appendix) may address this gap.

In this paper, we presented and validated a novel tool for the measurement of trust towards avatars. We observed that general trust measures correlate with the Interpersonal Trust Scale, and those for specific trust correlate with corresponding measures on the Specific Interpersonal Trust Scale. In terms of behavioral data, our scale showed sensitivity for the difference of trust in the trust game, while being mostly independent of investment in a divergent measure (the ultimatum game). In addition to some sensitivity to investment in the dictator game, this may indicate a variable common to specific and general trust as well as investment. Overall, we were able to create and validate a questionnaire measurement instrument for assessing trust in the context of avatars.

Methods

Ethical Statement

All subjects gave written informed consent in accordance with the Declaration of Helsinki before they participated in the study. The protocol was approved by the local ethics committee of the department of psychology of the Julius-Maximilians-University of Würzburg (GZEK 2022-06).

Preregistration and changes

The study was preregistered: https://osf.io/tdmq4. However, the current manuscript contains some changes to the original. First, a name change for the questionnaire was made from VR-SATS (Virtual Reality-–Avatar Trust Scale) to ATS (Avatar Trust Scale). Second, a subscale name change was made from "Knowledge of avatar nature" to "Avatar Literacy." Third, the proposed confirmatory factor analysis was not computed with all items that were intended to be used as indicators for the constructs, as the models did not converge due to Heywood cases occurring in the factor with nine items of *Specific Trust: Perceived trustworthiness of the avatar*.

To identify items that could cause this issue, an exploratory factor analysis was computed. The results showed that the indicator variables of all scales were as expected, except for items two and three of this scale (i.e., *I believe this avatar would act in my best interest* and *This avatar can be trusted in general*, respectively), where either no relevant loadings or a cross-loading to the general aspect of avatar trust was found. When excluded, the confirmatory analysis was computed using robust estimation. However, the second target model proposed in the preregistration could not be computed due to model non-identification. Therefore, we report results only for the 0-model, the basic model, and target model 1, which we will refer to as the *target model* throughout the remainder of this paper.

Participants

An a priori sample size calculation was performed using an expected effect of $r = .2$ as the desired effect size, with a statistical power of .95, nine latent variables, and 34 indicator variables. This led to a recommended sample size of 661[41]. Expecting a typical drop-out rate of 15%, we aimed to collect data from 780 participants in total. Our final sample encompassed 858 participants, of whom 840 agreed to data usage and were, thus, included in all analyses (279 male, mean age = 24.60, SD = 7.39). Please note that there were missing values for the Trust Game for two participants.

We restricted our sample in the data collection to the age range of 18—50 to ensure comparable familiarity with (digital) technology across all participants. Participants were recruited via university mailing lists and social media advertisements targeting adults in Germany. Most were university students or recent graduates. No course credit was required to participate; compensation consisted of course credit where applicable or entry into a lottery in which every tenth participant won 5 euros.

Procedure

After receiving an introduction to the study and completing demographic questions, all participants viewed the same female virtual human, shown in a short 10-second video clip waving its hand (Figure 2). The virtual human was not actually an avatar controlled by a real person but a pre-scripted animation. Participants were given a basic explanation of what avatars are and were instructed to imagine and interpret the virtual human as if it were an avatar controlled by another user in a virtual environment. All subjective and behavioral measures refer specifically to this brief interaction with this specific avatar. Participants then played three short economic games (the Trust Game, the Ultimatum Game, and the Dictator Game, see Figure 3). Following the games, they completed the questionnaires for the subjective measures (ATS, SITS, ITS, IRI, and VVIQ).

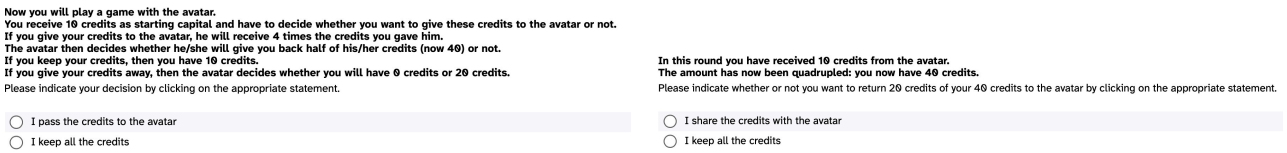
The virtual human used in the study (depicted in Figure 2) was selected through a pre-study, in which 50 participants (12 male and 38 female) were presented with 7 male and 7 female virtual humans[37]. Participants rated each virtual human's attractiveness and realism on a 7-point Likert scale. The female virtual human with medium attractiveness (M = 3.86, SD = 1.58) and high realism (M = 4.80, SD = 1.37) was selected for this study. We selected a single avatar appearance to maintain experimental control and reduce between--subject variability. Using multiple avatar appearances would have increased the number of conditions substantially (given our sample size) and complicated model estimation. The entire procedure, including initial instructions, gameplay, and completion of the questionnaires, took approximately 30 minutes.



Figure 2. The virtual human shown to participants.



(a) Ultimatum Game (left) and Dictator Game (right)



(b) Trust Game – Round 1

(c) Trust Game – Round 2

Figure 3. Questions presented to participants in the behavioral tasks. **a** Ultimatum and Dictator games (top row). **b, c** Trust Game rounds 1 and 2 (bottom row). Participants were instructed to respond as if they were interacting with a real player controlling the avatar (representation in Fig. 2).

Instruments

Avatar Trust Scale (ATS)

The ATS was constructed using a seven-step procedure: 1) Identification of the domain for the construct; 2) Item generation through brainstorming, focus group discussions, and review of the existing literature; 3) Categorization of the items into determinants; 4) Expert validation; 5) Scale modification, refinement, and finalization of the ATS; 6) First data collection; 7) Exploratory factor analysis (EFA), item-retest correlation, index of difficulty, and Lienert-item selection.

In the first draft, 155 items were generated and reduced to 106 after expert consideration in steps 4 and 5. After the factor analysis and Lienert-item selection (with a sample of 310 participants for the German version and 94 for the English version), 34 items resulted in six subfactors of the questionnaire: Specific trust: Perceived trustworthiness of the avatar (STPTA)[2,4];

Specific trust: Willingness to interact with the avatar (STWIA)[2,4]; Specific trust: Feeling secure with the avatar (STFSA)[15]; General trust: Propensity to trust avatars (GTPTA)[3,15]; General trust: Propensity to trust humans (GTPTH)[3,15]; General trust: Avatar literacy (GTAL).

All items included in this study can be found in the Supplementary Material. The final version of the questionnaire, including instructions and response scale, is provided in the Appendix. For further details regarding the item generation phase, see Supplemental Material S7.

Interpersonal Trust Scale (ITS)

The ITS[31] is a self--report measure for trust in other people. The ITS includes items referring to various social subjects such as politicians, teachers, physicians, parents, friends, and classmates. It consists of 25 items measuring trust and 15 filler items included to obscure the intent of the questionnaire[3]. Answers are indicated on a 5-point Likert scale. We used the German version by Amelang and Bartussek consisting of 27 items[42].

Specific Interpersonal Trust Scale (SITS)

The SITS measures interpersonal trust in meaningful relationships using a 9-point Likert scale ranging from "strongly disagree" to "strongly agree". Two versions exist: a male version with 21 items covering overall trust, emotional trust, reliability, and dependability, and a female version with 13 items addressing emotional trust, reliability, and physical trust[15].

Johnson and George developed separate versions to reflect gender-specific patterns in trust formation, as men and women tend to differ in how they perceive and respond to trust-related situations. These differences are supported by behavioral and theoretical evidence suggesting that women are generally more risk-sensitive in social contexts, while men are more inclined toward risk-taking[43,44].

Interpersonal Reactivity Index (IRI)

The Interpersonal Reactivity Index (IRI)[45] measures empathy across four subscales: fantasy, perspective-taking, personal distress, and empathic concern. We used the German version Saarbrücker Persönlichkeitsfragebogen including 16 items answered on a 5-point Likert scale ranging from "Never" to "Always". Results from this section are available in the Supplementary Material[46].

Vividness of Visual Imagery Questionnaire (VVIQ)

The Vividness of Visual Imagery Questionnaire (VVIQ)[32] measures the vividness of mental imagery with 16 items. Items are rated on a 5-point Likert scale. We used the German version by Jungmann et al.[47]

Trust Game

The trust game captures willingness to risk resources in an economic exchange and has been widely used in virtual--reality research to study trust behavior[25,37,48]. Our goal was to assess convergent validity: by comparing participants' investments and reciprocity in the game with their scores on the ATS, we could determine whether higher perceived trust correlates with greater trusting behavior.

Participants played a two--round, one--shot trust game following Berg et al.[5] (see Figure 3b and 3c). In Round 1, as trustors, participants received 10 credits and they could chose whether to keep the credits or send them to the trustee (the "avatar"). Any amount sent was quadrupled. In Round 2, the roles were reversed: participants received 40 credits (reflecting a full transfer by the trustor) and decided whether to return 20 credits to the avatar or keep all 40. Participants were instructed to imagine that they were interacting with the person represented by the avatar and that their decisions would affect both parties' payoffs. Although payoffs were hypothetical, we emphasized that choices should reflect their genuine willingness to trust and reciprocate.

Each decision was made independently and without feedback between rounds, in line with the standard one-shot trust game protocol. This structure captures both trust behavior (in the trustor role) and reciprocity (in the trustee role).

Ultimatum Game

In this game of bargaining[49], participants alternated roles as proposer and receiver. Proposers decided how to split 8 credits with the avatar (offers of 1, 3, 5 or 7 credits were allowed, see Figure 3a). The receivers were presented with each possible offer and indicated whether they would accept or reject it, knowing that rejection meant that both parties received zero. In our study, participants played the paradigm from both perspectives. We measured trust as proposer in terms of generosity (higher offers) and rejection thresholds (willingness to accept lower offers). Although this paradigm has not traditionally been associated with trust behavior, it serves as a useful divergent validation method[50].

Dictator Game

The Dictator Game is a variant of the Ultimatum Game[49], where the receiver cannot respond to the offer made by the proposer. In our implementation, participants played only as proposers and decided how much of 8 credits they wanted to give to the avatar (see Figure 3a). This game was used as a divergent validation measure, as it reflects generosity driven by consideration for others rather than strategic trust[50].

Statistics

Structural equation modeling was performed using the *lavaan* software package[51,52] to evaluate the structural hypotheses of the ATS. The robust estimator Weighted Least Square Mean and Variance Adjusted (WLSMV) was selected for all models. The following fit indices were reported for model comparison: Root Mean Square Error of Approximation (RMSEA; indicating good fit if $< .06$), Comparative Fit Index (CFI; good fit if $> .95$), Tucker-Lewis Index (TLI; good fit if $> .95$), and Standardized Root Mean Square Residual (SRMR; good fit if $< .08$)[53–55].

The evaluated models included the 0--model (an independence model with no subfactors), the basic model, and the target model (see Figure 1). Additionally, exploratory factor analysis (EFA) was computed to identify non-fitting indicator variables. The Kaiser-Meyer-Olkin (KMO) measure was .931, indicating that the dataset was highly suitable for factor analysis. Factor extraction was performed using the method of minimum residuals with oblimin rotation.

To assess scale reliability, McDonald's omega and Cronbach's alpha were calculated. High internal consistency was assumed for subscales with six or more items if values exceeded .7[56].

For construct validation, Pearson's correlations were computed for all validating constructs. A Bonferroni correction was applied for multiple comparisons, adjusting the critical p-value to .002 (based on 22 comparisons). For hypotheses that posit independence, no alpha correction was applied. Instead, a 90% confidence interval was used to assess whether zero was included in the interval around the mean correlation and TOSTER results are reported in the Supplemental Material S4- S6.

All analyses, except structural equation modeling, were conducted using *Jamovi*[57].

References

1. Balliet, D. & Van Lange, P. A. M. Trust, conflict, and cooperation: A meta-analysis. *Psychol. Bull.* **139**, 1090–1112, DOI: 10.1037/a0030939 (2013).
2. Thielmann, I. & Hilbig, B. E. Trust: An Integrative Review from a Person–Situation Perspective. *Rev. Gen. Psychol.* **19**, 249–277, DOI: 10.1037/gpr0000046 (2015).
3. Rotter, J. B. A new scale for the measurement of interpersonal trust. *J. Pers.* **35**, 651–665, DOI: 10.1111/j.1467-6494.1967.tb01454.x (1967).
4. Thielmann, I. & Hilbig, B. E. Trust in me, trust in you: A social projection account of the link between personality, cooperativeness, and trustworthiness expectations. *J. Res. Pers.* **50**, 61–65, DOI: 10.1016/j.jrp.2014.03.006 (2014).
5. Berg, J., Dickhaut, J. & McCabe, K. Trust, Reciprocity, and Social History. *Games Econ. Behav.* **10**, 122–142, DOI: 10.1006/game.1995.1027 (1995).
6. Müller, S., Schwehm, S., Hewig, J. & Paelecke, M. Influence of Trust-Enhancing Signals on Signal Processing in B2C E-Commerce. In *AMCIS 2020 Proceedings* (2020).
7. Hancock, P. A. *et al.* A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Hum. Factors: The J. Hum. Factors Ergonomics Soc.* **53**, 517–527, DOI: 10.1177/0018720811417254 (2011).
8. Bombari, D., Schmid Mast, M., Canadas, E. & Bachmann, M. Studying social interactions through immersive virtual environment technology: Virtues, pitfalls, and future challenges. *Front. Psychol.* **6**, DOI: 10.3389/fpsyg.2015.00869 (2015).
9. Oyekoya, O., Steptoe, W. & Steed, A. SphereAvatar: A situated display to represent a remote collaborator. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2551–2560, DOI: 10.1145/2207676.2208642 (ACM, 2012).
10. Steptoe, W., Steed, A. & Slater, M. Human Tails: Ownership and Control of Extended Humanoid Avatars. *IEEE Transactions on Vis. Comput. Graph.* **19**, 583–590, DOI: 10.1109/TVCG.2013.32 (2013).
11. Lugrin, B. *et al.* Challenge Discussion on Socially Interactive Agents: Considerations on Social Interaction, Computational Architectures, Evaluation, and Ethics. In Lugrin, B., Pelachaud, C. & Traum, D. (eds.) *The Handbook on Socially Interactive Agents*, 561–626, DOI: 10.1145/3563659.3563677 (ACM, 2022), 1 edn.
12. Yamagishi, T. *Trust: The Evolutionary Game of Mind and Society*. The Science of the Mind (Springer Tokyo, 2011).

13. Couch, L. L. & Jones, W. H. Measuring Levels of Trust. *J. Res. Pers.* **31**, 319–336, DOI: 10.1006/jrpe.1997.2186 (1997).

14. Yamagishi, T. & Yamagishi, M. Trust and commitment in the United States and Japan. *Motiv. Emot.* **18**, 129–166, DOI: 10.1007/BF02249397 (1994).

15. Johnson-George, C. & Swap, W. C. Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other. *J. Pers. Soc. Psychol.* **43**, 1306–1317, DOI: 10.1037/0022-3514.43.6.1306 (1982).

16. Dwivedi, Y. K. *et al.* Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *Int. J. Inf. Manag.* **66**, 102542, DOI: 10.1016/j.ijinfomgt.2022.102542 (2022).

17. Evans, A. M. & Revelle, W. Survey and behavioral measurements of interpersonal trust. *J. Res. Pers.* **42**, 1585–1593, DOI: 10.1016/j.jrp.2008.07.011 (2008).

18. Lewicki, R. J., Tomlinson, E. C. & Gillespie, N. Models of Interpersonal Trust Development: Theoretical Approaches, Empirical Evidence, and Future Directions. *J. Manag.* **32**, 991–1022, DOI: 10.1177/0149206306294405 (2006).

19. Riedl, R., Mohr, P. N. C., Kenning, P. H., Davis, F. D. & Heekeren, H. R. Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory. *J. Manag. Inf. Syst.* **30**, 83–114, DOI: 10.2753/MIS0742-1222300404 (2014).

20. Halbig, A. *et al.* Opportunities and Challenges of Virtual Reality in Healthcare – A Domain Experts Inquiry. *Front. Virtual Real.* **3**, 837616, DOI: 10.3389/frvir.2022.837616 (2022).

21. Van Rijn, B., Cooper, M. & Chryssafidou, E. Avatar-based counselling for young people within school counselling. Qualitative analysis of client experience. *Couns. Psychother. Res.* **18**, 59–70, DOI: 10.1002/capr.12155 (2018).

22. Lin, J. & Latoschik, M. E. Digital body, identity and privacy in social virtual reality: A systematic review. *Front. Virtual Real.* **3**, 974652, DOI: 10.3389/frvir.2022.974652 (2022).

23. Pan, Y. & Steed, A. The impact of self-avatars on trust and collaboration in shared virtual environments. *PLOS ONE* **12**, e0189078, DOI: 10.1371/journal.pone.0189078 (2017).

24. Pan, Y. & Steed, A. A Comparison of Avatar-, Video-, and Robot-Mediated Interaction on Users' Trust in Expertise. *Front. Robotics AI* **3**, DOI: 10.3389/frobt.2016.00012 (2016).

25. George, C., Eiband, M., Hufnagel, M. & Hussmann, H. Trusting strangers in immersive virtual reality. In *Companion Proceedings of the 23rd International Conference on Intelligent User Interfaces*, 1–2 (2018).

26. Philip, P. *et al.* Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients. *npj Digit. Medicine* **3**, 2, DOI: 10.1038/s41746-019-0213-y (2020).

27. Sun, N. & Botev, J. Intelligent autonomous agents and trust in virtual reality. *Comput. Hum. Behav. Reports* **4**, 100146, DOI: 10.1016/j.chbr.2021.100146 (2021).

28. Lee, J. J. & Dryjanska, L. Avatars and Digital Technology Literacy Applied in Psychology. In *Handbook of Research on Media Literacy Research and Applications Across Disciplines*, 402–420, DOI: 10.4018/978-1-5225-9261-7.ch023 (IGI Global, 2019).

29. Hong, T., Cabrera, J. & Beaudoin, C. E. Disentangling real-world and virtual-world social norms: The persuasive elements and social psychological effects of a serious game. *Telematics Informatics Reports* **9**, 100038, DOI: 10.1016/j.teler.2022.100038 (2023).

30. Pałosz, R. How do base rules constitute a virtual world? *Int. J. for Semiot. Law - Revue internationale de Sémiotique juridique* **35**, 1879–1901, DOI: 10.1007/s11196-022-09903-8 (2022).

31. Rotter, J. B. Interpersonal trust scale. *J. Pers.* (1967).

32. Marks, D. F. Visual imagery differences in the recall of pictures. *Br. J. Psychol.* **64**, 17–24, DOI: 10.1111/j.2044-8295.1973.tb01322.x (1973).

33. Parks, C. D., Joireman, J. & Van Lange, P. A. M. Cooperation, Trust, and Antagonism: How Public Goods Are Promoted. *Psychol. Sci. Public Interest* **14**, 119–165, DOI: 10.1177/1529100612474436 (2013).

34. Wang, W. Difference between the Real World and Virtual World. In *IS4SI 2019 Summit*, 35, DOI: 10.3390/proceedings2020047035 (MDPI, 2020).

35. Kaplan, A. D., Kessler, T. T., Brill, J. C. & Hancock, P. A. Trust in Artificial Intelligence: Meta-analytic findings. *Hum. Factors: The J. Hum. Factors Ergonomics Soc.* **65**, 337–359, DOI: 10.1177/00187208211013988 (2023).

36. Carlo, G., Randall, B. A., Rotenberg, K. J. & Armenta, B. E. A friend in need is a friend indeed: Exploring the relations among trust beliefs, prosocial tendencies, and friendships. In Rotenberg, K. J. (ed.) *Interpersonal Trust during Childhood and Adolescence*, 270–294, DOI: 10.1017/CBO9780511750946.013 (Cambridge University Press, 2010), 1 edn.

37. Lin, J., Cronjé, J., Käthner, I., Pauli, P. & Latoschik, M. E. Measuring Interpersonal Trust towards Virtual Humans with a Virtual Maze Paradigm. *IEEE Transactions on Vis. Comput. Graph.* **29**, 2401–2411, DOI: 10.1109/TVCG.2023.3247095 (2023).

38. Siegrist, M., Gutscher, H. & Earle, T. C. Perception of risk: The influence of general trust, and general confidence. *J. Risk Res.* **8**, 145–156, DOI: 10.1080/1366987032000105315 (2005).

39. Willis, J. & Todorov, A. First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychol. Sci.* **17**, 592–598, DOI: 10.1111/j.1467-9280.2006.01750.x (2006).

40. Zebrowitz, L. A. First Impressions From Faces. *Curr. Dir. Psychol. Sci.* **26**, 237–242, DOI: 10.1177/0963721416683996 (2017).

41. Soper, D. Free A-priori Sample Size Calculator for Structural Equation Models - Free Statistics Calculators (2022).

42. Amelang, M. & Bartussek, D. Zwischenmenschliches vertrauen. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)* DOI: 10.6102/ZIS127 (1997).

43. Wu, Y., Hall, A. S. M., Siehl, S., Grafman, J. & Krueger, F. Neural Signatures of Gender Differences in Interpersonal Trust. *Front. Hum. Neurosci.* **14**, 225, DOI: 10.3389/fnhum.2020.00225 (2020).

44. Macko, A. Gender Differences In Trust, Reactions To Trust Violation, And Trust Restoration. *Decyzje* **2020**, DOI: 10.7206/DEC.1733-0092.140 (2020).

45. Davis, M. H. Interpersonal Reactivity Index, DOI: 10.1037/t01093-000 (2011).

46. Paulus, C. Der saarbrücker persönlichkeitsfragebogen spf(iri) zur messung von empathie: Psychometrische evaluation der deutschen version des interpersonal reactivity index, DOI: 10.23668/PSYCHARCHIVES.9249 (2009).

47. Jungmann, S. M., Becker, F. & Witthöft, M. Erfassung der lebendigkeit mentaler vorstellungsbilder: Adaptation und validierung deutschsprachiger versionen des vividness of visual imagery questionnaire (vviq) und des plymouth sensory imagery questionnaire (psi-q). *Diagnostica* **68**, 125–136, DOI: 10.1026/0012-1924/a000291 (2022-07).

48. Bolton, G. E., Katok, E. & Ockenfels, A. Trust among Internet Traders: A Behavioral Economics Approach. *Analyse & Kritik* **26**, 185–202, DOI: 10.1515/auk-2004-0110 (2004).

49. Güth, W., Schmittberger, R. & Schwarze, B. An experimental analysis of ultimatum bargaining. *J. Econ. Behav. & Organ.* **3**, 367–388, DOI: 10.1016/0167-2681(82)90011-7 (1982).

50. Güth, W. On ultimatum bargaining experiments — A personal review. *J. Econ. Behav. & Organ.* **27**, 329–344, DOI: 10.1016/0167-2681(94)00071-L (1995).

51. R Core Team. R: A language and environment for statistical computing (2021).

52. Rosseel, Y. Lavaan: An R package for structural equation modeling. *J. Stat. Softw.* **48**, 1–20 (2012).

53. Bentler, P. M. Comparative fit indexes in structural models. *Psychol. Bull.* **107**, 238–246, DOI: 10.1037/0033-2909.107.2.238 (1990).

54. Hu, L. & Bentler, P. M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model. A Multidiscip. J.* **6**, 1–55, DOI: 10.1080/10705519909540118 (1999).

55. Kline, R. B. *Principles and Practice of Structural Equation Modeling*. Methodology in the Social Sciences (Guilford Press, New York, 2011), 3 edn.

56. Ponterotto, J. G. & Ruckdeschel, D. E. An Overview of Coefficient Alpha and a Reliability Matrix for Estimating Adequacy of Internal Consistency Coefficients with Psychological Research Measures. *Percept. Mot. Ski.* **105**, 997–1014, DOI: 10.2466/pms.105.3.997-1014 (2007).

57. The jamovi project. (version 2.7.37.0) [computer software] (2026).

Acknowledgements

This work is part of the "Privacy Matters" (PriMa) project. The PriMa project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860315, and the Faculty of Human Sciences, Julius-Maximilians Universität Würzburg, Germany. MFC was funded by a departmental studentship as part of the Psychology department at King's College London. Lastly, the European regional development fund 1.2-StMWK-F.4-UFR-002 [Europäischer Fonds für regionale Entwicklung (EFRE)] contributed to the funding of this project.

Author contributions statement

J.S.: Analysis, Writing, Review, Editing; J.C.: Conceptualization, Investigation, Writing, Review, Editing; J.L.: Creation of virtual human animations, Conceptualization; M.F.C.: Investigation, Writing, Review and Editing, Funding Acquisition; J.F.: Conceptualization, Investigation, Analysis; G.B.: Review, Editing; I.K.: Writing, Review, Editing; M.E.L.: Writing, Review, Editing, Funding Acquisition; J.R.: Conceptualization, Investigation, Writing, Review, Editing, Scientific Methods, Analysis, Funding information.

Additional information

Competing interests. The authors declare no competing interests. **Data availability**. The data is available from OSF at this link: https://osf.io/tz693/overview

Appendix: Items of the Avatar Trust Scale (ATS)

Instructions for participants: Below you will find a series of statements about attitudes and behavior. Please indicate the extent to which you agree or disagree with each statement by selecting one response option per item. There are no right or wrong answers. Respond to each statement independently, as if it were the only one. It is not necessary to be consistent across items. Try to answer all statements honestly and accurately, based on your personal feelings. Avoid leaving items blank and do not spend too much time on any single statement.

Response scale: 5-point Likert scale (Strongly disagree to Strongly agree).

Table 5. Items of the ATS

ID	German Version	English Version	Item
Subscale: Specific Trust Willingness to Interact with the Avatar (STWIA)			
STWIA1	Ich möchte mit diesem Avatar zusammen eine Aufgabe absolvieren.	I want to complete a task together with this avatar.	1
STWIA2	Ich möchte, dass dieser Avatar mein Teamkollege wird.	I want this avatar to become my teammate.	4
STWIA3	Ich erlaube diesem Avatar, mir in der virtuellen Welt zu folgen.	I allow this avatar to follow me in the virtual world.	10
STWIA4	Ich würde diesen Avatar zu einer virtuellen Veranstaltung von mir einladen.	I would invite this avatar to a virtual event of mine.	14
STWIA5	Ich würde mit diesem Avatar zu einer virtuellen Aktivität/Veranstaltung gehen.	I would go to a virtual activity/event with this avatar.	17
STWIA6	Ich möchte mit diesem Avatar zusammenarbeiten.	I want to cooperate with this avatar.	22
Subscale: Specific Trust Feeling Secure with the Avatar (STFSA)			
STFSA1*	Ich habe Angst vor diesem Avatar.	I am afraid of this avatar.	2
STFSA2*	Dieser Avatar wird mir gegenüber aggressiv sein.	This avatar will be aggressive towards me.	7
STFSA3*	Ich muss mich gegen diesen Avatar verteidigen.	I must defend myself against this avatar.	13
STFSA4*	Dieser Avatar wird mich verraten.	This avatar will betray me.	18
STFSA5*	Dieser Avatar wird mir schaden.	This avatar will harm me.	24
STFSA6*	Dieser Avatar wird mich bei einer Aufgabe ausnutzen.	This avatar will take advantage of me during a task.	32
Subscale: General Trust Propensity to Trust Humans (GTPTH)			
GTPTH1*	Es gibt viele Heuchler in dieser Gesellschaft.	There are many hypocrites in this society.	3
GTPTH2	Den meisten Menschen kann man vertrauen.	Most people can be trusted.	5
GTPTH3*	Nur wenige Personen sind vertrauenswürdig.	Only few people are trustworthy.	12
GTPTH4	Es ist leicht, Personen zu vertrauen.	It is easy to trust people.	27
GTPTH5*	Wenn ich mit Fremden zu tun habe, muss ich vorsichtig sein, bis sie sich als vertrauenswürdig erweisen.	When dealing with strangers, I need to be cautious until they prove to be trustworthy.	30
Subscale: General Trust Propensity to Trust Avatars (GTPTA)			

Table 5. Items of the ATS

ID	German Version	English Version	Item
GTPTA1	Die meisten Avatare sind vertrauenswürdig.	Most avatars are trustworthy.	8
GTPTA2	Die meisten Avatare sind grundsätzlich gut und freundlich.	Most avatars are basically good and kind.	9
GTPTA3	Avatare sind im Allgemeinen ehrlich.	Avatars are basically honest.	11
GTPTA4	Avataren kann im Allgemeinen vertraut werden.	Avatars can be trusted in general.	16
GTPTA5	Avatare sind im Allgemeinen zuverlässig.	Avatars are reliable in general.	21
Subscale: General Trust Avatar Literacy (GTAL)			
GTAL1	Ich kann spüren, ob ein Avatar vertrauenswürdig ist oder nicht.	I can sense whether an avatar is trustworthy or not.	23
GTAL2*	Es ist schwer für mich, zwischen einem vertrauenswürdigen und einem nicht vertrauenswürdigen Avatar zu unterscheiden.	It is hard for me to distinguish between a trustworthy and an untrustworthy avatar.	28
GTAL3*	Ich bin mir nicht sicher, ob dieser Avatar vertrauenswürdig ist.	I feel unsure about this avatar's trustworthiness.	31
Subscale: Specific Trust Perceived Trustworthiness of the Avatar (STPTA)			
STPTA1	Dieser Avatar hat im Allgemeinen gute Absichten.	This avatar has good intentions in general.	6
STPTA2	Dieser Avatar ist hilfsbereit.	This avatar is supportive.	15
STPTA3	Dieser Avatar ist fair.	This avatar is fair.	19
STPTA4	Dieser Avatar ist ehrlich.	This avatar is honest.	20
STPTA5	Dieser Avatar hat mir gegenüber gute Absichten.	This avatar has good intentions towards me.	25
STPTA6	Dieser Avatar ist im Allgemeinen zuverlässig.	This avatar is generally reliable.	26
STPTA7	Dieser Avatar ist vertrauenswürdig.	This avatar is trustworthy.	29

Note: (*) indicates inverted items. Items are grouped by subscale.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1InterpersonalReactivityIndex.docx
- S2ExploratoryFactorAnalysis.xlsx
- S4TOSTGoodVisualizers.xlsx
- S3Furtherreliabilitystatistics.xlsx
- S6TOSTDictatorGame.xlsx
- S5TOSTUltimatumGame.xlsx
- S7Questionairedevelopment.docx