Communications Psychology

https://doi.org/10.1038/s44271-025-00391-7

Article in Press

AI outperforms humans in establishing interpersonal closeness in emotionally engaging interactions, but only when labelled as human

Received: 17 June 2025

Accepted: 18 December 2025

Cite this article as: Kleinert, T., Waldschütz, M., Blau, J. *et al.* AI outperforms humans in establishing interpersonal closeness in emotionally engaging interactions, but only when labelled as human. *Commun Psychol* (2026). https://doi.org/10.1038/s44271-025-00391-7

Tobias Kleinert, Marie Waldschütz, Julian Blau, Markus Heinrichs & Bastian Schiller

AI outperforms humans in establishing interpersonal closeness in emotionally engaging interactions, but only when labelled as human

Tobias Kleinert[1]*, Marie Waldschütz[1], Julian Blau[1], Markus Heinrichs[1]*, Bastian Schiller[1,2]*

[1] Laboratory for Biological Psychology, Clinical Psychology, and Psychotherapy, Albert-Ludwigs University of Freiburg, Stefan-Meier-Straße 8, 79104 Freiburg, Germany

[2] Laboratory for Clinical Neuropsychology, Department of Psychology, Heidelberg University, Hauptstr. 47-51, 69117 Heidelberg, Germany

* Corresponding authors

Abstract

With the increasing accessibility of large language models to the public, questions arise about whether, and under what conditions, social-emotional interactions with artificial intelligence (AI) can lead to human-like relationship building. Across two double-blind randomised controlled studies with pre-registered analyses, 492 participants engaged in dyadic online interactions using a modified, text-based version of the 'Fast Friends Procedure' (a method designed to enable rapid relationship building), with pre-generated responses by either human partners or a minimally prompted large language model. When labelled as human, the AI outperformed human partners in establishing feelings of closeness during emotionally engaging 'deep-talk' interactions. This striking effect appears to stem from the AI's higher levels of self-disclosure, which in turn enhanced participants' perceptions of closeness. Labelling the partner as an AI reduced, but did not eliminate, relationship building, likely due to participants' lower motivation to engage in interactions with an AI, reflected in both shorter responses and reduced feelings of closeness. These findings highlight AI's potential to relieve overburdened social fields while underscoring the urgent need for ethical safeguards to prevent its misuse in fostering deceptive social connections.

Introduction

Fuelled by technological breakthroughs in neural network modelling and the rapidly advancing development and accessibility of large language models (LLMs), our experiences and evaluations of social interactions with artificial intelligence (AI) have fundamentally changed in recent years [1–6]. Indeed, growing evidence suggests that LLM-generated content can facilitate communication that not only feels similar to interaction with human agents but, in certain aspects and contexts, may even surpass it. For example, participants tend to evaluate LLM-generated content as more empathic than human-generated content, particularly when they are unaware of its non-human origin [7–10]. This has motivated research on the potential translational value of such findings, reporting beneficial effects of LLM-generated communication in therapeutic contexts [11,12]. These findings show that LLMs can be used to generate human-like responses during social communication and first pioneering evidence also suggests that some degree of human-AI relationship formation is possible [13–15]. However, it remains an open question whether, and under what conditions, humans build relationships with AI to the same extent as with other humans, especially in the early stages of building a new relationship to a previously unknown other. The present study aims to fill this research gap by investigating differences in relationship building between initial interactions with humans versus AI (i.e., LLM-generated content).

Driven by the rapid evolution of AI's communicative abilities, theoretical accounts of human-AI interaction are being refined. Early accounts explained social behaviour towards computer-mediated technology as the 'mindless' application of social heuristics (e.g., politeness, stereotyping) to encounters with machines that display interactive features, such as speech (e.g., Social Response Theory and the associated Computers-Are-Social-Actors paradigm [16,17]). With the emergence of more advanced AI technologies approaching human-level communication, existing theories of human interpersonal relationships (e.g., Social Penetration Theory and Social Exchange Theory [18–20]) have been applied to human-AI interactions. However, AI still lacks certain human characteristics such as agency, autonomy, memory, and a personal history. Therefore, some have questioned the applicability of human relationship theories to human-AI interactions, emphasising the need for new theoretical frameworks grounded in empirical evidence on the nature of these relationships [14].

An argument frequently raised in public media discussions about the differences in communicative abilities between humans and AI is that AI-generated content may be at the level of human-generated content in some domains, but not in the 'uniquely human' domains such as emotion [21–24]. Fittingly, research shows that human interaction partners are preferred over AI ones, particularly in domains involving emotion [25,26]. These caveats contrast with evidence from the field of 'emotional AI', which suggest that LLMs possess a marked ability to recognise and respond appropriately to emotions [27–30]. Thus, when trying to understand the differences in relationship building between humans and AI, it is essential to consider the emotional intensity of the communication content. In other words, are there differences in AI's ability to engage in superficial small-talk versus deep-talk communication on more personally meaningful and emotionally charged topics that require a higher level of self-disclosure [31]?

Existing theories of human-AI interaction also fail to consider the impact of attitudes towards AI. As AI's social capabilities continue to evolve, concerns and reservations about interactive technological devices are growing in parallel. Indeed, many hold negative attitudes towards AI, perceiving it as unnatural and sometimes even threatening [32,33]. This scepticism is reflected in intense media debates about AI potentially replacing humans' unique socio-emotional and cognitive abilities [34], creating a paradox: Although communicating with AI can foster relationship building, once people realise they are interacting with AI, these effects seem to dissipate. For example, the perceived superiority of AI-generated responses over human-generated responses to the description of an emotional situation was reversed once participants

realised that the response came from an AI [7]. When considering possible applications of AI in clinical and psychotherapeutic settings, such negative attitudes towards machine interaction become problematic, as the source of communication must be ethically revealed. Therefore, it is crucial to understand how labelling the communication source (as AI or human) influences relationship building.

The present pre-registered research consisting of two double-blind randomised controlled studies examines relationship building in social-emotional interactions with AI and humans (source human vs. source AI), and its modulation by the emotional intensity of the interaction (small-talk vs. deep-talk; study 1) and the labelling of the source of the content (label human vs. label AI; study 2; for details on specific hypotheses, see pre-registration available at https://osf.io/chdx7). Both studies were conducted online simultaneously and included some shared data to enhance comparability between studies, prevent participants from taking part in both studies, and avoid history effects, which are likely in a rapidly evolving field such as AI interactions. History effects are important to consider, as participants' attitudes towards, and interactions with AI can change significantly over time [35], potentially leading to observed differences in AI interactions that reflect broader societal shifts rather than effects of experimental conditions. To mimic everyday relationship building, 492 participants engaged in a 15-minute online communication task (i.e., the Fast Friends Procedure, or FFP) designed to induce interpersonal closeness to an unfamiliar interaction partner through escalating mutual self-disclosure in a series of turn-taking questions [36–38]. We specifically selected the FFP because it was designed to enable the rapid development of interpersonal closeness in the early stages of relationship building between previously unacquainted partners [36], which was the focus of our study. Unbeknownst to participants, responses to FFP items were pre-generated either by a minimally prompted LLM or by human interaction partners who performed the FFP in a laboratory environment (for all items and responses, see Supplementary Table 1). We prompted the AI to respond from the perspective of fictional characters rather than in its original form to enable it to answer personal questions and keep basic character information (name, age, place of residence, field of study) consistent with human partners. Relationship building was operationalised by self-reports of perceived interpersonal closeness to the interaction partner. Furthermore, we applied exploratory automated linguistic analysis using the Linguistic Inquiry and Word Count system (LIWC [39,40]) to investigate whether any identified differences in relationship building between conditions could be explained by variations in self-disclosure levels of interaction partners (i.e., AI-generated characters or humans) and/or participants. In the LIWC, self-disclosure is operationalised as the number of words related to the self, emotions, and social processes [41]. We also analysed the response length of interaction partners and participants as a measure of social motivation, with longer responses indicating greater motivation.

In summary, the hypotheses we tested in study 1 were: Hypothesis 1: We expect that deep talk interactions with an AI will lead to a significant increase in interpersonal closeness compared to a baseline measure. This hypothesis is being tested to validate findings suggesting that relationship-building with AI is possible [7]. We additionally examine whether small-talk interactions will also generate increases in closeness. Hypothesis 2: We expect higher interpersonal closeness towards the interaction partner after interactions with humans compared to interactions with an AI across both small-talk and deep-talk interactions (main effect of the factor 'source identity'). This assumption builds on the traditional view that interactions with humans should elicit greater closeness than interactions with AI, as social interactions are fundamentally rooted in human behaviour [24]. Hypothesis 3: We expect higher interpersonal closeness towards the interaction partner after deep-talk interactions compared to small-talk interactions across interactions with humans and AI (main effect of the factor 'emotional intensity'). This hypothesis relies on the idea that self-disclosure on emotional topics is a key driver of early relationship-

building [36]. Hypothesis 4: We expect that the differences (human > AI) regarding interpersonal closeness towards the interaction partner are larger after deep talk interactions compared to small talk interactions (interaction effect between the factors 'emotional intensity' and 'source identity'). This hypothesis assumes that AI may adequately mimic non-personal small-talk but not personal deep-talk, as emotions are widely considered a uniquely human domain [22], although more recent research (some published after our pre-registration) suggests otherwise [10].

The following hypotheses were tested in study 2: Hypothesis 5: We expect that deep talk interactions with an AI will lead to a significant increase in interpersonal closeness compared to a baseline measure, even if participants are informed that they will interact with an AI. Although research indicates that people often hold reservations about AI interactions [7], we anticipate that participants will still develop some degree of perceived closeness with the AI, reflecting the human tendency to respond to human-like artificial agents as if they were real social partners, a phenomenon known as anthropomorphism [16,17]. Hypothesis 6: Analogous to study 1, we expect higher interpersonal closeness towards the interaction partner after interactions with humans compared to interactions with an AI across both 'label human' and 'label AI' interactions (main effect of the factor 'source identity'). Hypothesis 7: We expect higher interpersonal closeness towards the interaction partner when participants are informed that they interact with a human compared to when participants are informed that they interact with an AI across actual interactions with humans and AI (main effect of the factor 'source label'). This hypothesis is based on findings indicating reservations towards social interactions with AI [7]. Hypothesis 8: We expect that the differences (human > AI) regarding interpersonal closeness towards the interaction partner are larger when participants are informed that they are interacting with a human compared to when they are informed that they are interacting with AI (interaction effect between the factors 'source identity' and 'source label'). This hypothesis draws on the assumption that the anti-AI bias in the AI-labelled condition would reduce feelings of closeness regardless of the actual source identity, whereas in the human-labelled condition, the expected advantage of genuine human responses would be more apparent.

Methods

Study Design

Study 1 is a double-blind randomized controlled trial with a 2 x 2 between-subjects design including the factors 'source identity' (human interaction partner [hereafter referred to as 'source human'] vs. fictional characters created by a large language model [hereafter referred to as 'source AI']) and 'emotional intensity' (small-talk vs. deep-talk). Accordingly, participants were randomly (but evenly) distributed among the four treatment groups 'source human/small-talk', 'source human/deep-talk', 'source AI/small-talk', and 'source AI/deep-talk'. All participants of study 1 were informed they would interact with a human, which was a deception for the two treatment groups who actually interacted with an AI (for details on the debriefing, see Procedure/Online experiment). In study 2, we re-analysed selected data from study 1 along with other data collected for study 2. Specifically, study 2 was another double-blind randomised controlled trial with a 2 x 2 between-subjects design including the factors 'source identity' (source AI vs. source human) and 'source label' (label human vs. label AI). This study focused on deep-talk interactions only. Participants were randomly (but evenly) distributed among the four treatment groups 'source human/label human', 'source AI/label human' (both collected in study 1), 'source human/label AI', and 'source AI/label AI' (both collected in study 2). Accordingly, participants in the treatment group 'source human/label AI' were deceived about the identity of their interaction partner. In both studies, perceived interpersonal closeness to the interaction partner was used as the main dependent variable.

Sample

We recruited 18- to 35-year-old heterosexual female and male university students for our study, focussing on friendly, non-romantic relationship building among individuals of the same self-reported gender. Gender was assessed using the following item: 'Please indicate your gender' (response options: 'female', 'male', 'diverse'). Participants were excluded if they had mental health challenges, were undergoing current psychotherapeutic, neurological, or psychiatric treatment, or were abusing alcohol or drugs (for details on exclusion criteria, see [42]). Using G*Power [43], we performed an a priori power analysis for ANCOVA (fixed effects, main effects and interactions; numerator df: 2, number of groups: 4; number of covariates: 4; alpha = .05, power = .80, small to medium effect size f = .175; based on average effect sizes in social psychology and neuroscience [44,45]). This analysis yielded a required sample size of n = 318 for each study.

Expecting a drop-out rate of approximately 10%, we recruited a total sample of 359 participants for study 1. From this initial sample, 37 participants were excluded due to the following reasons: answering incorrectly to at least one out of two attention control items during the questionnaires (n = 28); an unrealistic total duration of the experiment of less than 50% of the expected minimum duration based on pre-tests (i.e., < 20 minutes, n = 2); expressing doubts about the cover story of a live interaction in the experiment (n = 6); responding in a different language (n = 1). Thus, the final sample size analysed in study 1 was n = 322 (age: M = 23.46, SD = 3.19, range: 18-35; 168 female participants, 154 male participants). For study 2, we recruited an additional sample of 179 participants, out of which 9 participants were excluded due to either incorrectly answering to the attention control items (n = 8), or an unrealistic total duration of the experiment (n = 1), leaving a sample of n = 170 (age: M = 23.25, SD = 2.96, range: 18-35; 98 female participants, 72 male participants). Together with the relevant data collected in study 1 (i.e., the two groups 'human/label human' and 'AI/label human'), the final sample size of study 2 was n = 334 (age: M = 23.22, SD = 2.98, range: 18-35; 194 female participants, 140 male participants). The total sample size of both studies was n = 492 (age: M = 23.38, SD = 3.11, range: 18-35; 266 female participants, 226 male participants).

Procedure

Preparation of human-generated responses. Participants engaged in a text-based online version of the Fast Friends Procedure (FFP), designed to generate interpersonal closeness between two unfamiliar interaction partners through escalating mutual self-disclosure in a series of turn-taking questions [36]. Human responses were generated in two in-person laboratory sessions in which individuals of the same gender took part simultaneously (session 1: 10 women; session 2: 10 men; n = 20; age: M = 21.60; SD = 2.11; range: 19-25). Note that in the main online experiment, the age of all human interaction partners was standardised to 25 to match the age of AI-generated partners, as similarity in age can influence perceived closeness. Participants were recruited with flyers and public notices and were pre-screened via an online screening questionnaire to assess exclusion criteria, which were analogous to the criteria of the two main studies. They then responded to three warm-up items on basic character information (i.e., 'What is your name and how old are you?'; 'Where do you live?'; 'What do you study?'), followed by the eight small-talk items and the eight deep-talk items from the FFP, as required for the respective conditions in the main online experiment (all items available in Supplementary Table 1). Participants were instructed to respond within 90 seconds to each item to enable short, spontaneous responses. No actual time limit was set to ensure complete responses. Participants in the laboratory sessions were informed that they would be matched with another person in the same room, and that each interaction partner would receive the responses of the other after the experiment. We chose this procedure for three reasons. First, we wanted participants to have the experience of communicating with a real person, thereby maximising their motivation to respond in a way they would find accurate in a social interaction. Second, we wanted to parallelize the procedure of this

appointment with that of the online experiment. Third, in order to use laboratory-generated answers in the online experiment, it was necessary to avoid references to previous responses of the partner that would not be meaningful in the context of other interactions.

Participants in the in-person laboratory sessions were compensated with €20 plus additional earnings from a Trust-Game (which was no further analysed here due to extreme ceiling effects, for details see section 'Measures of relationship building') of M = €3.13 on average (SD = .886). After the experiment, they received the responses of their interaction partner to the FFP items. From the total sample, the responses of three men and three women were randomly selected as the interaction partner's responses for the online experiment. Although only three sets of responses were needed from each group, we invited 10 participants to each laboratory session to enhance the feeling of actual interactions and to account for potential dropouts due to inadequate responses. In the end, none of the participants provided inadequate responses.

Preparation of AI-generated responses. AI responses were generated using the large language model PaLM 2 (interface: Google BARD; Google LLC, CA, USA; date of access: February 19, 2024). The AI was instructed to create six fictional student characters (three men and three women) aged 25 and then answer the eight small-talk and the eight deep-talk items of the FFP from the perspective of these fictional characters. As we found that the plausibility of AI-generated characters decreased as their number increased, we set six as a compromise between plausibility and representability. The following minimal prompt was used to keep AI responses as close as possible to its default style: 'Create six different biographies of typical students (three women and three men) aged 25 and answer the following questions from the perspective of the six students.' Importantly, responses of all AI-generated partners were generated in a single session to ensure character consistency across FFP items. The age of 25 was chosen as it was the approximate average age expected for the study sample, ensuring minimal age difference (potentially affecting relationship building) with participants ranging from 18 to 35 years. Note that responses to warm-up items for both human and AI interaction partners were drawn from the aforementioned AI-generated characters to maintain consistency across conditions in terms of name, age, place of residence, and field of study. This approach ensured that any observed differences between conditions could be attributed to variations in responses rather than demographic differences. Biographies of the six AI-generated characters and responses of both AI-generated characters and the six human interaction partners are shown in Supplementary Table 1.

Recruitment. To recruit participants, we contacted student councils across German universities, requesting them to forward the study flyer to their students. We also used flyers, public notices and print and online social media to recruit participants in Freiburg, Germany. Via a URL or QR code, interested persons were forwarded to a screening questionnaire, where exclusion criteria were assessed. Suitable participants immediately continued with the online experiment after finishing the screening.

Online experiment. This study was approved by the Ethics Committee of the University of Freiburg (ETK-Freiburg application code: 23-1479-S2, February 15, 2024) and conducted in accordance with the principles of the Declaration of Helsinki. No data on race or ethnicity was collected. Data for both study 1 and study 2 were collected simultaneously to prevent history effects (data collection period: March 3 to June 12, 2024). Participants first read and signed an informed consent form, which provided details about the study's procedure and specified whether they would be interacting with another human (studies 1 and 2) or an AI (study 2). They then entered a virtual waiting room for 50 seconds, during which they were informed that another participant of the same gender was being located for their upcoming online interaction and that this process might take a few minutes. As this study focussed on friendly, non-romantic relationships, participants were always assigned to an interaction partner (real or fictional) of the same gender. Participation in the online study was limited to the hours between 4:00 and 9:00 pm to enhance

the plausibility of a real-time interaction. Prior to starting the FFP, participants were instructed to respond to each item within three minutes to avoid lengthy waiting times for their interaction partner. However, no actual time limit was enforced. During the FFP, participants first responded to three warm-up items and, after each response, read the corresponding response to the same item from their interaction partner. To ensure consistency in basic character information across conditions, participants in both the human and AI interaction groups received AI-generated responses to the warm-up items, while no personal information from human partners was used. Next, participants completed a brief baseline questionnaire measuring perceived interpersonal closeness to their partner. They then proceeded with the main FFP phase, consisting of either eight small-talk or eight deep-talk items. Here, they were required to read the partner's response to each item for a minimum of 25 seconds before they could continue. To enhance the feeling of a live interaction, we introduced random waiting periods of 3 to 10 seconds after participants submitted their responses. This setup was designed to create the impression that sometimes participants submitted their responses first (resulting in a waiting period), while at other times, their partner responded first (resulting in no waiting period). To reduce potential pressure from sending responses slower than the partner, waiting periods were included slightly more often than not (i.e., in 5 out of 8 cases). After responding to the last item, participants completed the brief questionnaire again to assess perceived interpersonal closeness to their interaction partner after completion of the FFP. Next, participants engaged in an interactive trust game where they could earn additional monetary gains based on their own decisions and those of their interaction partner [46]. Finally, participants completed questionnaires measuring individual trait characteristics evaluated elsewhere. Two attention control questions were included in the questionnaire battery to identify inattentive participants, explicitly requesting them to select a specific response option. In total, the experiment had a duration of approximately 40 minutes. Participants of each study received a compensation of €15 plus the additional average gain from the trust game of M = €3.54 (SD = .969). As this research involved a degree of deception about the true source of the responses and the real-time nature of the social interaction, participants were debriefed via email about the true source of the responses they had read and about the fact that the responses had been generated before the actual experiment took place. Debriefing took place two weeks after the end of data collection.

Measures of relationship building

To obtain an intuitive measure of perceived interpersonal closeness to the interaction partner (or simply 'interpersonal closeness'), we applied an adapted version of the widely used Inclusion of Other in the Self Scale (IOS [47–49]), a one-item pictorial scale including nine images depicting two progressively overlapping circles to represent the 'self' and the 'other'. Here, we labelled the 'other' as the 'interaction partner' to specifically measure closeness to the interaction partner. More overlapping circles indicate a higher 'inclusion of the other in the self' and thus a higher perceived interpersonal closeness. The IOS shows good psychometric properties, including two-week retest-reliability (r = .86) and convergent, discriminant and predictive validity [47,50]. Note that we refrained from using difference scores (post minus pre) in our main analyses, as pre-measures are likely already influenced by the information from the warm-up items of the FFP (i.e., name, age, residence, and field of study), which can bias social perception [51,52]. Subtracting these initial impressions could thus reduce the variance of interest in the post measures of closeness. This issue is further amplified in the two 'label AI' conditions, where pre-measures are additionally shaped by participants' expectations of interacting with AI, meaning that differencing would remove precisely those initial attitudes that are central to our research question. Consistent with our preregistration, we therefore relied on post measures only.

We pre-registered the use of the Interpersonal Closeness Questionnaire [53] as an additional measure of interpersonal closeness. However, we chose to focus on the IOS for our

analyses because (a) the two measures showed high redundancy ($r_{(490)}$ = .723, p < .001, 95% CI [.678, .762]), (b) the IOS is more widely used than the ICQ, and (c) the IOS was more sensitive to the FFP across both studies and all conditions (as indicated by increases from pre- to post-measures analysed in hierarchical linear models for repeated measures [for details, see section 'Statistical analyses']; IOS: $t_{(493)}$ = 19.40, p < .001, R^2_m = .106; ICQ: $t_{(493)}$ = 14.50, p < .001, R^2_m = .072). Note that the key results of both studies are similar when using the ICQ. We also planned to include trust towards the interaction partner, assessed in an interactive decision-game with real monetary consequences (i.e., the trust game [46]), as a measure of relationship building. However, this measure exhibited limited variability due to extreme ceiling effects, with 47.4% of participants choosing the maximum transfer. As a result, we decided not to include the analysis of trust in the study.

Linguistic analysis

To explore potential reasons for identified differences in relationship building between treatment groups in both studies, we conducted automated linguistic analyses of participants' as well as their human or AI interaction partners' responses to FFP items using the Linguistic Inquiry and Word Count system (LIWC-22 [39,40]). The LIWC is a software that counts words in text files that match specific categories, and quantifies them relative to the total word count. In the current study we focused on analysing self-disclosure [41], which is assumed to be the main mechanism underlying the establishment of interpersonal closeness in the FFP [36]. In the LIWC, the category self-disclosure consists of the three subscales 'self-related personal pronouns' (e.g., I, me, my), 'emotion' (e.g., happy, cry, abandon), and 'social processes' (e.g., friend, talk, family), together forming the self-disclosure variable. Consistent with Callaghan and colleagues [41], who developed the self-disclosure measure within the LIWC, we also analysed the total response length as an indicator of the motivation to engage in the social interaction.

Statistical analyses

Research questions and analyses were pre-registered at the OSF repository (https://osf.io/chdx7). The data and code used to generate the results of this study are freely available there as well (https://osf.io/qs6yf/). Closeness measures after the interaction (i.e., post measures) are used as the main dependent variable across both studies. As basic analyses in both studies, we computed HLM for repeated measures to investigate whether any increase in interpersonal closeness would take place over the course of the FFP in specific experimental conditions (e.g., in masked deep-talk interactions with AI [study 1, pre-registered hypothesis 1] and unmasked deep-talk interactions with AI [study 2, pre-registered hypothesis 5]). These analyses included the repeated measures factor 'time' (two levels 'pre' and 'post' per participant), the dependent variable 'interpersonal closeness', the covariates age and gender, and a random intercept across participants (as pre- and post-measures of closeness were nested within participants). To measure effect size in these models, we computed marginal R-squared values following the recommended procedure by Nakagawa & Schielzeth [54]. To test whether increases in closeness differed meaningfully between human and AI interactions, we conducted equivalence testing (TOST) with bounds of ±.35 pooled standard deviations, representing small-to medium effect sizes. This analysis was complemented by a Bayesian two-sample t-test to quantify evidence for or against a meaningful difference.

Participants were nested within groups, each interacting with one out of 24 partners (2 source identities [human vs. AI] × 2 emotional intensities [small-talk vs. deep-talk] × 2 genders [female vs. male] × 3 human or AI-generated partners per condition). To assess data dependency in interpersonal closeness ratings, we used an ANOVA to compare a standard model of closeness with a random intercept model allowing variation across interaction partners [55]. As the model with random variation across intercepts did not surpass the standard model (p = .356), hierarchical

linear modelling (HLM) was not required for our main analyses. Thus, ANCOVAs were conducted to test main and interaction effects of the independent variables ‘source identity’ (human vs. AI; studies 1 and 2), ‘emotional intensity’ (small-talk vs. deep-talk; study 1), and ‘source label’ (label human vs. label AI, study 2) on the dependent variable ‘interpersonal closeness’ while controlling for the effects of age and gender (pre-registered hypotheses 2, 3, 4 [study 1], and 6, 7, and 8 [study 2]). Analogous analyses were performed within specific treatment groups as post-hoc tests. Whenever relevant, we formally tested the normality of residuals and equality of variances. The assumption of equal variances was met in all analyses, as indicated by Levene’s test (all $p > .05$). Although interpersonal closeness was not normally distributed across studies (Shapiro-Wilk test, $p < .001$), and residuals in some analyses deviated from normality, both HLM and ANOVA were applied, as they are generally robust to violations of normality, especially with large sample sizes [56,57].

We then investigated whether treatment group differences in closeness could be explained by differences in self-disclosure and/or response length. Specifically, we conducted (a) ANCOVA to determine whether treatment group differences in closeness would parallel treatment group differences in the participant’s and/or their interaction partner’s self-disclosure and/or response length, and (b) partial correlation analyses to examine whether the participant’s and/or their interaction partner’s self-disclosure and/or response length were associated with the participant’s perceived interpersonal closeness. Age and gender were controlled for in both analyses. P-values smaller than .05 (two-tailed) were considered statistically significant across all analyses.

Results

Do people build human-like relationships with AI?

As a basic analysis of study 1, we first tested whether interpersonal closeness increased at all in masked interactions with an AI (i.e., the interaction partner was labelled as a human). Repeated measures HLM analyses revealed statistically significant increases in closeness from pre- to post-interaction measures in human-labelled AI interactions across both deep-talk and small-talk conditions ($t_{(165)}$ = 11.92, b = 1.16, 95% CI [.970, 1.36], p < .001, R^2_m = .121; see Supplementary Figure 1A) and separately within each condition (small-talk: $t_{(81)}$ = 7.26, b = .988, 95% CI [.717, 1.26], p < .001, R^2_m = .112; deep-talk: $t_{(83)}$ = 9.67, b = 1.33, 95% CI [1.06, 1.61], p < .001, R^2_m = .153). As predicted, deep-talk interactions with AI significantly increased interpersonal closeness compared to a baseline measure, confirming hypothesis 1. For comparison, we also analysed the increase in closeness in interactions with humans. Again, we found statistically significant increases in closeness across both deep-talk and small-talk conditions ($t_{(155)}$ = 11.25, b = 1.10, 95% CI [.909, 1.30], p < .001, R^2_m = .114; see Supplementary Figure 1B) and within each condition (small-talk: $t_{(75)}$ = 7.47, b = 1.13, 95% CI [.830, 1.43], p < .001, R^2_m = .100; deep-talk: $t_{(79)}$ = 8.48, b = 1.07, 95% CI [.823, 1.33], p < .001, R^2_m = .138). An equivalence test followed by a Bayesian t-test revealed that differences in closeness increases between human and AI interactions across levels of emotional intensity were practically negligible when testing against the presence of a small-to-medium effect ($t_{(139.58)}$ = -2.71, p = .004, Hedges' g = .048, 95% CI [-.170, .266]). The Bayes-Factor, using the default medium-sized JZS prior (r = .707; BF_{01} = 7.43; posterior mean effect size δ = .046, based on 20,000 iterations; 95% credible interval [-.168, .258]), indicated that the data were 7.43 times more likely to occur under the null hypothesis than under the alternative hypothesis. To assess robustness, we repeated the analysis with a wider prior (r = 1.00), resulting in a similar conclusion (BF_{01} = 10.37).

As the main analysis of study 1, we compared interpersonal closeness as measured after FFP interactions by running an ANCOVA with the independent variables 'source identity' (levels: source human and source AI), 'emotional intensity' (levels: small-talk and deep-talk), their interaction, and the covariates age and self-reported gender. While there were no statistically significant main effects of 'source identity' ($F_{(1, 316)}$ = .097, p = .756, η^2_p < .001, 95% CI [0.00, .014]; $M_{source\ human}$ = 4.08, SD = 1.71; $M_{source\ AI}$ = 4.15, SD = 1.94) or 'emotional intensity' ($F_{(1, 316)}$ = .420, p = .518, η^2_p = .001, 95% CI [0.00, .021]; $M_{small\text{-}talk}$ = 4.06, SD = 1.90; $M_{deep\text{-}talk}$ = 4.17, SD = 1.77), we found a statistically significant interaction effect between the two variables ($F_{(1, 316)}$ = 8.17, p = .005, η^2_p = .025, 95% CI [.002, .068]). Post-hoc tests revealed higher closeness after interacting with an AI compared to a human within the deep-talk condition ($F_{(1, 160)}$ = 4.89, p = .028, η^2_p = .030, 95% CI [0.00, .096]; $M_{source\ human}$ = 3.85, SD = 1.69; $M_{source\ AI}$ = 4.48, SD = 1.80), but no statistically significant difference in closeness within the small-talk condition ($F_{(1, 162)}$ = 2.89, p = .091, η^2_p = .018, 95% CI [0.00, .078]; $M_{source\ human}$ = 4.33, SD = 1.71; $M_{source\ AI}$ = 3.82, SD = 2.04; see Figure 1A). Surprisingly, our results provide neither evidence that interactions with humans yield stronger feelings of closeness than interactions with AI, nor that deep-talk interactions yield stronger feelings of closeness than small-talk interactions. Consequently, hypotheses 2 and 3 are rejected. Furthermore, and contrary to hypothesis 4, AI interactions yielded stronger feelings of closeness than human interactions, but only in the deep-talk condition, not in the small-talk condition.

To further explore why people feel closer to the AI than to humans following deep-talk interactions, we tested whether AI-generated responses differed from human-generated responses regarding self-disclosure as measured by the Linguistic Inquiry and Word Count system (LIWC-22 [39,40]). An ANOVA (using a dataset including all human- and AI-generated responses within the deep-talk condition; n = 12) revealed that AI-generated responses showed

considerably higher levels of self-disclosure than human-generated responses ($F_{(1, 10)}$ = 18.57, p = .002, η^2_p = .650, 95% CI [.167, .801]; M_{AI} = 38.03, SD_{AI} = 2.62; M_{human} = 32.31, SD_{human} = 1.92; Figure 1B; no statistically significant difference regarding response length; $F_{(1, 10)}$ = 18.57, p = .392, η^2_p = .074, 95% CI [.000, .407]; M_{AI} = 313.83, SD_{AI} = 18.89; M_{human} = 295.50, SD_{human} = 46.45). Next, we tested whether self-disclosure shown in partner responses was related to participants' perceived interpersonal closeness using partial correlation analysis controlling for age and gender (using a dataset of deep-talk interactions in study 1; n = 164). We found that participants felt closer to their interaction partner when their partners' responses displayed higher levels of self-disclosure ($r_{p(160)}$ = .242, 95% CI [.091, .382], p = .002, R^2 = .058; Figure 1C).

We then tested whether the participants' own responses differed between interactions with an AI and interactions with a human. An ANCOVA demonstrated that participants showed higher levels of self-disclosure in interactions with an AI ($F_{(1, 160)}$ = 3.92, p = .049, η^2_p = .024, 95% CI [0.00, .087]; M_{AI} = 34.80, SD_{AI} = 5.00; M_{human} = 33.26, SD_{human} = 4.61; Figure 1D; no statistically significant difference regarding response length; $F_{(1, 160)}$ = .007, p = .932, η^2_p < .001, 95% CI [0.00, .012]; M_{AI} = 197.62, SD_{AI} = 62.58; M_{human} = 199.15, SD_{human} = 64.58). However, participants own self-disclosure was not significantly associated with their perceived closeness in deep-talk interactions ($r_{p(160)}$ = .117, 95% CI [-.038, .266], p = .139). Lastly, we analysed whether the degree of self-disclosure shown in the responses of the interaction partners related to participants' own self-disclosure. Indeed, participants' self-disclosure was positively associated with their interaction partner's self-disclosure ($r_{p(160)}$ = .183, 95% CI [.030, .328], p = .020, R^2 = .033; Figure 1E).

To summarise the findings of study 1, the AI-generated content outperformed human-generated content in establishing feelings of closeness during emotionally engaging deep-talk interactions. This effect appears to be driven by higher levels of self-disclosure by AI partners compared to human partners, which, in turn, enhanced participants' perceived interpersonal closeness. Moreover, participants disclosed more information themselves in interactions with AI and self-disclosure levels of both parties were associated with each other. These findings suggest that the AI's increased self-disclosure motivates participants to disclose more personal information themselves, ultimately leading to more intimate interactions and a stronger sense of closeness.

[FIGURE 1]

Does the belief of interacting with an AI hinder relationship building (study 2)?

Building on the basic analysis from study 1, which demonstrated that relationship building occurs in masked interactions with an AI (i.e., when the AI was labelled as a human), we first examined in study 2 whether the belief of interacting with an AI prevents relationship building. Using repeated measures HLM analyses as described before, we analysed whether interpersonal closeness increased in interactions in which participants were informed they would interact with an AI, both when they actually interacted with the AI ('source AI/label AI') and when they actually interacted with a human ('source human/label AI'). Repeated measures HLM analyses revealed statistically significant effects of 'time' (pre vs. post interaction) on closeness both across levels of 'source identity' ($t_{(169)}$ = 10.41, b = 1.03, 95% CI [.834, 1.22], p < .001, R^2_m = .120, SD = 1.29; see Supplementary Figure 1C) and separately within each condition (interactions with humans: $t_{(83)}$ = 6.93, b = 1.03, 95% CI [.738, 1.33], p < .001, R^2_m = .187; interactions with AI: $t_{(82)}$ = 7.91, b = 1.02, 95% CI [.767, 1.28], p < .001, R^2_m = .080). These findings demonstrate that relationship building occurred even when participants were informed they were interacting with an AI, supporting hypothesis 5. For comparison, we also analysed the increase in closeness in

interactions in which participants were informed they would interact with a human (across levels of 'source identity'). The analysis revealed a statistically significant increase in closeness across sources when participants were informed they would interact with a human ($t_{(163)}$ = 12.82, b = 1.21, 95% CI [1.02, 1.39], p < .001, R^2_m = .123; see Supplementary Figure 1D).

As the main analysis of study 2, we compared closeness after AI-labelled and human-labelled interactions by running an ANCOVA with the independent variables 'source identity' (pre-registered hypothesis 6), 'source label' (pre-registered hypothesis 7), their interaction (pre-registered hypothesis 8), and the covariates age and gender. The analysis revealed a statistically significant main effect of 'source label', with lower interpersonal closeness following AI-labelled interactions ($F_{(1, 328)}$ = 4.90, p = .028, η^2_p = .015, 95% CI [0.00, .050]; $M_{label\ human}$ = 4.17, SD = 1.77, $M_{label\ AI}$ = 3.72, SD = 1.90; Figure 2A), but no statistically significant main effect of 'source identity' ($F_{(1, 328)}$ = 3.05, p = .082, η^2_p = .009, 95% CI [0.00, .041]; $M_{source\ human}$ = 3.77, SD = 1.79, $M_{source\ AI}$ = 4.11, SD = 1.90) and no statistically significant interaction ($F_{(1, 328)}$ = 2.03, p = .155, η^2_p = .006, 95% CI [0.00, .033]). Again, these results provide no evidence that participants establish stronger feelings of closeness with human interaction partners than with AI interaction partners. Thus, hypothesis 6 is rejected. As expected, labelling the interaction partner as an AI led to lower interpersonal closeness ratings after the interaction compared to when the partner was labelled as human, demonstrating an anti-AI bias and confirming hypothesis 7. Hypothesis 8 was rejected, as we found no evidence for differences in interpersonal closeness (human > AI) following human interactions compared to AI interactions.

As participants were presented with responses from the same human or AI interaction partners in both human-labelled and AI-labelled conditions, the difference in interpersonal closeness cannot be attributed to variations in the partners' responses. We therefore tested whether people themselves communicated differently with AI-labelled partners than human-labelled partners in an exploratory fashion. Indeed, while there were no statistically significant group differences regarding self-disclosure ($F_{(1, 330)}$ = .050, p = .823, η^2_p < .001, 95% CI [.000, .012]; M_{AI} = 34.14, SD_{AI} = 4.39; M_{human} = 34.05, SD_{human} = 4.86), people wrote significantly longer responses when assuming they would interact with a human ($F_{(1, 330)}$ = 5.32, p = .022, η^2_p = .016, 95% CI [< .001, .052]; M_{AI} = 181.78, SD_{AI} = 67.55; M_{human} = 198.37, SD_{human} = 63.37; Figure 2B). In turn, longer responses were also related to higher levels of perceived closeness ($r_{p(330)}$ = .257, 95% CI [.154, .355], p < .001, R^2 = .066; Figure 2C). AI-labelled interactions thus resulted in both shorter responses of participants and lower levels of perceived closeness.

We found evidence that individuals form social bonds with AI even when being aware of interacting with an artificial agent, yet also observed an anti-AI bias leading to lower feelings of closeness after the interaction compared to human-labelled interactions. To explore why some people form social bonds to AI while others do not, we examined whether AI scepticism is more pronounced in individuals who value natural human communication. Indeed, we found a statistically significant interaction effect between 'source label' and *universalism*, a personal value centred on concern for the welfare of people and nature [58], in predicting interpersonal closeness ($F_{(1, 328)}$ = 4.11, p = .043, η^2_p = .012, 95% CI [0.00, .046]). This effect was driven by a positive association between universalism and closeness in human-labelled interactions ($r_{p(162)}$ = .240, 95% CI [.136, .339], p = .002, R^2 = .058), which was not present in AI-labelled interactions ($r_{p(168)}$ = .017, 95% CI [-.091, .125], p = .825, R^2 < .001). These findings indicate that personal values may modulate relationship building with AI.

To summarise the findings of study 2, relationship building occurred even when participants believed that they interacted with AI. However, participants felt less close to AI-labelled partners compared to human-labelled partners. This effect seems to stem from lower motivation to engage with AI partners, as evidenced by shorter responses.

[FIGURE 2]

Discussion

Can we 'befriend' an AI? The present study examined the effects of source identity (AI-generated responses vs. human-generated responses), source label (interaction partner labelled as an AI vs. interaction partner labelled as a human), and emotional intensity (small-talk vs. deep-talk) on relationship building. We found that humans indeed build relationships with fictional characters created by AI. Strikingly, AI-generated content outperformed human-generated content in establishing feelings of closeness during emotionally engaging deep-talk interactions (including topics such as the most treasured memory of your life or what you value most in a friendship). Follow-up analyses suggest that AI-responses featured higher levels of self-disclosure in emotional interactions, which in turn elicited higher levels of self-disclosure and perceived interpersonal closeness in participants. Being explicitly informed that one would interact with an AI led to an anti-AI bias, which reduced, but did not prevent, relationship building. This effect might be due to lower motivation of participants to engage with AI, as evidenced by both shorter responses and lower levels of reported closeness.

One of the key innovations of our study is the direct comparison between human-to-AI and human-to-human relationship building. Recent studies have demonstrated that people do form some sort of relationship with 'AI social companions' [59], and highlight AI's socio-emotional capabilities, such as emotional awareness [27], empathetic responses [60], and the ability to offer relationship advice [11]. Expanding upon this research, we provide evidence that, when assuming they are interacting with a human, people form relationships with AI to a similar extent as with fellow humans, as shown by a similar increase in interpersonal closeness over the course of the interaction. Moreover, we found that people felt even closer to AI than to fellow humans after emotionally engaging interactions. At first glance, this seems counterintuitive, as AI lacks the emotion-related bodily sensations that underpin human emotional experiences [61]. On second thought, however, this 'deficit' could also create an advantage for AI. For humans, disclosing personal information on emotionally charged topics is risky and requires trust, as the recipient might fail to show the desired empathic reaction or even misuse the information to one's disadvantage. As a result, humans often avoid discussing emotional topics to protect themselves [62,63]. In contrast, as AI cannot experience emotions, there are no such restrictions when opening up about emotionally charged topics. Indeed, follow-up linguistic analyses suggest that AI-generated responses showed higher levels of self-disclosure, which, in turn, also encouraged more self-disclosure by participants [41]. Importantly, both the partners' and participants' own self-disclosure were associated with higher levels of perceived interpersonal closeness. Thus, our findings highlight not only AI's ability to excel in emotionally charged communication, but also its potential to help humans feel more comfortable opening up compared to interactions with another human.

Does the higher level of self-disclosure shown by AI imply that AI is generally superior to humans in emotional conversations? Probably not. First, emotional conversations between humans serve many purposes, and building a relationship is just one of them. As neither human nor AI responses in our study were generated with a specific goal in mind, humans may still outperform AI when actively trying to build a relationship. Second, high self-disclosure at certain levels and in certain situations may be perceived as unnatural (e.g., when interacting with complete strangers), unprofessional (e.g., in the workplace), or even risky (e.g., when interacting with an untrustworthy person [62,63]). AI models may be less flexible and precise than humans in appropriately adjusting self-disclosure across different situations. Third, AI-generated content only resulted in greater feelings of closeness when disguised as human-generated, which is not the case in everyday applications. Consistent with recent findings showing that labelling content

as AI-generated reduces its positive perception [7], participants reported less interpersonal closeness when they were informed they would interact with an AI. This is not surprising, as humans generally prefer human over AI partners, particularly in areas often considered uniquely 'human' such as emotional interactions [25,26,64,65]. Additional analyses showed that participants provided shorter responses under these circumstances, indicating less motivation to engage in personal interactions with an AI. These findings align with the Social Need Fulfillment Model for Human-AI relationships [66], which suggests that human-AI interactions typically satisfy only concrete social needs (e.g., pleasure) rather than deeper, symbolic needs (e.g., feeling genuine care). Humans possess a fundamental, evolutionary need for social connection that involves shared emotions, mutual understanding, and the comfort of knowing someone else truly comprehends our feelings [67,68]. The knowledge that another person feels what we feel, understands our intentions, and responds with empathy and authenticity is essential to humanity, something that AI, at least at the current state, cannot truly replicate. So why do individuals form relationships with AI at all, even when being aware of its artificial nature (also see [59])? One explanation is the phenomenon of anthropomorphism, the tendency to attribute human traits, emotions, or intentions to non-human entities [16,17]. Humans are inherently social, so when presented with AI-generated cues that resemble human-generated cues, the brain may intuitively respond to these artificial cues much as it would to genuine social cues. At the same time, our results highlight that at least some individuals remain reluctant to engage with AI, leading to the question of how these people differ from those who are more receptive. In an exploratory analysis, we found that universalism moderated the difference in interpersonal closeness following AI- versus human-labelled interactions. Specifically, individuals high in universalism felt closer to humans, but not to AI. This indicates that traits linked to natural social interaction, and, vice versa, potentially to negative attitudes towards artificial interaction, may reduce the likelihood of forming bonds with AI. However, this result requires further validation in future research.

As has often been the case with technological advances throughout history [69,70], the rise of AI brings both benefits and risks to society. Our findings underscore these two sides of the coin in social applications, highlighting AI's potential for relationship building in overburdened social fields, while also emphasising the risks of its misuse, especially when disguised as human. Healthcare is struggling to meet the growing demand for services due to factors such as aging populations, reduced funding, and the growing psychological toll on healthcare workers [71–73]. Conversational AI may help alleviate this burden. As demonstrated by our finding that AI excels particularly in emotional conversation, it could be effective in psychotherapy and medical settings where relationship building and adequate interaction on emotional topics is key (for reviews, see [74,75]). These settings include health- or psychoeducation, providing care for individuals with limited access to therapy, offering social contact to alleviate loneliness in the elderly, bridging waiting times until the start or between psychotherapy sessions, and facilitating communication with patients [75–78]. Importantly, AI should assist, not replace, therapists, as exclusive reliance on AI for addressing health issues may lead to over-reliance, addiction, or withdrawals from human relationships [79–82], as well as other unforeseeable harmful effects (e.g., not adequately reacting to the expression of suicidal intentions [83]). Therefore, a human introduction and ongoing monitoring are imperative for safe and effective use. Our results also demonstrate negative attitudes towards social interaction with AI, even in a relatively young sample aged 18 to 35 years. To fully unlock AI's potential in healthcare applications, efforts are needed to increase AI acceptance in these fields, e.g., by providing a clear and transparent explanation of the reasons for using AI by a human before an AI intervention begins, or by implementing follow-up human-to-human sessions to discuss and integrate previous human-to-AI interactions [84–88].

Beyond AI's potential benefits in healthcare, our findings also highlight the risks AI poses for society, especially when AI-generated information is presented as human-created. AI-generated content is already flooding social media [89]. With the quality of AI content improving to

the point where it even outperforms humans in certain social contexts (such as relationship building in emotional interactions), the risk of people falling for deceptive traps continues to grow. Specifically, AI's ability to build social-emotional relationships can be misused to establish deceptive emotional connections, steal personal data, and enable exploitation by unethical actors, including both individuals and corporations [90–92]. Obviously, people are far more likely to buy products, disclose personal information, or send money when a request appears to come from a friend rather than from an easily recognisable scam email. Effectively tackling AI misuse requires a combination of transparent regulations, ethical guidelines, robust detection mechanisms, and public awareness to ensure responsible development and use [85,93].

Limitations

We acknowledge several limitations of this study. Building on our findings, future studies could explore relationship building with AI in real-time interactions by combining conversational AI with avatars and voice generation [94,95]. However, while this could enhance efficiency and acceptance by making interactions become more life-like, it may also backfire if AI becomes too human-like as suggested by the Uncanny Valley hypothesis [96]. Furthermore, while the Fast Friends Procedure used here provides a well-stablished and effective framework for establishing relationships in a semi-standardised manner [36], less structured interactions may offer greater ecological validity in future research. AI responses to the FFP items were generated using the large language model PaLM 2 in February 2024. As a result, the findings may not generalize to other AI systems. However, this also suggests that we may have underestimated the communicative capabilities of more advanced models available at the time of submission (May 2025), such as GPT-4 (OpenAI, CA, USA) or Gemini 2.5 (Google DeepMind, CA, USA). The finding that AI outperformed humans in fostering emotional connections, even when using now-outdated software, speaks volumes about the potential of future AI systems in this domain. Additionally, the AI was prompted to respond from the perspective of six students. Although we used only a one-sentence minimal prompt to keep responses as close as possible to standard AI output, this approach may still affect how our findings relate to typical AI interactions, which do not involve such prompting. Importantly, however, the prompt did not include instructions regarding the tone of the interaction (e.g., the degree of self-disclosure, empathy, or emotionality), demonstrating that AI-generated responses showed self-disclosure and fostered relationship building even without specific prompting to do so. As noted by a reviewer, an alternative and less minimal prompt could be to instruct the AI to respond from the perspective of the specific human profiles used in this study, which could provide broader control for the characters presented. Relatedly, including more than six human and AI interaction partners per condition could be beneficial in future studies to better represent typical human and AI responding. We also note that this study features a WEIRD sample (Western, educated, industrialised, rich, and democratic [97]), limiting the generalisability of our findings. However, given that AI technologies are predominantly developed and adopted in WEIRD contexts, these settings provide a meaningful foundation for studying human-AI interactions, even if broader generalisability remains an open question. Future research may also benefit from incorporating neuroimaging techniques such as EEG or fMRI to investigate the neural underpinnings of human–AI bonding or attitudes towards AI (for related research from our group, see [98–103]. Finally, longitudinal studies are needed to examine whether human-AI relationships can be sustained over time and whether they can reach or even surpass the well-documented long-term mental and physical benefits of human social bonding [68,98,104–108].

Conclusion

In conclusion, we present three key findings: First, people form relationships with AI to a similar extent as with other humans when the partner is labelled as human. Second, even minimally prompted AI can outperform humans in establishing feelings of closeness in emotional conversations, partly due to higher levels of self-disclosure. Third, people show an anti-AI bias,

as evidenced by weaker relationship building when the partner is explicitly labelled as AI. Together, these findings highlight the dual role of conversational AI as both a powerful tool and a potential risk for society. On one hand, AI shows great promise in alleviating strain in overburdened social fields such as psychotherapy, medical care, and elder care. To foster acceptance in these areas, we recommend transparent human-led introduction, continuous monitoring, and systematic evaluation of human-AI interactions. On the other hand, our results underscore the risk of AI being misused for manipulation by fostering deceptive emotional connections. Clear ethical guidelines and safeguards are therefore crucial to ensure that conversational AI is leveraged responsibly and for societal benefit.

Author contributions (CRediT statement)
Conceptualization: Tobias Kleinert, Marie Waldschütz, Bastian Schiller, Markus Heinrichs; Methodology: Tobias Kleinert, Marie Waldschütz, Julian Blau, Bastian Schiller; Software: Tobias Kleinert, Marie Waldschütz, Julian Blau; Investigation: Tobias Kleinert, Marie Waldschütz, Julian Blau; Data Curation: Tobias Kleinert, Marie Waldschütz, Julian Blau; Formal Analysis: Tobias Kleinert, Marie Waldschütz, Bastian Schiller; Resources: Markus Heinrichs, Bastian Schiller; Writing – Original Draft: Tobias Kleinert, Bastian Schiller; Writing – Review and Editing: Marie Waldschütz, Julian Blau, Markus Heinrichs; Visualization: Tobias Kleinert; Supervision: Markus Heinrichs, Bastian Schiller; Project Administration: Tobias Kleinert, Markus Heinrichs, Bastian Schiller; Funding Acquisition: Bastian Schiller

Competing interest information
The authors declare no competing interests.

Data availability statement
The data used to generate the results of this study are freely available in the OSF repository (https://osf.io/qs6yf/).

Code availability statement
The code used to generate the results of this study are freely available in the OSF repository (https://osf.io/qs6yf/).

Acknowledgments
Funded by the European Union (European Research Foundation; ERC Starting Grant; SODI, ‘From face-to-face to face-to-screen: Social animals interacting in a digital world’, awarded to Prof. Dr. Bastian Schiller, Project 101076414). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. Elsholz, E., Chamberlain, J., & Kruschwitz, U. (2019). Exploring Language Style in Chatbots to Increase Perceived Product Value and User Engagement. *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 301–305. https://doi.org/10.1145/3295750.3298956

2. Inzlicht, M., Cameron, C. D., D'Cruz, J., & Bloom, P. (2024). In praise of empathic AI. *Trends in Cognitive Sciences*, *28*(2), 89–91. https://doi.org/10.1016/j.tics.2023.12.003

3. Kambeitz, J., & Meyer-Lindenberg, A. (2025). Modelling the impact of environmental and social determinants on mental health using generative agents. *npj Digital Medicine*, *8*(1), 36. https://doi.org/10.1038/s41746-024-01422-z

4. Kellmeyer, P. (2019). Artificial Intelligence in Basic and Clinical Neuroscience: Opportunities and Ethical Challenges. *Neuroforum*, *25*(4), 241–250. https://doi.org/10.1515/nf-2019-0018

5. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473. https://doi.org/10.1016/j.artint.2021.103473

6. Skjuve, M., Følstad, A., & Brandtzaeg, P. B. (2023). The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–10. https://doi.org/10.1145/3571884.3597144

7. Yin, Y., Jia, N., & Wakslak, C. J. (2024). AI can help people feel heard, but an AI label diminishes this impact. *Proceedings of the National Academy of Sciences*, *121*(14), e2319112121. https://doi.org/10.1073/pnas.2319112121

8. Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., & Hogarth, M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, *183*(6), 589–596.

9. Lee, Y. K., Suh, J., Zhan, H., Li, J. J., & Ong, D. C. (2024). *Large Language Models Produce Responses Perceived to be Empathic*. 63–71. https://doi.org/10.1109/ACII63134.2024.00012

10. Ovsyannikova, D., de Mello, V. O., & Inzlicht, M. (2025). Third-party evaluators perceive AI as more compassionate than expert humans. *Communications Psychology*, *3*(1), 4. https://doi.org/10.1038/s44271-024-00182-6

11. Vowels, L. M. (2024). Are chatbots the new relationship experts? Insights from three studies. *Computers in Human Behavior: Artificial Humans*, 100077. https://doi.org/10.1016/j.chbah.2024.100077

12. Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Mental Health Research*, *3*(1), 4. https://doi.org/10.1038/s44184-023-00047-6

13. He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q., & Hou, X. (2023). Conversational agent interventions for mental health problems: Systematic review and meta-analysis of randomized controlled trials. *Journal of Medical Internet Research*, *25*, e43862. https://www.jmir.org/2023/1/e43862

14. Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, *140*, 107600. https://doi.org/10.1016/j.chb.2022.107600

15. Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2022). A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, *168*, 102903. https://doi.org/10.1016/j.ijhcs.2022.102903

16. Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

17. Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge University Press, 10*, 19–36. https://psycnet.apa.org/record/1996-98923-000

18. Altman, I., & Taylor, D. (1973). Communication in interpersonal relationships: Social penetration theory. *Interpersonal processes: New Directions in Communication Research*, *14*, 257–277.

19. Blau, P. M. (1968). Social exchange. In D.L. Sills (Ed.), *International Encyclopedia of the Social Sciences* (Vol. *7*(4), pp. 452–457). Macmillan.

20. Carpenter, A., & Greene, K. (2016). Social penetration theory. In *The International Encyclopedia of Interpersonal Communication* (pp. 1–5). John Wiley & Sons. https://sites.comminfo.rutgers.edu/kgreene/wp-content/uploads/sites/28/2018/02/ACGreene-SPT.pdf

21. Emerging India Analytics. (2023, Juni 21). The Battle Of The Minds: Why AI Will Never Fully Replicate Human Emotions. *Medium*. https://medium.com/@analyticsemergingindia/the-battle-of-the-minds-why-ai-will-never-fully-replicate-human-emotions-db08bdeea61a

22. Martinez-Miranda, J., & Aldea, A. (2005). Emotions in human and artificial intelligence. *Computers in Human Behavior*, 323–341. https://doi.org/10.1016/j.chb.2004.02.010

23. Rothman, J. (2024, August 6). In the Age of A.I., What Makes People Unique? *The New Yorker*. https://www.newyorker.com/culture/open-questions/in-the-age-of-ai-what-makes-people-unique

24. Wu, J. (2024). Social and ethical impact of emotional AI advancement: The rise of pseudo-intimacy relationships and challenges in human interactions. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1410462

25. Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

26. Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

27. Elyoseph, Z., Refoua, E., Asraf, K., Lvovsky, M., Shimoni, Y., & Hadar-Shoval, D. (2024). Capacity of generative AI to interpret human emotions from visual and textual data: Pilot evaluation study. *JMIR Mental Health*, *11*, e54369. https://mental.jmir.org/2024/1/e54369

28. Montemayor, C., Halpern, J., & Fairweather, A. (2022). In principle obstacles for empathic AI: Why we can't replace human empathy in healthcare. *AI & SOCIETY*, *37*(4), 1353–1359. https://doi.org/10.1007/s00146-021-01230-z

29. Broekens, J., Hilpert, B., Verberne, S., Baraka, K., Gebhard, P., & Plaat, A. (2023). Fine-grained affective processing capabilities emerging from large language models. *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8. https://ieeexplore.ieee.org/abstract/document/10388177/

30. Tak, A. N., & Gratch, J. (2023). Is GPT a Computational Model of Emotion? *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8. https://doi.org/10.1109/ACII59096.2023.10388119

31. Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, *68*(4), 712–733. https://doi.org/10.1093/joc/jqy026

32. De Freitas, J., Agarwal, S., Schmitt, B., & Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour*, *7*(11), 1845–1854. https://doi.org/10.1038/s41562-023-01734-2

33. Stein, J.-P., Messingschlager, T., Gnambs, T., Hutmacher, F., & Appel, M. (2024). Attitudes towards AI: Measurement and associations with personality. *Scientific Reports*, *14*(1), 2909. https://doi.org/10.1038/s41598-024-53335-2

34. Brooks, D. (2023). Human Beings Are Soon Going to Be Eclipsed. *International New York Times*. https://www.nytimes.com/2023/07/13/opinion/ai-chatgpt-consciousness-hofstadter.html

35. Modhvadia, R., Sippy, T., Field Reid, O., & Margetts, H. (2025). *How Do People Feel About AI? Wave Two of a Nationally Representative Survey of UK Attitudes to AI Designed Through a Lens of Equity and Inclusion*. Ada Lovelace Institute and The Alan Turing Institute. https://attitudestoai.uk/

36. Aron, A., Melinat, E., Aron, E. N., Vallone, R. D., & Bator, R. J. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin*, *23*(4), 363–377. https://doi.org/10.1177/0146167297234003

37. Tönsing, D., Schiller, B., Vehlen, A., Spenthof, I., Domes, G., & Heinrichs, M. (2022). No evidence that gaze anxiety predicts gaze avoidance behavior during face-to-face social interaction. *Scientific Reports*, *12*(1), 21332. https://doi.org/10.1038/s41598-022-25189-z

38. Tönsing, D., Schiller, B., Vehlen, A., Nickel, K., Tebartz van Elst, L., Domes, G., & Heinrichs, M. (2025). Altered interactive dynamics of gaze behavior during face-to-face interaction in autistic individuals: A dual eye-tracking study. *Molecular Autism, 16*(1), 12. https://doi.org/10.1186/s13229-025-00645-5

39. Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22.* Austin, TX: University of Texas at Austin, 10. https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf

40. Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

41. Callaghan, D. E., Graff, M. G., & Davies, J. (2013). Revealing All: Misleading Self-Disclosure Rates in Laboratory-Based Online Research. *Cyberpsychology, Behavior, and Social Networking*, *16*(9), 690–694. https://doi.org/10.1089/cyber.2012.0399

42. Schiller, B., Tönsing, D., Kleinert, T., Böhm, R., & Heinrichs, M. (2022). Effects of the COVID-19 Pandemic Nationwide Lockdown on Mental Health, Environmental Concern, and Prejudice Against Other Social Groups. *Environment and Behavior*, *54*(2), 516–537. https://doi.org/10.1177/00139165211036991

43. Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

44. Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

45. Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

46. Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, *435*(7042), 673–676. https://doi.org/10.1038/nature03701

47. Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*(4), 596. https://doi.org/10.1037/0022-3514.63.4.596

48. Sprecher, S. (2014). Initial interactions online-text, online-audio, online-video, or face-to-face: Effects of modality on liking, closeness, and other interpersonal outcomes. *Computers in Human Behavior*, *31*, 190–197. https://doi.org/10.1016/j.chb.2013.10.029

49. Sprecher, S. (2021). Closeness and other affiliative outcomes generated from the Fast Friends procedure: A comparison with a small-talk task and unstructured self-disclosure and the moderating role of mode of communication. *Journal of Social and Personal Relationships*, *38*(5), 1452–1471. https://doi.org/10.1177/0265407521996055

50. Gächter, S., Starmer, C., & Tufano, F. (2015). Measuring the closeness of relationships: A comprehensive evaluation of the'inclusion of the other in the self'scale. *PloS One*, *10*(6), e0129478. https://doi.org/10.1371/journal.pone.0129478

51. Sidhu, D. M., & Pexman, P. M. (2015). What's in a name? Sound symbolism and gender in first names. *PloS one*, *10*(5), e0126809. https://doi.org/10.1371/journal.pone.0126809

52. Vedel, A. (2016). Big Five personality group differences across academic majors: A systematic review. *Personality and individual differences*, *92*, 1–10. https://doi.org/10.1016/j.paid.2015.12.011

53. Boothby, E. J., Smith, L. K., Clark, M. S., & Bargh, J. A. (2016). Psychological Distance Moderates the Amplification of Shared Experience. *Personality and Social Psychology Bulletin*, *42*(10), 1431–1444. https://doi.org/10.1177/0146167216662869

54. Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

55. Field, A., Field, Z., & Miles, J. (2012). Multilevel Linear Models. In *Discovering Statistics Using R* (pp. 855–909). SAGE Publications, Inc.

56. Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, *53*(6), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

57. Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

58. Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, *2*(1), 11. https://doi.org/10.9707/2307-0919.1116

59. Chaturvedi, R., Verma, S., Das, R., & Dwivedi, Y. K. (2023). Social companionship with artificial intelligence: Recent trends and future avenues. *Technological Forecasting and Social Change*, *193*, 122634. https://doi.org/10.1016/j.techfore.2023.122634

60. Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., & Klang, E. (2024). Large language models and empathy: systematic review. *Journal of Medical Internet Research*, 26, e52597. https://www.jmir.org/2024/1/e52597

61. Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I. P., Hari, R., & Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences*, *109*(24), 9599–9604. https://doi.org/10.1073/pnas.1206095109

62. Krämer, N. C., & Schäwel, J. (2020). Mastering the challenge of balancing self-disclosure and privacy in social media. *Current Opinion in Psychology*, *31*, 67–71. https://doi.org/10.1016/j.copsyc.2019.08.003

63. Vogel, D. L., & Wester, S. R. (2003). To seek help or not to seek help: The risks of self-disclosure. *Journal of Counseling Psychology*, *50*(3), 351. https://doi.org/10.1037/0022-0167.50.3.351

64. Esmaeilzadeh, P., Mirzaei, T., & Dharanikota, S. (2021). Patients' perceptions toward human–artificial intelligence interaction in health care: Experimental study. *Journal of Medical Internet Research*, *23*(11), e25856. https://www.jmir.org/2021/11/e25856

65. Rubin, M., Li, J. Z., Zimmerman, F., Ong, D. C., Goldenberg, A., & Perry, A. (2025). Comparing the Value of Perceived Human versus AI-Generated Empathy. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-025-02247-w

66. Machia, L. V., Corral, D., & Jakubiak, B. K. (2024). Social Need Fulfillment Model for Human–AI Relationships. *Technology, Mind, and Behavior, 5*(4). https://doi.org/10.1037/tmb0000141

67. Baumeister, R. F., & Leary, M. R. (2017). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Interpersonal Development*, 57–89. https://doi.org/10.1037/0033-2909.117.3.497

68. Holt-Lunstad, J., Smith, T. W., & Layton, J. B. (2010). Social Relationships and Mortality Risk: A Meta-Analytic Review. *Plos Medicine*. https://doi.org/10.1371/journal.pmed.1000316

69. Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., & Combs, B. (1978). How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sciences*, *9*(2), 127–152. https://doi.org/10.1007/BF00143739

70. Starr, C. (1969). Social Benefit versus Technological Risk: What is our society willing to pay for safety? *Science*, *165*(3899), 1232–1238. https://doi.org/10.1126/science.165.3899.1232

71. Jones, C. H., & Dolsten, M. (2024). Healthcare on the brink: Navigating the challenges of an aging society in the United States. *npj Aging*, *10*(1), 22. https://doi.org/10.1038/s41514-024-00148-2

72. Park, E.-Y. (2021). Meta-Analysis of Factors Associated with Occupational Therapist Burnout. *Occupational Therapy International*, *2021*, 1–10. https://doi.org/10.1155/2021/1226841

73. Phillips, L. (2023). *A closer look at the mental health provider shortage*. Www.Counseling.Org. https://www.counseling.org/publications/counseling-today-magazine/article-archive/article/legacy/a-closer-look-at-the-mental-health-provider-shortage

74. Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

75. Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019). Artificial Intelligence for Mental Health and Mental Illnesses: An Overview. *Current Psychiatry Reports*, *21*(11), 116. https://doi.org/10.1007/s11920-019-1094-0

76. Maurya, R. K., Montesinos, S., Bogomaz, M., & DeDiego, A. C. (2025). Assessing the use of ChatGPT as a psychoeducational tool for mental health practice. *Counselling and Psychotherapy Research*, *25*(1), e12759. https://doi.org/10.1002/capr.12759

77. Wang, A., Qian, Z., Briggs, L., Cole, A. P., Reis, L. O., & Trinh, Q.-D. (2023). The Use of Chatbots in Oncological Care: A Narrative Review. *International Journal of General Medicine*, *Volume 16*, 1591–1602. https://doi.org/10.2147/IJGM.S408208

78. Yang, Y., Wang, C., Xiang, X., & An, R. (2025). AI Applications to Reduce Loneliness Among Older Adults: A Systematic Review of Effectiveness and Technologies. *Healthcare*, *13*(5), 446. https://www.mdpi.com/2227-9032/13/5/446

79. Aktan, M. E., Turhan, Z., & Dolu, I. (2022). Attitudes and perspectives towards the preferences for artificial intelligence in psychotherapy. *Computers in Human Behavior*, *133*, 107273. https://doi.org/10.1016/j.chb.2022.107273

80. Hu, B., Mao, Y., & Kim, K. J. (2023). How social anxiety leads to problematic use of conversational AI: The roles of loneliness, rumination, and mind perception. *Computers in Human Behavior*, *145*, 107760. https://doi.org/10.1016/j.chb.2023.107760

81. Marriott, H. R., & Pitardi, V. (2024). One is the loneliest number… Two can be as bad as one. The influence of AI Friendship Apps on users' well-being and addiction. *Psychology & Marketing*, *41*(1), 86–101. https://doi.org/10.1002/mar.21899

82. Sedlakova, J., & Trachsel, M. (2023). Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent? *The American Journal of Bioethics*, *23*(5), 4–13. https://doi.org/10.1080/15265161.2022.2048739

83. Roose, K. (2024, Oktober 23). Can A.I. Be Blamed for a Teen's Suicide? *The New York Times*. https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html

84. Heizmann, C., Gleim, P., & Kellmeyer, P. (2025). Participatory Co-Creation of an AI-Supported Patient Information System: A Multi-Method Qualitative Study. *Studies in health technology and informatics*, *327*, 338–342.

85. Huang, C., Zhang, Z., Mao, B., & Yao, X. (2022). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, *4*(4), 799–819. https://ieeexplore.ieee.org/document/9844014

86. Kim, J. K., Chua, M., Rickard, M., & Lorenzo, A. (2023). ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in

academic medicine. *Journal of Pediatric Urology*, *19*(5), 598–604. https://doi.org/10.1016/j.jpurol.2023.05.018

87. Ognibene, D., Wilkens, R., Taibi, D., Hernández-Leo, D., Kruschwitz, U., Donabauer, G., Theophilou, E., Lomonaco, F., Bursic, S., & Lobo, R. A. (2023). Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. *Frontiers in Artificial Intelligence*, *5*, 654930. https://doi.org/10.3389/frai.2022.654930

88. Silva, G. R. S., & Canedo, E. D. (2024). Towards User-Centric Guidelines for Chatbot Conversational Design. *International Journal of Human–Computer Interaction*, *40*(2), 98–120. https://doi.org/10.1080/10447318.2022.2118244

89. Rao, S., Verma, A. K., & Bhatia, T. (2021). A review on social spam detection: Challenges, open issues, and future directions. *Expert Systems with Applications*, *186*, 115742. https://doi.org/10.1016/j.eswa.2021.115742

90. Feuerriegel, S., DiResta, R., Goldstein, J. A., Kumar, S., Lorenz-Spreen, P., Tomz, M., & Pröllochs, N. (2023). Research can help to tackle AI-generated disinformation. *Nature Human Behaviour*, *7*(11), 1818–1821. https://doi.org/10.1038/s41562-023-01726-2

91. Gumusel, E. (2025). A literature review of user privacy concerns in conversational chatbots: A social informatics approach: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, *76*(1), 121–154. https://doi.org/10.1002/asi.24898

92. Köbis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, *5*(6), 679–685. https://doi.org/10.1038/s41562-021-01128-2

93. Yuste, R., Goering, S., Arcas, B. A. Y., Bi, G., Carmena, J. M., Carter, A., Fins, J. J., Friesen, P., Gallant, J., & Huggins, J. E. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*, *551*(7679), 159–163. https://doi.org/10.1038/551159a

94. André, E., & Pelachaud, C. (2010). Interacting with Embodied Conversational Agents. In F. Chen & K. Jokinen (Hrsg.), *Speech Technology* (pp. 123–149). Springer US. https://doi.org/10.1007/978-0-387-73819-2_8

95. Sundar, A., Russell-Rose, T., Kruschwitz, U., & Machleit, K. (2024). The AI Interface: Designing for the Ideal Machine-Human Experience. *Computers in Human Behavior*, 165, 108539. Elsevier. https://doi.org/10.1016/j.chb.2024.108539

96. Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://ieeexplore.ieee.org/document/6213238

97. Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a

98. Schiller, B., Kleinert, T., Teige-Mocigemba, S., Klauer, K. C., & Heinrichs, M. (2020). Temporal dynamics of resting EEG networks are associated with prosociality. *Scientific Reports*, *10*(1), 13066. https://doi.org/10.1038/s41598-020-69999-5

99. Kleinert, T., Nash, K., Leota, J., Koenig, T., Heinrichs, M., & Schiller, B. (2022). A self-controlled mind is reflected by stable mental processing. *Psychological Science*, *33*(12), 2123–2137. https://doi.org/10.1177/09567976221110

100. Kleinert, T., & Nash, K. (2022). Trait Aggression is Reflected by a Lower Temporal Stability of EEG Resting Networks. *Brain Topography*, 1–10. https://doi.org/10.1007/s10548-022-00929-6

101. Nash, K., Kleinert, T., Leota, J., Scott, A., & Schimel, J. (2022). Resting-state networks of believers and non-believers: An EEG microstate study. *Biological Psychology*, *169*, 108283. https://doi.org/10.1016/j.biopsycho.2022.108283

102. Nash, K., Leota, J., Kleinert, T., & Hayward, D. A. (2023). Anxiety disrupts performance monitoring: Integrating behavioral, event-related potential, EEG microstate, and sLORETA evidence. *Cerebral Cortex*, *33*(7), 3787–3802. https://doi.org/10.1093/cercor/bhac307

103. Schiller, B., Sperl, M. F. J., Kleinert, T., Nash, K., & Gianotti, L. R. R. (2023). EEG Microstates in Social and Affective Neuroscience. *Brain Topography*. https://doi.org/10.1007/s10548-023-00987-4

104. Fiske, A. P., & Fiske, S. T. (2007). Social relationships in our species and cultures. *Handbook of Cultural Psycholog* (pp. 283–306). The Guildford Press.

105. Holt-Lunstad, J. (2018). Why Social Relationships Are Important for Physical Health: A Systems Approach to Understanding and Modifying Risk and Protection. *Annual Review of Psychology*, *69*(1), 437–458. https://doi.org/10.1146/annurev-psych-122216-011902

106. Hostinar, C. E., & Gunnar, M. R. (2018). Future directions in the study of social relationships as regulators of the HPA axis across development. *Future Work in Clinical Child and Adolescent Psychology*, 333–344. https://doi.org/10.1080/15374416.2013.804387

107. Løseth, G. E., Eikemo, M., Trøstheim, M., Meier, I. M., Bjørnstad, H., Asratian, A., Pazmandi, C., Tangen, V. W., Heilig, M., & Leknes, S. (2022). Stress recovery with social support: A dyadic stress and support task. *Psychoneuroendocrinology*, *146*, 105949. https://doi.org/10.1016/j.psyneuen.2022.105949

108. Spengler, F. B., Scheele, D., Marsh, N., Kofferath, C., Flach, A., Schwarz, S., Stoffel-Wagner, B., Maier, W., & Hurlemann, R. (2017). Oxytocin facilitates reciprocity in social communication. *Social Cognitive and Affective Neuroscience*, *12*(8), 1325–1333. 10.1093/scan/nsx061

Figure captions and notes

Figure 1

AI-generated content leads to higher levels of interpersonal closeness than human-generated content in deep-talk interactions

[FIGURE 1]

Note. * p < .05, ** p < .010, *** p < .001. A: Violin plots of perceived interpersonal closeness in the four different conditions of study 1 (*n* = 322). There was significantly higher interpersonal closeness following interactions with AI compared to interactions with humans within deep-talk interactions. B: Violin plots of self-disclosure (as measured via the LIWC-22) displayed by the six human and the six AI interaction partners within the deep-talk condition (*n* = 12). AI interaction partners showed significantly more self-disclosure. C: Scatterplot showing a statistically significant positive association between the partner's self-disclosure and interpersonal closeness (*n* = 164). Participants felt closer to their partner when the partner showed more self-disclosure. D: Violin plots of participants' own self-disclosure when interacting with a human or with an AI (*n* = 164). Participants showed significantly more self-disclosure when interacting with an AI. E: Scatterplot showing a statistically significant positive association between the partner's self-disclosure and participants' own self-disclosure (*n* = 164). Participants showed more self-disclosure when their partner showed more self-disclosure. This suggests that participants tend to disclose more personal information to AI interaction partners in response to the AI's higher level of self-disclosure, indicating a reciprocal effect. All violin plots include individual data points, means and standard deviations. All scatterplots include individual data points, regression slopes, and 95% confidence intervals.

Figure 2

The anti AI-bias

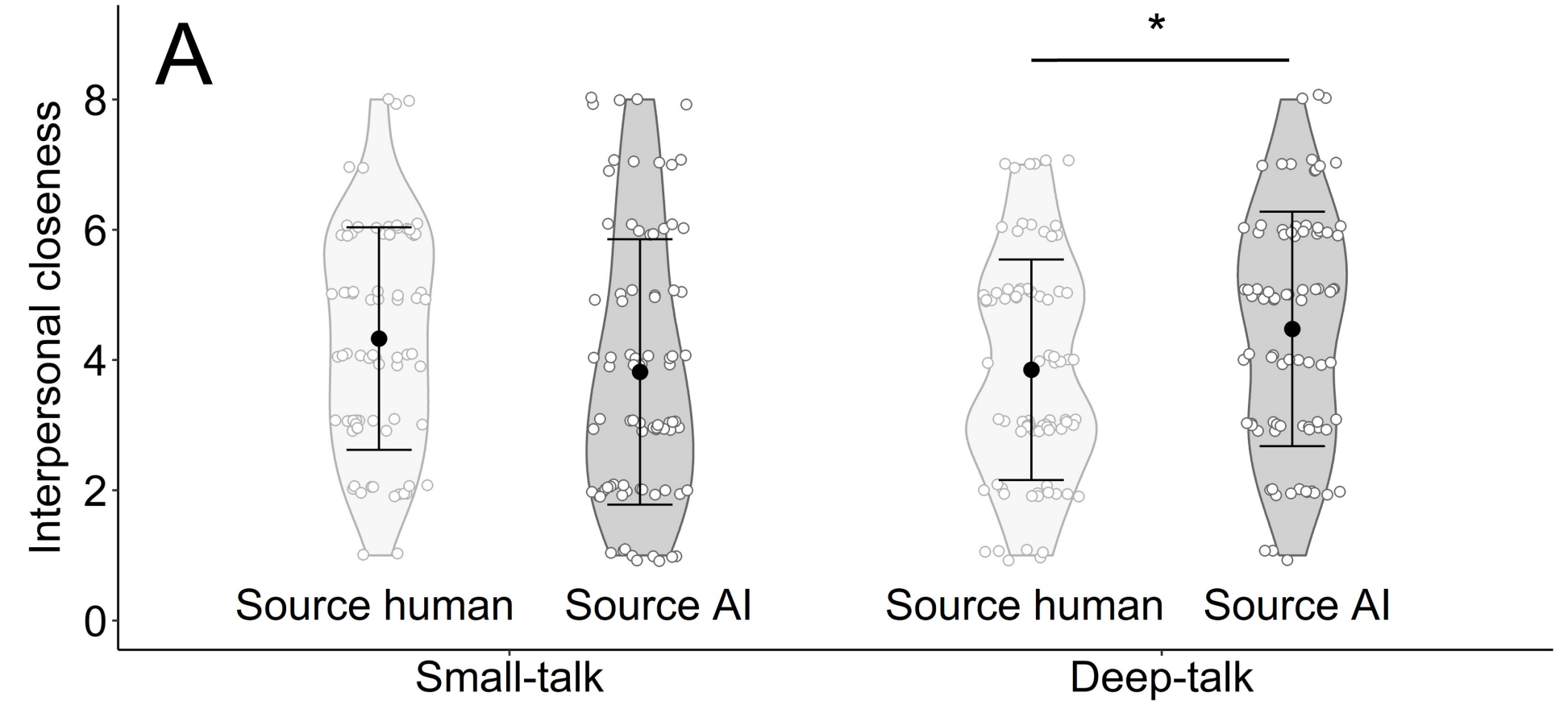
[FIGURE 2]

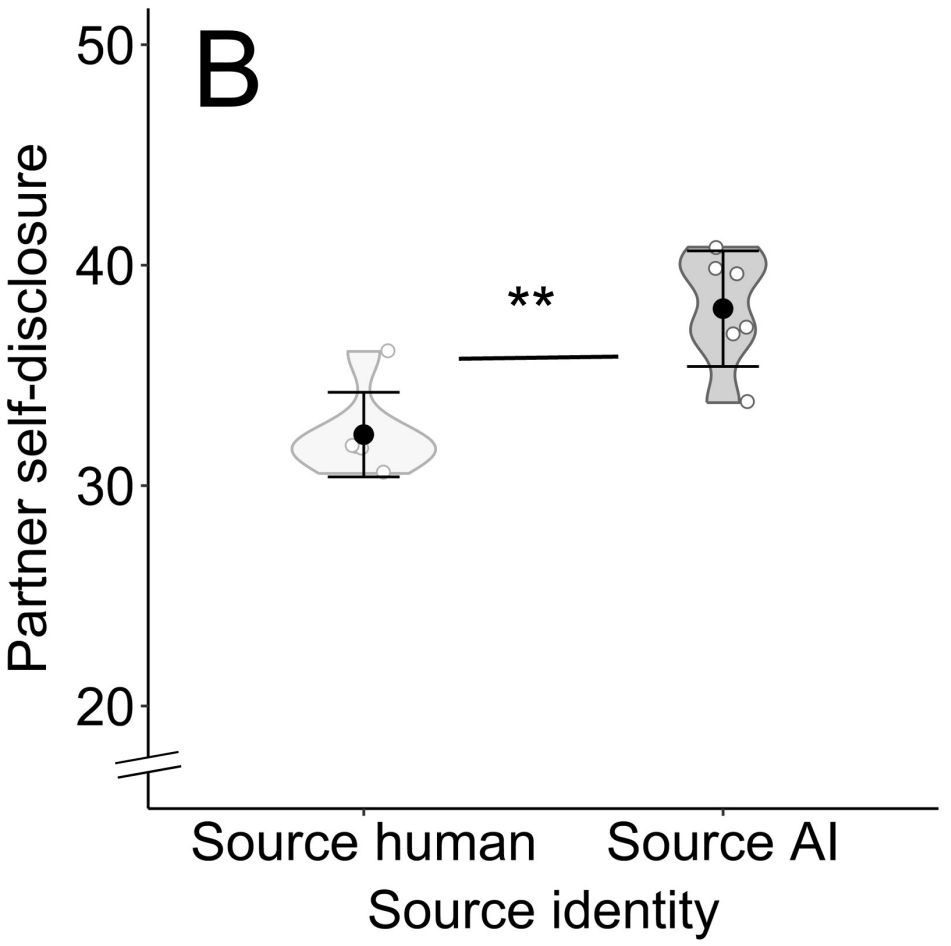
Note. * p < .05, ** p < .010, *** p < .001. A: Violin plots of perceived interpersonal closeness following AI- and human-labelled interactions (*n* = 334). Participants who believed they would interact with an AI reported significantly lower interpersonal closeness than participants who believed they would interact with a human. B: Violin plots of the participants' response length in AI- and human-labelled interactions (*n* = 334). Participants who believed they would interact with an AI wrote significantly shorter responses than participants who believed they would interact with a human. All violin plots include individual data points, means and standard deviations. C: Scatterplot with individual data points, regression slope, and 95% confidence interval, illustrating a statistically significant positive association between the participants' response length and perceived interpersonal closeness (*n* = 334). Participants who wrote longer responses felt closer to their interaction partner.

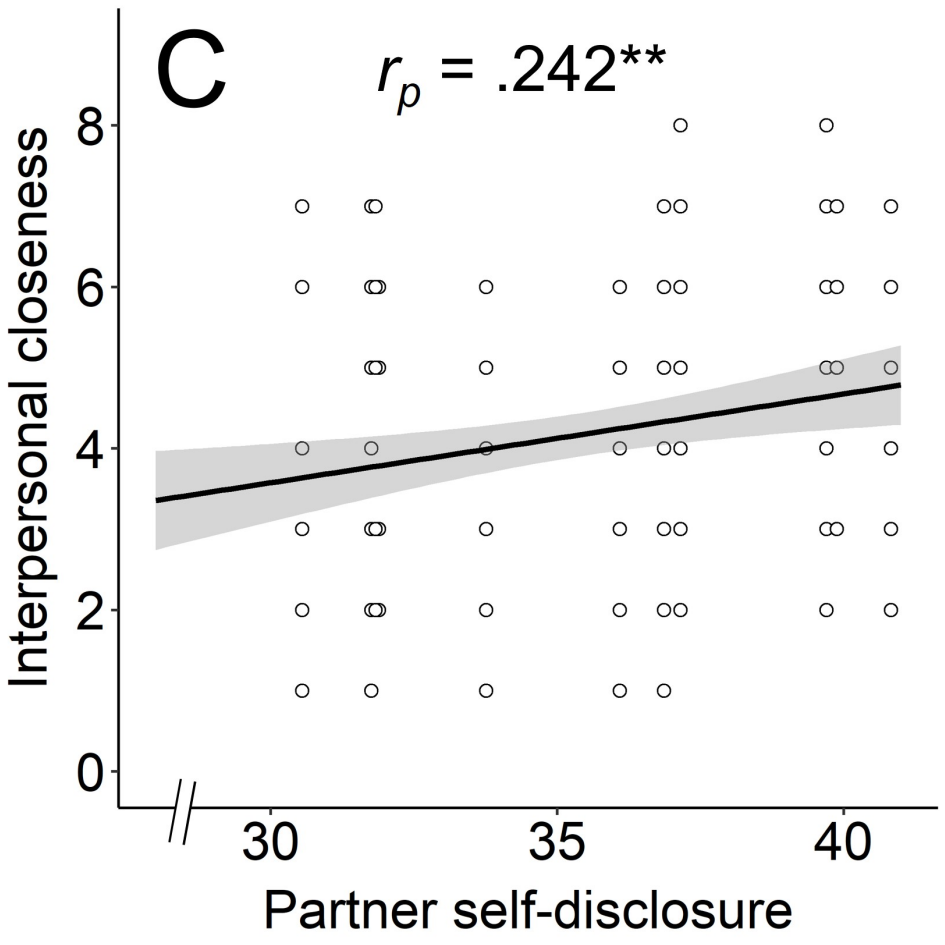
Editorial Summary:

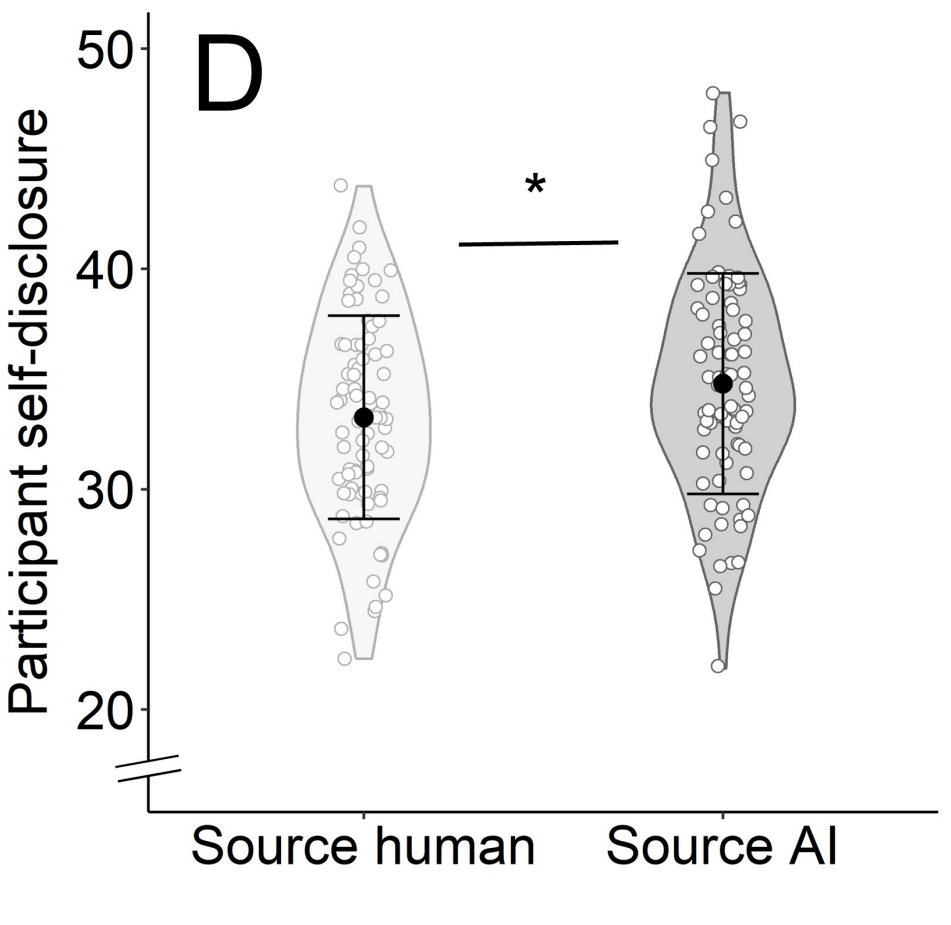
Groups of two or four chimpanzees encountered a collective resource sustainability problem. Quartets avoided resource collapse for longer than dyads, with group social tolerance playing a supportive role.

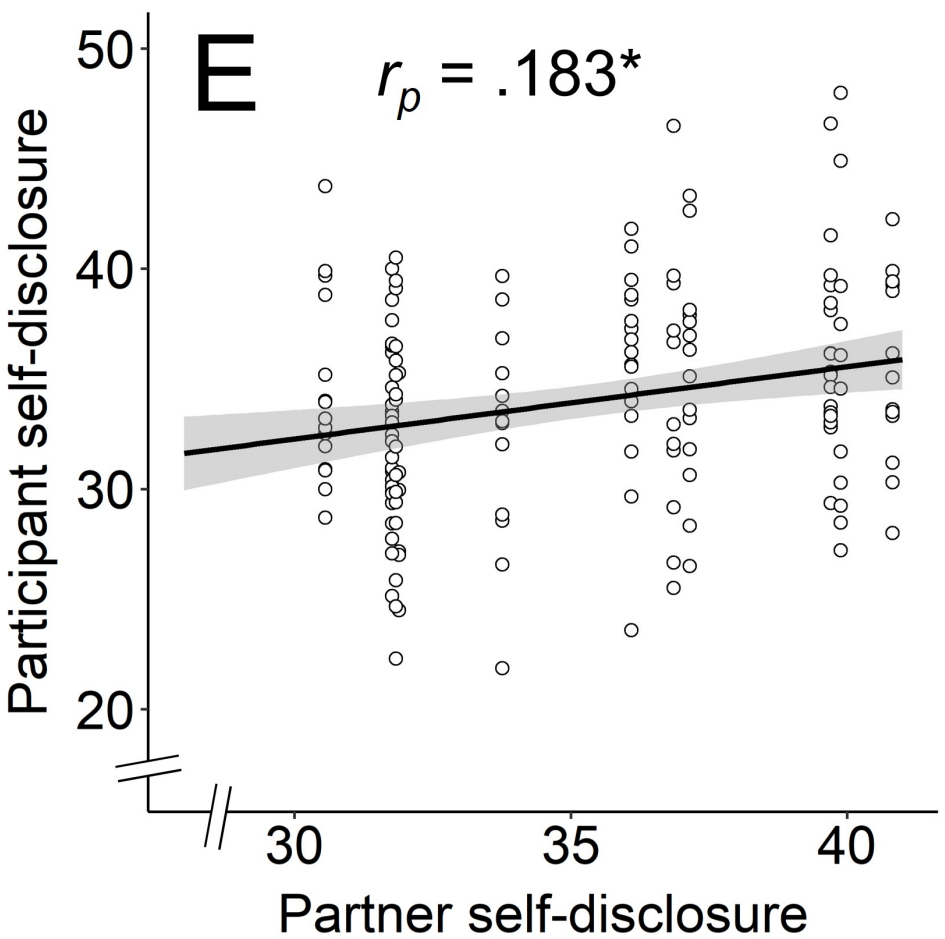
Peer Review:
Communications Psychology thanks Rebecca Koomen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Troby Ka-Yan Lui. A peer review file is available.

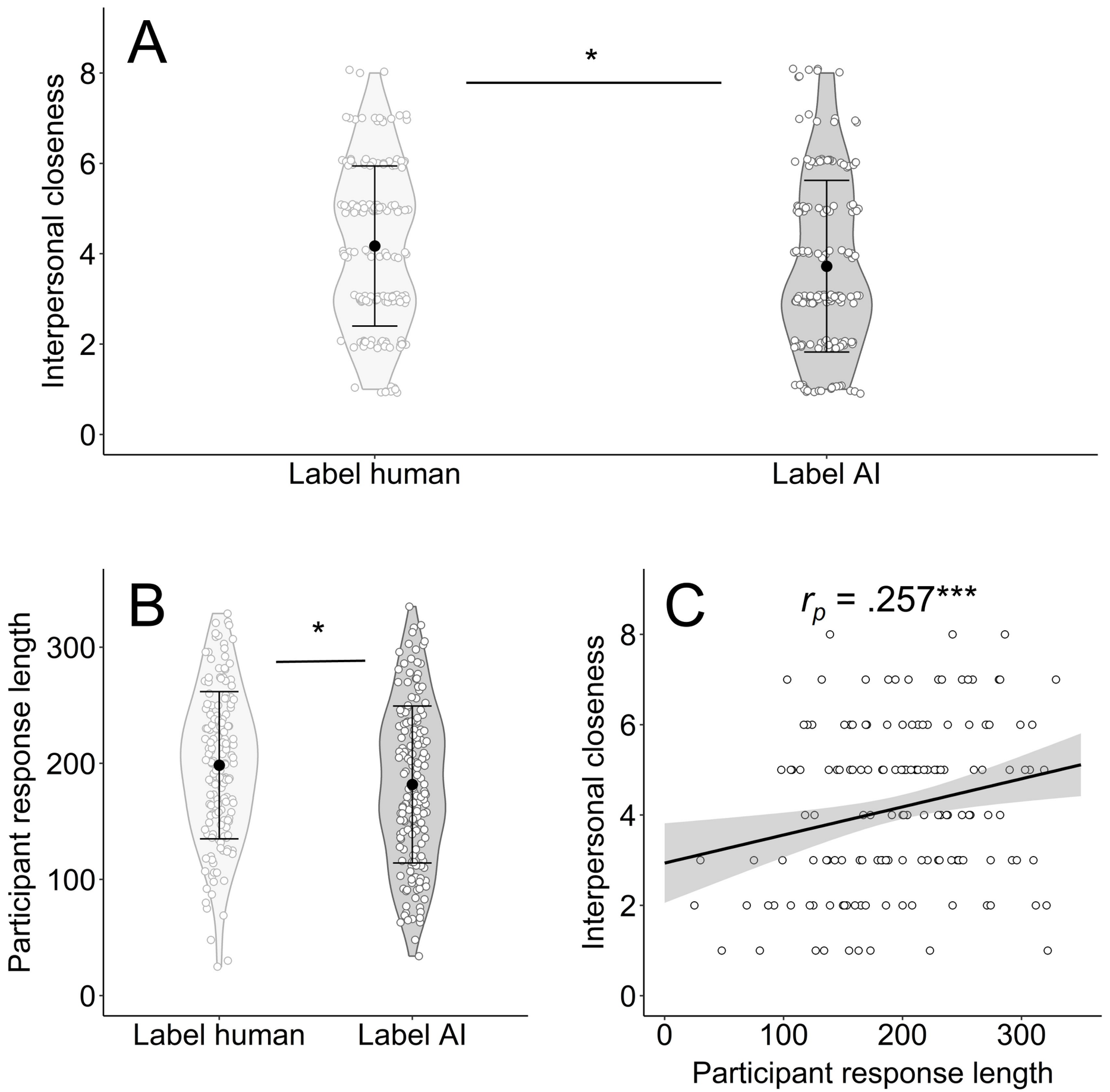

A
Interpersonal closeness
0
2
4
6
8
*
Source human
Source AI
Source human
Source AI
Small-talk
Deep-talk


B
Partner self-disclosure
20
30
40
50
**
Source human
Source AI
Source identity


C
r_p = .242**
Interpersonal closeness
0
2
4
6
8
30
35
40
Partner self-disclosure


D
Participant self-disclosure
20
30
40
50
*
Source human
Source AI


E
r_p = .183*
Participant self-disclosure
20
30
40
50
30
35
40
Partner self-disclosure


A
*
Interpersonal closeness
0
2
4
6
8
Label human
Label AI
B
*
Participant response length
0
100
200
300
Label human
Label AI
C
r_p = .257***
Interpersonal closeness
0
2
4
6
8
0
100
200
300
Participant response length