nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02247-w

Comparing the value of perceived human versus AI-generated empathy

Received: 10 July 2024

Accepted: 14 May 2025

Published online: 30 June 2025

Check for updates

Matan Rubin[1] ✉, Joanna Z. Li[2,3], Federico Zimmerman[2,3], Desmond C. Ong[4], Amit Goldenberg[2,3,5] & Anat Perry[1,5] ✉

Artificial intelligence (AI) and specifically large language models demonstrate remarkable social–emotional abilities, which may improve human–AI interactions and AI's emotional support capabilities. However, it remains unclear whether empathy, encompassing understanding, 'feeling with' and caring, is perceived differently when attributed to AI versus humans. We conducted nine studies (n = 6,282) where AI-generated empathic responses to participants' emotional situations were labelled as provided by either humans or AI. Human-attributed responses were rated as more empathic and supportive, and elicited more positive and fewer negative emotions, than AI-attributed ones. Moreover, participants' own uninstructed belief that AI had aided the human-attributed responses reduced perceived empathy and support. These effects were replicated across varying response lengths, delays, iterations and large language models and were primarily driven by responses emphasizing emotional sharing and care. Additionally, people consistently chose human interaction over AI when seeking emotional engagement. These findings advance our general understanding of empathy, and specifically human–AI empathic interactions.

Artificial intelligence (AI) and particularly large language models (LLMs) are increasingly involved in our social lives. AI chatbots already demonstrate impressive abilities, including language analyses, translation, creative writing and problem solving[1–4]. Perhaps more intriguing, these models show remarkable abilities to engage with social–emotional situations and to mimic expressions of interpersonal emotional reactions, such as empathy, compassion and care[5,6]. Since the release of these models, researchers and industrial companies have explored using LLMs' empathic responses to provide social, emotional and peer support in medicine, mental health and daily life[7–11]. At the extreme, Replika, an app for building virtual friendships with AI avatars, promises an 'AI companion who cares', is 'always on your side' and is 'always ready to chat when you need an empathetic friend'[12,13].

Empathy is a multifaceted construct that includes a cognitive component (understanding another's emotional state), an affective component ('feeling with' another) and a motivational component (feeling concern for another with a willingness to invest effort in improving their well-being)[14]. When successful, human empathy offers substantial benefits. Sharing another's pain or joy fosters a profound sense of mutual understanding and togetherness, associated with physiological and neural synchrony[14,15]. It enhances resilience[16] and serves as a catalyst for prosocial and altruistic behaviour in humans[17–20] as well as other mammals[21,22]. However, empathy is also hard work[23]—people often do not succeed in taking the other's perspective[24], fail to empathize when fatigued or burned out[25–29] and frequently avoid empathizing due to empathy's taxing nature[20,23,30].

In contrast to humans, LLMs can accurately capture our emotional situation by analysing patterns and associations of phrases, and they do so instantly, without tiring or being preoccupied. These models show impressive abilities to infer emotional states[31–35]; and,

[1]Psychology Department, Hebrew University of Jerusalem, Jerusalem, Israel. [2]Harvard Business School, Harvard University, Boston, MA, USA. [3]Digital, Data, and Design Institute, Harvard University, Boston, MA, USA. [4]Psychology Department, The University of Texas at Austin, Austin, TX, USA. [5]These authors contributed equally: Amit Goldenberg, Anat Perry. ✉e-mail: Matan.rubin@mail.huji.ac.il; Anat.perry@mail.huji.ac.il

from a third-person perspective, AI-generated expressions are rated as more empathic than human ones[36–39]. AI can also be experienced as empathic in a first-person interaction, at least in the short term, if presented as caring[40]. However, although AI-generated responses can make people 'feel heard', this effect is diminished when AI involvement is disclosed[8,41]. Similar claims have been made in the domain of AI-mediated communication, revealing that when AI facilitates human communication it is perceived as responsible for unsuccessful communication[42]. Moreover, AI mediation of and intervention in human communication are not always acceptable[43], may reduce positivity and social attraction towards the human partner[44,45], and raise concerns about deception and inauthenticity in the communication[46,47]. In sum, realizing that an emotional response was AI-generated can reduce its social–emotional benefits[8,41].

This may be because, despite expressing empathy, current AI models fundamentally lack the capacity to truly 'feel with' us[48–53]. They share neither our pain nor our joy, and their response does not require any effort (mental, emotional, in time or space). In short, AIs do not genuinely care[52,53]. This was well phrased by philosopher John Haugeland: 'The trouble with artificial intelligence is that computers don't give a damn'[54]. Given these distinctions, a fundamental question arises: do we indeed value human empathy more than empathy generated by AI, and perhaps more importantly, what dimensions of empathy are valued more in their human form, and why? In other words, what motivates and influences a preference for human empathy? With greater knowledge of how individuals perceive empathy on the basis of its source, and also considering which dimensions of empathy—understanding, 'feeling with' or caring—influence a preference for human empathy and what motivates such a preference, we can better elucidate the boundaries of AI capabilities while also advancing our understanding of human empathy and the qualities we value in human social connections.

The current project

We conducted nine studies (n = 6,282) to examine whether AI-generated responses are perceived as more empathic when thought to be written by a human as opposed to an AI. In the first set of studies (1a–1d), we established the phenomena, hypothesizing that labelling an AI-generated response as produced by AI versus a human would lead to a reduction in perceived empathy (study 1a), a reduction in feelings of support and positive emotions, and a rise in negative emotions (study 1b). We also examined how these measures are influenced by the participants' belief that the AI responses were edited by humans and the human responses were aided by AI. We replicated the effects with a different, open-source, LLM (study 1c) and with a four-turn continuous interaction (study 1d).

The second set of studies was designed to address an alternative explanation for the differences found in studies 1a–1d. Studies 2a and 2b aimed to determine whether high-quality AI-generated responses led to a 'halo effect'. High-quality responses may have emphasized the differences between the AI and human labels, either because participants were then particularly disappointed to learn that the response was produced by AI, or because they were positively surprised that it was actually a person who produced such a high-quality response, and not because of actual differences in perceived empathy. In study 2a, we examined whether differences between conditions remained even when participants received simple, brief responses. In study 2b, we explored whether differences in conditions were sustained with an increased waiting time for a response.

Study 3 was designed to further understand why perceived human responses are rated higher than AI ones—specifically, whether these differences are dependent on particular aspects of empathy (cognitive, affective or motivational). To achieve this, we prompted the AI to separately emphasize understanding, emotional sharing or caring. We hypothesized that human empathy would be valued more when emotional sharing or caring was stressed (study 3).

After we observed that people value human empathy more than artificial empathy, we aimed to further examine how much this difference mattered to people and why. We did this by giving participants a choice: either wait for a human to respond to their experience or receive an immediate response from an AI (study 4). We expanded this choice in Study 5 by offering participants the option to either wait for a human to simply read about their emotional experience or to receive an immediate response from an AI (study 5). We then asked the participants to explain their preference. These studies aimed to find the differences in value between artificial empathy, actual human empathy and just being heard. Together, these results shed light on what we value in human empathy and offer insights into the underlying motivations to receive an empathic human response.

Results

Studies 1a–1d perceived source affects perceived empathy

In study 1a, we tested whether people would perceive the empathy of a response differently if they thought the response was written by an AI or a human. We told the participants they were paired with either an AI or another participant. The participants shared a recent emotional experience and waited for 60 seconds; they were told either that the AI was generating a response or that the other participant was writing one, depending on the experimental condition. Participants in both conditions were then shown an AI-generated response, with the AI having been prompted to respond to their specific experience and include all three aspects of empathy (cognitive, affective and motivational; see Supplementary Information section 13 for the specific prompts). The only difference was the information on the source of the response. The participants then rated how much empathy they felt, using both a single-item scale ('general empathy') and a 15-item scale that described specific aspects of empathy (cognitive, affective and motivational empathy; Methods). Lastly, the participants answered three questions rating positivity resonance in the interaction[55] (see Methods for specific questions); these answers were averaged to score positivity resonance, a measure of synchrony and positivity in the interaction.

Study 1b was designed to replicate study 1a, with one change to the procedure and several additional measures. In study 1a, participants' sharing could have been influenced by whether they thought what they wrote would be read by an AI or a human; and so in study 1b, we had the participants first share an emotional experience before being told whether they were paired with a human or an AI. Even with this more conservative design, we predicted that we would replicate the results from study 1a. We included additional questions as well: the participants rated the levels of positive and negative emotions the response elicited, the degree to which they felt the response was helpful and could help in the future, and their desire to keep conversing. The final addition to study 1b was asking participants in the human condition to report the degree to which they believed the response was aided by an AI, and participants in the AI condition to report the degree to which they believed it was edited by a human.

In both studies, when participants were told the response was human, they found the response to be more empathic (study 1a: Welch's t-test on the general empathy question: $t_{685.77} = -4.56$; $P < 0.001$; Cohen's $d = 0.34$; mean difference, −0.40; 95% confidence interval (CI) for difference, (−0.57, −0.23); study 1b: $t_{552.95} = -4.36$; $P < 0.001$; Cohen's $d = 0.36$; mean difference, −0.45; 95% CI, (−0.65, −0.24); Fig. 1a,c). To test whether the manipulation affected the various aspects of empathy differently (cognitive, affective and motivational), we then fitted a linear mixed-effect model to predict empathy with condition, aspect of empathy (cognitive, affective or motivational) and their interaction. The results in both studies showed a significant main effect for condition (study 1a: $F_{1,694.36} = 29.74$, $P < 0.001$, partial $\eta^2 = 0.04$; study 1b: $F_{1,567.53} = 51.31$, $P < 0.001$, partial $\eta^2 = 0.08$). In study 1a, we found no significant effect of aspects of empathy ($P = 0.77$) or interactions

https://doi.org/10.1038/s41562-025-02247-w

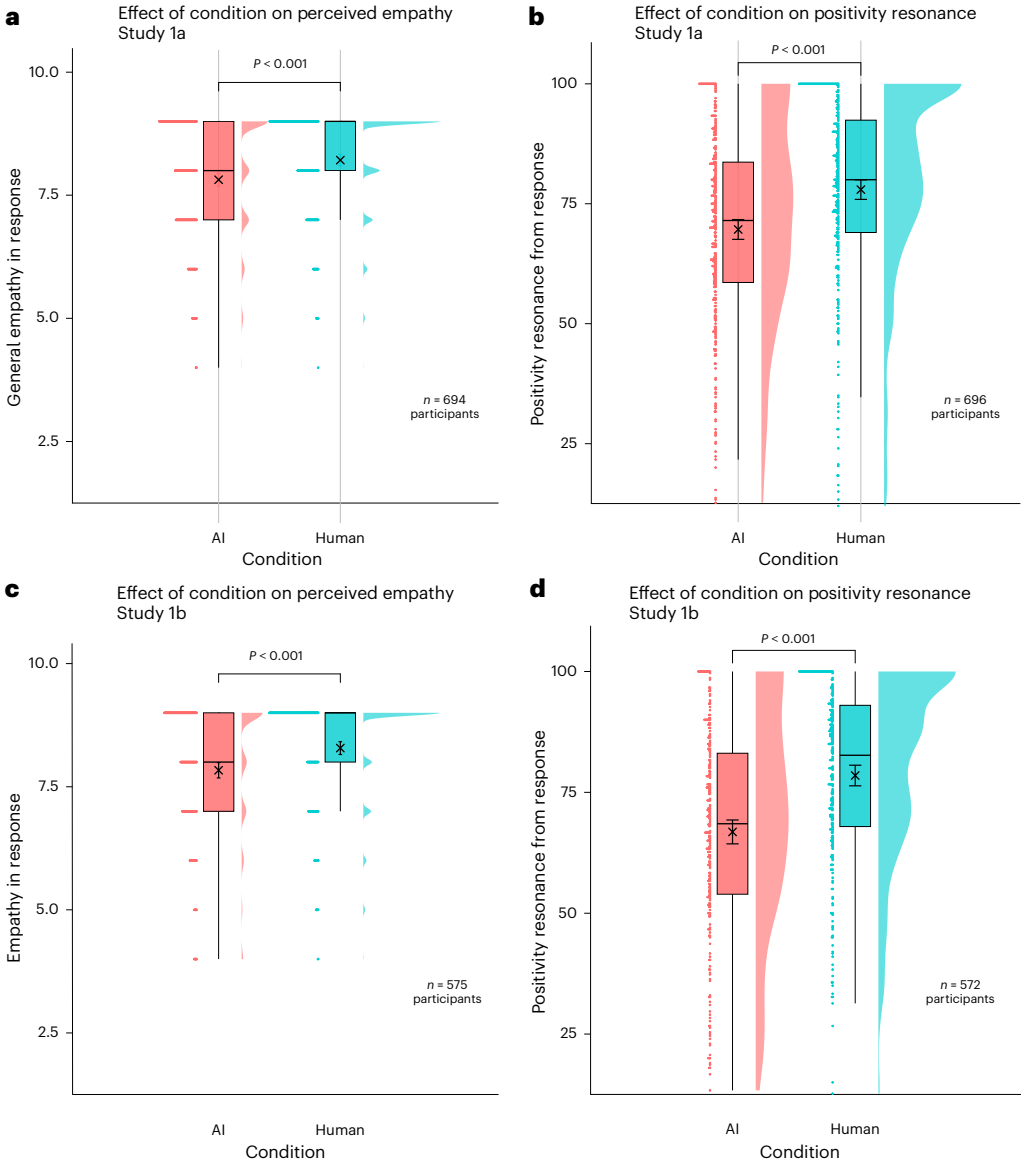

Fig. 1 | Effects of condition on perceived empathy and positivity resonance. **a**–**d**, Two-sided *t*-tests revealed significant differences between conditions (belief that one is conversing with an AI versus with a human) in perceived empathy ($t_{685.77} = -4.56$; $P < 0.001$; Cohen's $d = 0.34$, 95% CI for difference, (−0.57, −0.23)) (**a**) and positivity resonance ($t_{694} = -5.63$; $P < 0.001$; Cohen's $d = 0.43$; 95% CI, (−11.20, −5.41)) (**b**) in study 1a, and in perceived empathy ($t_{552.95} = -4.36$; $P < 0.001$; Cohen's $d = 0.36$; 95% CI, (−0.65, −0.24)) (**c**) and positivity resonance ($t_{556.99} = -7.04$; $P < 0.001$; Cohen's $d = 0.59$; 95% CI, (−14.92, −8.41)) (**d**) in study 1b. Within each box, the X marks the mean, with error bars that mark the 95% CI around the mean. The central line marks the median, the box shows the 75% interquartile range (IQR) and the whiskers indicate the total range.

($P = 0.46$), suggesting that people attributed lower levels of all three aspects to responses labelled as coming from an AI. In study 1b, however, there was a significant interaction between condition and aspect of empathy ($F_{2,1126.97} = 4.09$, $P = 0.02$, partial $\eta^2 = 0.007$). Post hoc contrasts revealed that the difference between the conditions was significantly smaller for cognitive empathy than for affective empathy ($t_{1147} = 2.82$; s.e. = 0.06; $P = 0.02$; $\beta = 0.16$; 95% CI, (0.02, 0.29)). In other words, the belief that they were communicating with a human increased participants' perception of affective empathy more than their perception of cognitive empathy in the response.

We also observed a significant difference between conditions in positivity resonance (study 1a: $t_{694} = -5.63$; $P < 0.001$; Cohen's $d = 0.43$; mean difference, −8.31; 95% CI, (−11.20, −5.41); Fig. 1b; study 1b: $t_{556.99} = -7.04$; $P < 0.001$; Cohen's $d = 0.59$; mean difference, −11.67; 95% CI, (−14.92, −8.41); Fig. 1d), such that participants who believed they were conversing with a human rated their bond as more positive.

Study 1b also showed that responses perceived to be from a human source elicited more positive emotions ($t_{567.47} = -4.71$; $P < 0.001$; Cohen's $d = 0.39$; mean difference, −0.66; 95% CI, (−0.94, −0.38); Fig. 2a) and fewer negative emotions ($t_{504.58} = 5.67$; $P < 0.001$; Cohen's $d = -0.47$; mean difference, 0.34; 95% CI, (0.22, 0.46); Fig. 2b). Perceived human responses were also deemed more authentic ($t_{574.75} = -2.81$; $P = 0.005$; Cohen's $d = 0.23$; mean difference, −0.21; 95% CI, (−0.36, −0.06); Fig. 2c). There were even greater differences in the interpersonal value of the experience between conditions. Questions on helpfulness, ability to help in the future and desire to keep conversing showed excellent internal consistency ($r = 0.78$–0.85, all $P < 0.001$, Cronbach's $\alpha = 0.93$) and were averaged into a measure called Support. A *t*-test revealed significant differences in Support between conditions, such that participants felt more support following a perceived human response ($t_{583.37} = -3.30$; $P = 0.001$; Cohen's $d = 0.27$; mean difference, −0.65; 95% CI, (−1.03, −0.26); Fig. 2d).

https://doi.org/10.1038/s41562-025-02247-w

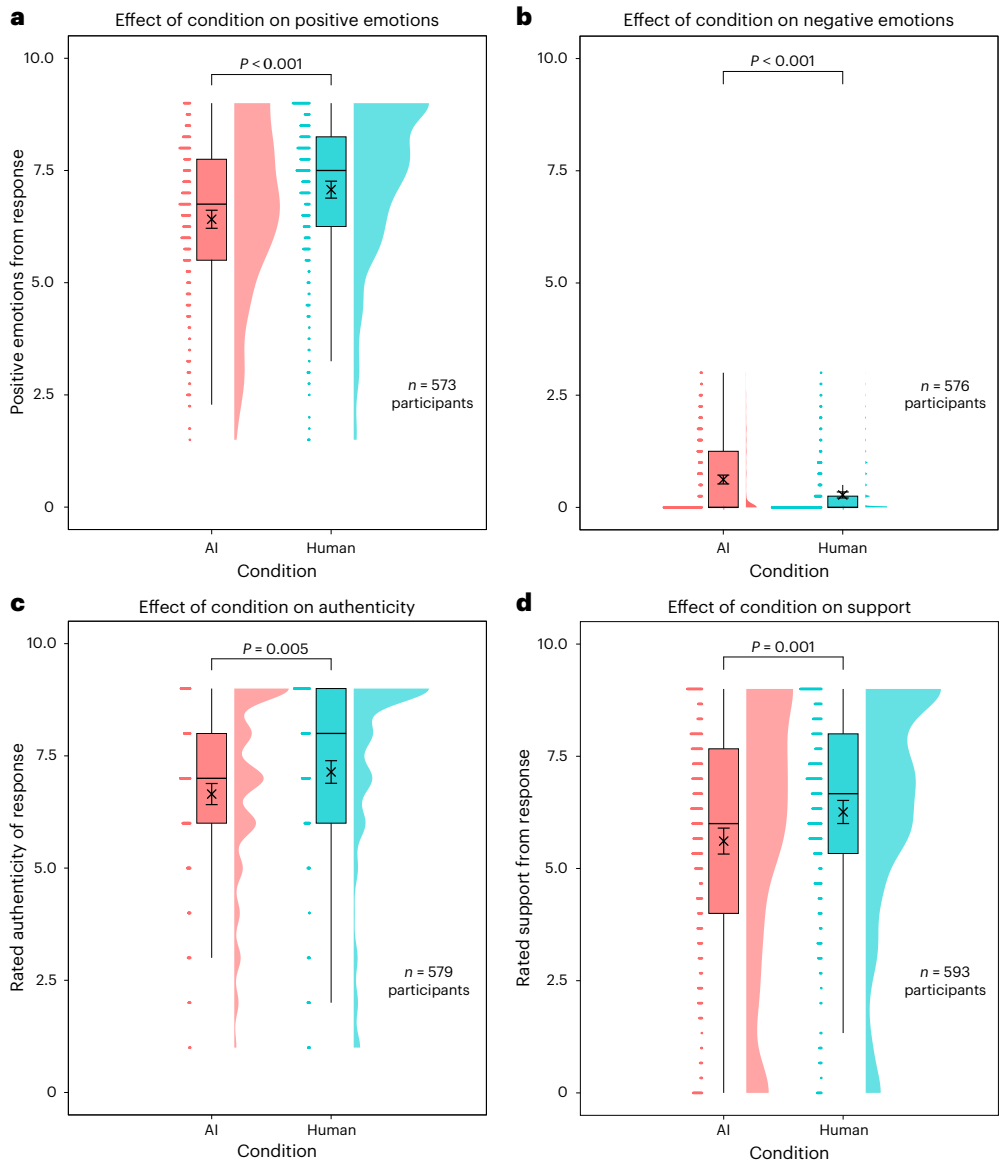

Fig. 2 | Effects of condition on emotions, authenticity and support. **a**–**d**, Two-sided *t*-tests revealed significant differences between conditions in positive emotions ($t_{567.47} = -4.71$; $P < 0.001$; Cohen's $d = 0.39$; 95% CI, (−0.94, −0.38)) (**a**), negative emotions ($t_{504.58} = 5.67$; $P < 0.001$; Cohen's $d = -0.47$; 95% CI, (0.22, 0.46)) (**b**), authenticity ($t_{574.75} = -2.81$; $P = 0.005$; Cohen's $d = 0.23$; 95% CI, (−0.36, −0.06)) (**c**) and support ($t_{583.37} = -3.30$; $P = 0.001$; Cohen's $d = 0.27$; 95% CI, (−1.03 −0.26)) (**d**). In all measures, the affective experience when the response was perceived to be human was rated more positively than when it was perceived to be AI-generated. Within each box, the X marks the mean, with error bars that mark the 95% CI around the mean. The central line marks the median, the box shows the 75% IQR and the whiskers indicate the total range.

Next, we wanted to test whether these results were moderated by the degree to which participants believed AI was involved in a human response, or a human in an AI response (which we term 'assumed aid from the other source (AI/human)'). We used a linear regression model to predict empathy, using condition, assumed aid from the other source and their interaction. Assumed aid did not have an effect on perceived empathy in the AI condition ($P = 0.88$) but showed a significant negative effect in the human condition ($t_{571} = -6.13$; s.e. = 0.05; $P < 0.001$; $\beta = -0.33$; 95% CI, (−0.39, −0.20)). These results suggest that in the human condition, the more participants thought an AI was involved, the less empathic they perceived the response to be (Fig. 3a; see also ref. 51). We found similar effects between assumed aid from the other source and condition on most other measures (Supplementary Information section 4b).

Additionally, to enhance both generalizability and replicability of the results, in study 1c we replicated all of our findings with an open-source LLM (Llama 3.1-405B) (see Supplementary Information section 5 for the results).

Lastly, in study 1d, we aimed to examine whether the gap in the perception of empathy between conditions goes beyond a single-response interaction and is apparent in more realistic and detailed back-and forth interactions. To do this, we again replicated the general structure of study 1b but this time allowed for a four-turn continuous empathic interaction. After describing their experience, the participants waited five seconds and were told they were being connected with another participant or with an AI. They were then instructed to enter the chatroom, in which they would receive a response to their experience and be able to discuss it further by conversing with the other participant/AI for three more messages, receiving a response for each message. Response times and lengths were randomly generated to create variance, giving an authentic impression of an online interaction (see Methods for details).

https://doi.org/10.1038/s41562-025-02247-w

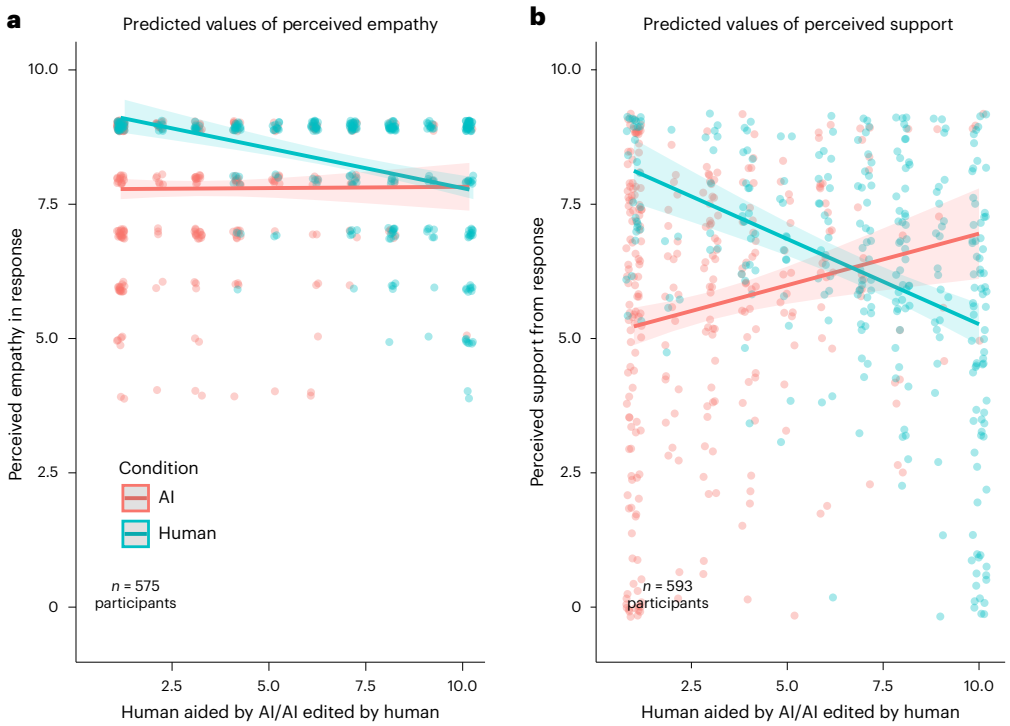

Fig. 3 | Effects of assumed aid from the other source on empathy in the response and perceived support. a,**b**, Contrasts from linear models show significant effects of assumed aid from the other source in each condition on empathy in the response (**a**) and perceived support (**b**). These results show that perceived AI involvement in a response presented as human lowers perceived empathy ($t_{571} = -6.13$; s.e. = 0.05; $P < 0.001$; $\beta = -0.33$; 95% CI, (−0.39, −0.20)) and support ($t_{589} = -6.90$; s.e. = 0.06; $P < 0.001$; $\beta = -0.43$; 95% CI, (−0.55, −0.31), while perceived human involvement in an AI response increases perceived support ($t_{589} = 3.29$; s.e. = 0.08; $P = 0.001$; $\beta = 0.26$; 95% CI, (0.10, 0.41)). Similar effects of assumed aid can be seen in most of the variables—see Supplementary Information section 4b for the results and figures. The regression lines are based on the mean of the dependent variable for each level of assumed aid, and the ribbons represent the 95% CIs around the regression lines.

We first looked at the difference between conditions. A *t*-test did not find a significant difference between conditions in the general empathy rating ($P = 0.25$; Fig. 4a). However, a linear mixed-model analysis using the multiple-item empathy scale, similar to studies 1a and 1b, revealed a significant difference between conditions ($F_{1,687.42} = 5.38$, $P = 0.02$, partial $\eta^2 = 0.008$). Importantly, when focusing on the specific aspects of empathy, contrasts revealed that the effect of condition held only for affective empathy ($t_{975} = -2.59$; s.e. = 0.07; $P = 0.01$; $\beta = -0.18$; 95% CI, (−0.32, −0.04)) and motivational empathy ($t_{977} = -2.41$; s.e. = 0.07; $P = 0.02$; $\beta = -0.17$; 95% CI, (−0.31, −0.03)), but not for cognitive empathy ($P = 0.17$). These results suggest that the affective and motivational aspects of empathy have unique additional value when perceived as human (Fig. 4b).

We then used a linear model to predict the general empathy rating from condition, assumed aid from the other source and their interaction. We replicated the result from studies 1b and 1c, finding a significant interaction between condition and assumed aid, with contrasts showing that in the human condition, empathy was significantly lowered the more participants perceived AI involvement ($t_{673} = -5.87$; s.e. = 0.05; $P < 0.001$; $\beta = -0.32$; 95% CI, (−0.43, −0.21)), while in the AI condition, perceived human involvement had a positive effect on ratings of empathy ($t_{673} = 2.23$; s.e. = 0.05; $P = 0.03$; $\beta = 0.12$; 95% CI, (0.01, 0.22); Fig. 4c). Running the same analysis with the multiple-item empathy scale revealed similar results. Assumed aid from the other source had significant negative effects on all aspects of empathy in the human condition (cognitive empathy: $t_{1005} = -7.00$; s.e. = 0.07; $P < 0.001$; $\beta = -0.46$; 95% CI, (−0.59, −0.33); affective empathy: $t_{994} = -7.82$; s.e. = 0.07; $P < 0.001$; $\beta = -0.51$; 95% CI, (−0.64, −0.38); motivational empathy: $t_{1002} = -6.97$; s.e. = 0.07; $P < 0.001$; $\beta = -0.46$; 95% CI, (−59, −0.33)), but it had a positive effect in the AI condition specifically for affective empathy ($t_{999} = 2.90$; s.e. = 0.06; $P = 0.004$; $\beta = 0.18$; 95% CI, (0.06, 0.31); Fig. 4d). Thus, perceived AI involvement decreased the levels of empathy in a response that was presented as human, and perceived human involvement in a response presented as artificial increased specifically the emotional aspect of empathy.

We found similar effects of condition and interactions on positivity resonance, positive emotions, authenticity and support, which are reported in full in Supplementary Information section 6a.

Studies 2a and 2b addressing concerns of a halo effect

One concern in interpreting the results from the previous studies is that the effect may not be caused solely by differences in perceived empathy. Instead, it could be influenced by a halo effect[56], caused by the speed and eloquence of the response. Receiving such a detailed response so quickly may have led participants to be particularly impressed that a human managed to write it, or to be disappointed when it was an AI response, leading to emphasized differences in ratings that are not indicative of differences in perceived empathy. To address this possibility, we conducted two follow-up studies. Study 2a replicated the results when the participants were given responses limited to two short sentences, written in informal day-to-day language (see Methods for the exact prompt). Study 2b replicated the results when the participants had to wait three minutes for the response. The full results for these two studies are reported in the Supplementary Information (sections 7 and 8).

Together, the results from these six studies show that AI-generated responses, or even AI involvement in a response perceived as human, reduce the perceived emotional value, in terms of experienced empathy as well as general positive emotions and feelings of support. This held true using a different LLM, in longer continuous interactions and

https://doi.org/10.1038/s41562-025-02247-w

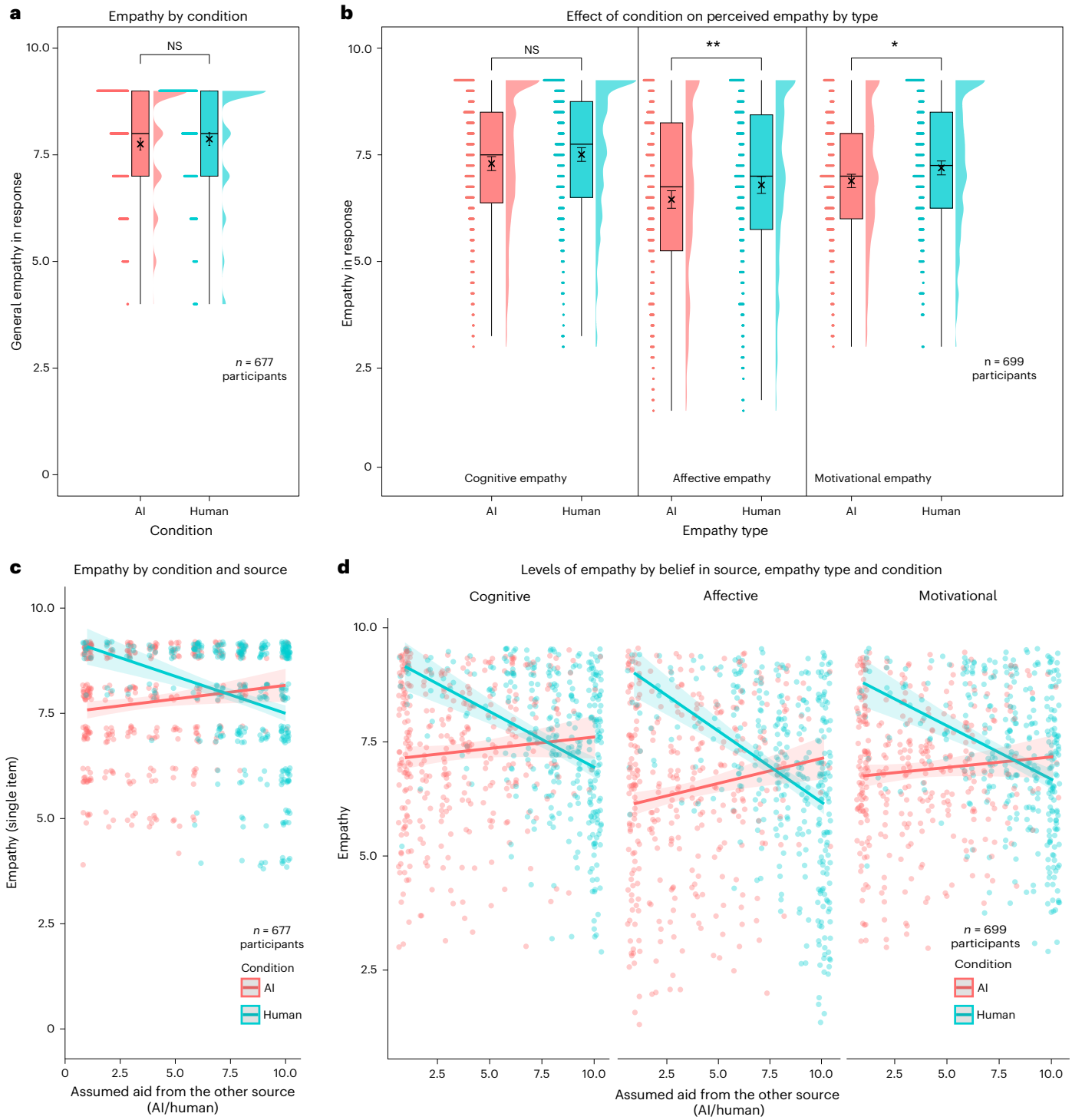

Fig. 4 | Differences in perceived empathy in a multi-turn interaction. **a**, A two-tailed *t*-test revealed no effect of condition on the general empathy rating ($P = 0.25$; NS, not significant). **b**, However, a linear model accounting for aspects of empathy found a significant effect of condition ($F_{1,687.42} = 5.38$, $P = 0.02$, partial $\eta^2 = 0.008$). Follow-up contrasts indicated that this effect was driven by differences in affective empathy ($t_{975} = -2.59$; s.e. = 0.07; $P = 0.01$; $\beta = -0.18$; 95% CI, (−0.32, −0.04)) and motivational empathy ($t_{977} = -2.41$; s.e. = 0.07; $P = 0.02$; $\beta = -0.17$; 95% CI, (−0.31, −0.03)). Within each box, the X marks the mean, with error bars indicating the 95% CI. The central line marks the median, the boxes show the IQR and the whiskers denote the full range. **c**,**d**, A linear model accounting for assumed aid revealed significant interactions for all measures of empathy. Follow-up contrasts indicated that perceived AI involvement in the human condition lowered perceived empathy (**c**—general empathy: $t_{673} = -5.87$; s.e. = 0.05; $P < 0.001$; $\beta = -0.32$; 95% CI, (−0.43, −0.21); **d**—specific aspects: cognitive: $t_{1005} = -7.00$; s.e. = 0.07; $P < 0.001$; $\beta = -0.46$; 95% CI, (−0.59, −0.33); affective: $t_{994} = -7.82$; s.e. = 0.07; $P < 0.001$; $\beta = -0.51$; 95% CI, (−0.64, −0.38); motivational: $t_{1002} = -6.97$; s.e. = 0.07; $P < 0.001$; $\beta = -0.46$; 95% CI, (−59, −0.33)). Conversely, perceived human involvement in the AI condition increased affective empathy (**c**—general empathy: $t_{673} = 2.23$; s.e. = 0.05; $P = 0.03$; $\beta = 0.12$; 95% CI, (0.01, 0.22); **d**—affective empathy: $t_{999} = 2.90$; s.e. = 0.06; $P = 0.004$; $\beta = 0.18$; 95% CI, (0.06, 0.31)). The regression lines are based on the mean of the dependent variable for each level of assumed aid, and the ribbons represent the 95% CIs around the regression lines.

even when participants received very brief responses or had to wait longer to receive a response.

Study 3 stressing different aspects of empathy

The goal of study 3 was to examine whether the gap in the perception of empathy between conditions is dependent on specific aspects of empathy. To do this, we replicated the general structure of study 1b but prompted the AI models to respond in a way that would mostly convey a specific aspect of empathy—understanding (cognitive), feeling with (affective) or caring for (motivational) the participant—by providing the model with a description of what each aspect of empathy would contain (hereafter 'cognitive response', 'affective response' and 'motivational response'; Methods).

First, we wanted to verify whether the different responses from different prompts actually correspond to the different aspects of perceived empathy. We found that every response type was rated significantly higher for its corresponding type of empathy than for other types of empathy (Fig. 5b; see Supplementary Information section 8a for the full comparisons). For example, if the prompt to generate the response was cognitive and aimed to express 'recognizing' emotions, cognitive empathy was rated higher than affective and motivational empathy. These results provide evidence for the validity of the prompts for generating each response type and for the validity of each construct in the questionnaire.

We next analysed the main research question: is the difference in perceived empathy between conditions dependent on the specific response type? First, we found that responses produced by models prompted to give cognitive responses were rated as less empathic than affective responses ($t_{1156} = -4.28$; s.e. = 0.06; $P < 0.001$; $\beta = -0.25$; 95% CI, (−0.39, −0.11)) or motivational responses ($t_{1158} = -4.86$; s.e. = 0.06; $P < 0.001$; $\beta = -0.28$; 95% CI, (−0.42, −0.14)). We replicated the general findings from study 1a and study 1b, where participants in the human condition rated the response as more empathic than did those in the AI condition, for motivational responses ($t_{1117} = 4.02$; s.e. = 0.07; $P < 0.001$; $\beta = -0.30$; 95% CI, (−0.45, −0.15)) and for affective responses ($t_{1117} = -2.00$; s.e. = 0.07; $P = 0.05$; $\beta = -0.15$; 95% CI, (−0.29, −0.001)). But this condition manipulation was not significant for cognitive responses ($P = 0.25$; Fig. 5a). In other words, the models prompted to give a motivational or affective response produced responses that were perceived to be more empathic when presented as human responses, but this gap was not true for models prompted to give a cognitive response.

We found similar results on positivity resonance. We also found that the effect of condition on empathy, positivity resonance, positive emotions and other variables had significant interactions with assumed aid from the other source, similar to study 1b. In other words, the more influence participants thought AI had on a response perceived to be human, the lower they rated empathy, positivity resonance, positive emotions and support. See Supplementary Information section 8d for a full report.

To summarize, the results of studies 1–3 reveal that AI-generated empathic responses are rated quite highly; however, participants who believed they were communicating with a human (as opposed to an AI) attributed higher value overall to the response they received. This effect was more apparent in affective and motivational empathy. Participants also felt more positively and less negatively as a result of the response and believed it to be more helpful and supportive. These effects are driven by the affective (feeling) and motivational (caring) components of empathy in the response.

Studies 4 and 5 participants' choice of AI or human interactions

These next studies aimed to quantify how much more people value human empathy than artificial empathy in terms of time and reasons to wait for a human response over an AI response. We asked the participants to describe a recent emotional experience. We then offered them a choice: in study 4, we offered the participants either an immediate AI response or the option to wait a certain time (2 hours, 24 hours, 1 week, 2 months or 2 years) to receive a human response. An actual empathic human response (written by M.R.) was indeed sent to them afterwards. Study 5 was identical, but instead of receiving a human response after the proposed time, the participants were only offered confirmation that a human had read their story. The goal here was to try to differentiate between the desire to be heard (or in this case, read) and the longing for a genuine human response.

In both studies, we asked the participants how they felt about AI in general. We also asked them to rate the degree to which certain considerations influenced their choices: belief that the AI/human would better understand them, care for them, share their experience or alleviate loneliness; curiosity about the AI/human perspective; hesitation about the alternative; and preference for an immediate response.

We used logistic regression to test whether and how much the time frame offered would increase the probability of participants choosing AI, and we found no effect of time frame ($P = 0.46$ for study 4, $P = 0.70$ for study 5; Supplementary Information section 10a). However, we found significant differences in the considerations that participants expressed as related to their choice. In both studies, participants who chose a human response thought another person would better understand them, share their experience, care about them and alleviate their loneliness. They were also significantly more hesitant about an AI response compared with the hesitation that those who chose AI responses expressed about human responses. Those who chose an AI response rated their preference for an immediate response much higher than did those who chose a human response (see Table 1 for the results).

To summarize, studies 4 and 5 reveal that people are willing to wait a substantial amount of time to receive a human response, with a smaller percentage willing to wait to simply have their experience read by another human. The main motivations for choosing to wait for a human response are related to empathic needs—seeking understanding, sharing feelings and being cared for.

Discussion

The current results consistently show that there are meaningful differences in our perception of empathy between identically generated responses that are thought to be human as opposed to AI. On a basic level, AI responses are deemed empathic overall, as can be seen by participants' high ratings for empathy, positive emotions and support provided by the responses in the AI condition.

However, participants who thought the responses they received were written by a human valued them more, felt more empathy and more support, and wanted to keep conversing with their correspondent more than those who thought the responses were generated by AI. Importantly, this additional value of perceived human communication over communication with AI has to do specifically with affective and motivational components of empathy—the fact that people are able to feel and care, while AI does not. In more realistic four-turn continuous interactions (study 1d), the differences in the affective and motivational aspects of empathy were the ones that mattered most, with participants attributing greater emotional value to them when perceiving them as human. Moreover, when the AI was prompted to answer only in a cognitive manner (study 3), participants valued the perceived human and AI responses similarly, strengthening the notion that people differentiate, at least to some degree, between knowledge of their emotional states (which the AI can give them) and feeling with or caring about them.

The overall positive perception of AI as empathic, as expressed in the high ratings participants gave, is in line with recent papers suggesting that AI could replace humans as emotional supporters to some degree, simulating care and understanding without the burden of fatigue, human avoidance and inherent human biases[5,36,57]. Importantly, this study also shows AI's limits. While AI may be able to replace most forms of human communication, in those cases where one seeks to

https://doi.org/10.1038/s41562-025-02247-w

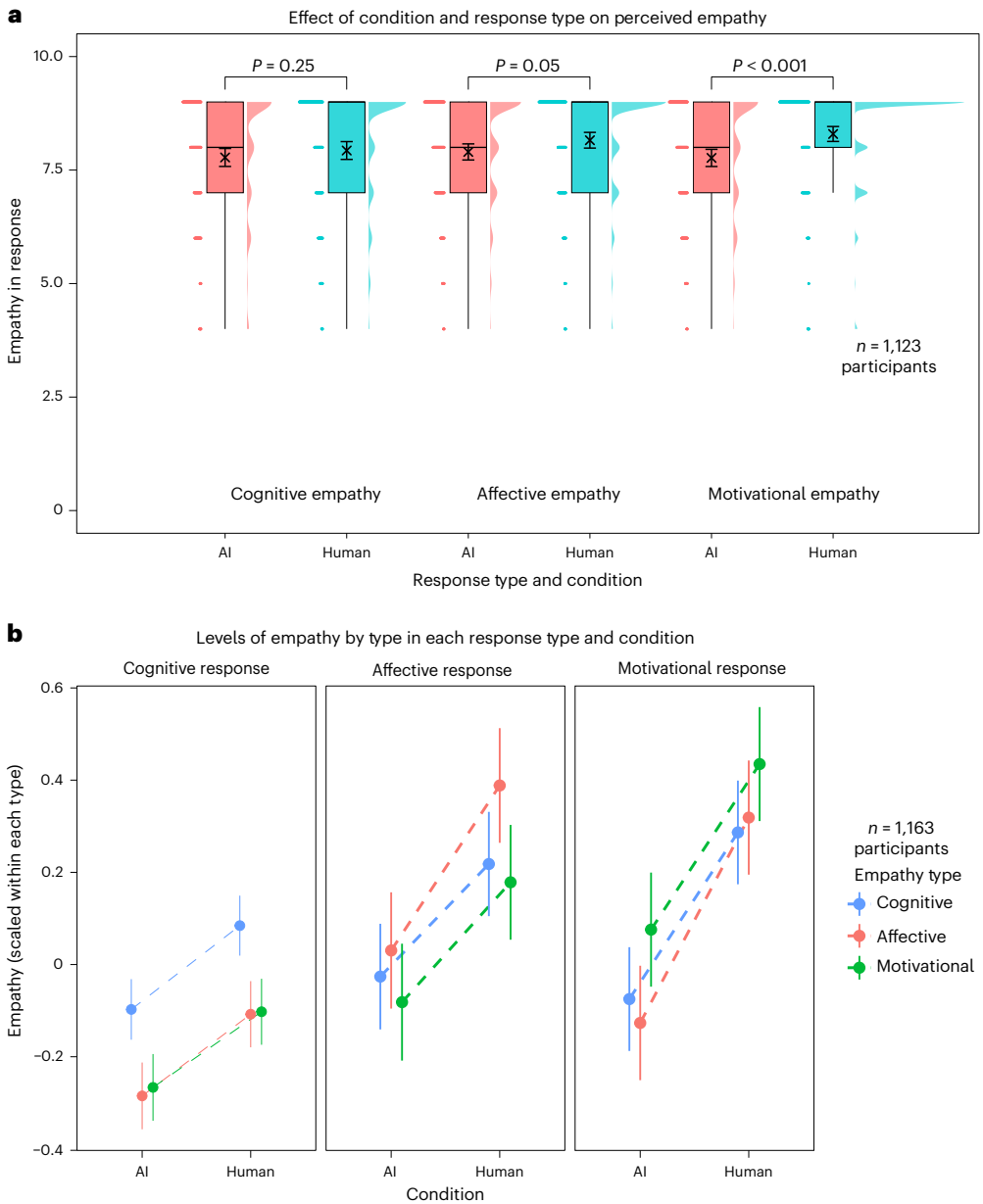

Fig. 5 | Linear model reveals condition-dependent differences in perceived empathy by response type. a, Contrasts show that the effect of condition on overall empathy remained significant only in the responses that conveyed affective ($t_{1117} = -2.00$; s.e. = 0.07; $P = 0.05$; $\beta = -0.15$; 95% CI, (−0.29, −0.001)) or motivational empathy ($t_{1117} = 4.02$; s.e. = 0.07; $P < 0.001$; $\beta = -0.30$; 95% CI, (−0.45, −0.15)). Within each box, the X marks the mean, the error bars represent the 95% CI, the central line indicates the median, the box shows the IQR and the whiskers represent the full data range. **b**, Within each response type, the corresponding type of empathy was rated highest. Contrasts showed significant differences between conditions for affective and motivational responses (see Supplementary Information section 9a for details). The circle marks the mean for each group, and the bars represent the 95% CIs. Individual data points are omitted for readability; the full plot with data points is available in Supplementary Fig. 8b.

share emotions and receive genuine emotional sharing and care, human empathy will be valued more than that of AI[53].

Studies 4 and 5 further reveal that people choose to communicate with people or AI for different reasons. Those who chose a response from an AI did so mainly due to the immediacy of the response or due to curiosity regarding AI responses. Interestingly, a notable percentage of the participants reported that they chose to receive a response from AI because they were hesitant about a human response, which supports the possibility that in some cases people will in fact prefer talking to AI, possibly to avoid disclosing personal information to another person, or to not feel judged[58]. In contrast, people choose human communication when they want to receive genuine empathy and emotional support—understanding, feeling and caring—as well as to alleviate loneliness, even if they must wait a considerable amount of time for it.

For some people, the need to be seen (or heard/read) is enough to make them willing to wait longer for a human reader. Just the thought that another human being knows what you are going through allows for a greater sensation of feeling understood and cared for than an immediate AI response does. Waiting a considerable amount of time also allows for the idea that someone is putting time aside (that is, mental and emotional effort) to answer you or think about your experience. The investment of time and energy into a relationship is thought to be critical for its maintenance and improvement[59,60]. Thus, thinking that another human is potentially taking the time to do so gives room for

https://doi.org/10.1038/s41562-025-02247-w

Table 1 | Differences in reasons for choices in studies 4 and 5

Study	Reason given	*t*	Mean difference	CI for difference	d.f.	*P*	Cohen's *d*
Study 4	Understanding	−37.4	−5.31	−5.59, −5.03	450.10	<0.001***	3.24
	Sharing	−40.55	−5.52	−5.79, −5.25	471.02	<0.001***	3.54
	Caring	−33.43	−5.01	−5.30, −4.71	468.57	<0.001***	2.92
	Loneliness	−22.72	−4.10	−4.45, −3.74	480.06	<0.001***	2.05
	Timing	14.58	2.82	2.44, 3.20	409.04	<0.001***	−1.36
	Hesitation	−4.53	−0.97	−1.38, −0.55	456.42	<0.001***	0.41
	Curiosity	1.31	0.19	−0.10, 0.48	434.55	0.19	0.12
Study 5	Understanding	−37.46	5.45	−5.73, −5.16	442.51	<0.001***	3.18
	Sharing	−37.65	−5.56	−5.85, −5.27	430.97	<0.001***	3.24
	Caring	−24.90	−4.63	−4.99, −4.26	330.74	<0.001***	2.37
	Loneliness	−16.45	−3.56	−3.98, −3.13	278.58	<0.001***	1.67
	Timing	21.92	3.96	3.61, 4.32	352.36	<0.001***	−2.04
	Hesitation	−4.57	−1.11	−1.58, −0.63	320.37	<0.001***	0.44
	Curiosity	1.93	0.34	−0.01, 0.70	293.82	0.05†	0.19

Detailed results for the differences in reasons given for choice of either the AI or the human response in studies 4 and 5. Two-tailed *t*-tests were used to compare the rating of each reason between groups. ***$P < 0.001$; †$P < 0.10$.

imagining a closer bond. However, this is an experience that AI cannot (genuinely) deliver.

Importantly, even the notion that a perceived human response was aided by AI reduces its positive effects[8,41,44]. This resonates with the idea that genuine empathy and care require human experience and effort. When it becomes apparent, whether explicitly stated, subtly hinted or even just believed, that a response has not solely originated from an individual person, its perceived value diminishes. While we may not always accurately detect when a response is generated by AI[44,61], it is becoming increasingly evident that a substantial portion of our daily interactions will soon be augmented or reviewed if not entirely generated by AI. We can therefore anticipate a decrease in emotional resonance or felt empathy in these experiences. While this can be seen as a question of framing, this framing is crucial, and it has important theoretical and applicative implications. Similar to the experience of receiving a generic birthday or condolence card, which often lacks a personal touch and emotional meaning, any interaction, even when personalized, may soon carry a sense of detachment and lack of authenticity, should it be perceived as involving AI.

The understanding that human empathy possesses a unique emotional value that AI cannot truly provide is a serious consideration when planning how to implement AI ethically in different fields. Certainly, there are moments when understanding, a courteous and uplifting response or sound advice go a long way. Moreover, AI responses were rated as highly empathic and can definitely provide these expressions of polite understanding and advice in an effective, cost-effective, scalable manner. In these instances, AI is not constrained by human limitations such as schedules, fatigue, burnout or distractions; and as the technology advances and improves, it may soon (if it has not already) reach a human-level understanding of others' emotions in context (cognitive empathy). Indeed, with careful and proper application, AI could be used wisely to improve medical and mental health outcomes[7,62–64]. Similar considerations arise when applying AI to education, benefiting from its advantages, including personalized teaching, at scale and any time. However, all these implementations need to be done while using valuable human communication when it is most needed—that is, in specific places within the interaction that require emotional connection, feeling and care. For these situations we suggest that humans (teachers, caregivers, partners, friends or at times even a listening stranger) will have greater value. The challenge now becomes realizing when this is the case[49,57].

Limitations and future directions

The current study was limited to relatively short, text-based interactions. Consequently, the potential effects observed may differ in the context of lengthier emotional conversations. It is unclear whether individuals would develop increased trust and positive reactions over long-term human–AI empathic communication, or experience heightened negativity and unease when engaging in more intimate ongoing communication with AI. Moreover, as advancements in AI applications progress, particularly with the introduction of AI avatars or human-like robots, distinguishing between interactions with humans and those with AI will become increasingly challenging. Even if potential regulatory measures mandate disclosure of conversational partners[65,66], individuals may easily forget this distinction amid immersive conversations. Future research should investigate the impacts of extended dialogues and the integration of more lifelike features and modes of communication on individuals' perceptions and valuation of emotional interactions with AI.

Another limitation is that we did not compare actual human responses to the AI responses. This was done to keep responses identical on every level, except the perception of who the response is from, and to allow us to control the specific aspects and empathic content expressed in the empathic responses in real time. Whereas some studies have compared AI and human responses[7,41,67,68], they did not do so in real-time interactions, instead splitting the interaction into two sections and gathering (and sometimes coding) human responses between them. These studies also did not differentiate between the various aspects of empathy. While the use of deception in behavioural studies should be limited, we believe that in this case it was vital to answer our research question. Specifically, the current design enabled us to differentiate the value of a perceived human versus AI-generated response, keeping all else equal, and doing so in ecological real-time personalized interactions, as similar as possible to real interactions. We leveraged AI to keep the content, tone and timing of the responses constant and controlled, while manipulating only the perceived source and context. In studies 4 and 5, where the responses were not in real time, a meaningful limitation is the believability of such an offer of a delayed response. Future research could further assess how actual responses from AI versus humans are perceived and preferred in real-time settings as a function of the aspects and levels of empathy expressed in them. It would also be useful to explore in which contexts the various aspects of empathy are valued more.

The study was also limited to online samples. While using these paradigms in online settings is highly ecological, the samples analysed are not necessarily representative. Future studies should replicate these findings in different contexts with different populations.

Finally, it should be acknowledged that what we value today may not be what we value in the future. With the rise of AI–human or AI-mediated intimate interactions, our perception of empathy—how we share emotions and care—could shift[69]. We might start valuing different aspects of the interaction more or define empathy differently altogether. This evolving landscape underscores the need for ongoing research to keep pace with our changing values and culture in the realm of AI-mediated relationships.

Conclusions

Across nine studies (n = 6,282) involving brief emotional interactions and manipulating the perceived source of the responder, we have demonstrated that people attribute greater emotional value to communication they perceive as human than to interactions with an AI chatbot or those perceived as assisted by AI. These differences stem from humans' unique potential to feel and care, which AI currently does not possess. These findings have important implications for the integration of AI into any domain that requires emotional sensitivity, from health care, education, and counselling to marketing and sales. While AI can play a valuable role in augmenting human efforts—such as offering cognitive empathy or providing practical support—it is less effective in contexts requiring authentic emotional sharing and deep care. In the spaces that demand a deep sense of emotional sharing and caring, human connection is still valued more.

Methods

Ethical statement

All studies were approved by the ethics committee of the social sciences faculty of the Hebrew University of Jerusalem (IRB approval number 17112023 for studies 1–3 and approval number IRB_2023_048 for studies 4 and 5).

Preregistrations

All studies were preregistered. Study 1a was preregistered on 19 November 2023 (https://aspredicted.org/DCV_J2L). Study 1b was preregistered on 25 December 2023 (https://aspredicted.org/75G_XKZ). Study 1c was preregistered on 31 October 2024 (https://aspredicted.org/gxzw-zyqr.pdf). Study 1d was preregistered on 14 November 2024 (https://aspredicted.org/25km-8sr3.pdf). Study 2a was preregistered on 23 October 2024 (https://aspredicted.org/kqkx-gjwj.pdf). Study 2b was preregistered on 29 October 2024 (https://aspredicted.org/nw6d-48d8.pdf). Study 3 was preregistered on 7 December 2023 (https://aspredicted.org/QRX_QHM). Study 4 was preregistered on 4 January 2024 (https://aspredicted.org/4N4_DDW). Study 5 was preregistered on 17 January 2024 (https://aspredicted.org/7XD_YJF).

Deviations from preregistrations

We did not specify the full details of our power analyses in the preregistrations. Power analyses for each study are reported fully in detail below and were run prior to writing the preregistrations. We also did not specify that we would exclude participants who failed attention checks concerning the instructions they received, as reported below. The main hypothesis for study 1a was preregistered as a directional hypothesis, stating that people would attribute more empathy to responses perceived as human. For statistical conservatism, the tests were analysed with no directionality, but the effect was still significant and in the hypothesized direction.

We initially planned to analyse study 1b using a series of t-tests, treating each aspect of empathy as a separate dependent variable in a separate analysis. To comply with the preregistration, we also conducted these t-tests on the specific aspects of empathy. However, these are identical in result and less powerful than the mixed-model analysis described earlier. These analyses can be found in Supplementary Information section 4a.

Participants

All participants were recruited via Prolific. The participants in studies 1a, 1b, 1d, 3, 4 and 5 were recruited from the UK, and the participants in studies 1c, 2a and 2b were recruited from the USA. While we did not collect information on ethnicity, here we report information on participants' ethnicity as it appears on the Prolific platform. All participants were compensated for their time at a rate of 9 GBP an hour.

For study 1a, following power analyses looking for an effect size of Cohen's d = 0.2 at 80% power, we aimed to reach a final sample of 800 participants. Starting with 998 participants, we filtered out those who failed attention checks (n = 263) or reported technical issues (n = 10), leading to a final sample of 725 participants (mean age, 28.42 ± 4.51 (s.d.); 53% female; AI condition n = 371; mean years of education, 15.78 ± 3.10 (s.d.); 77% White, 9% Asian, 9% Black, 4% mixed, 1% other).

For study 1b, following power analyses to find the same effect size as in study 1a (Cohen's d = 0.36) at 80% power, we aimed to reach a final sample of 600 participants. Starting with 827 participants, we filtered out those who failed attention checks (n = 190) or reported technical issues (n = 44), leading to a final sample of 593 participants (mean age, 28.06 ± 4.57 (s.d.); 53% female; AI condition n = 296; mean years of education, 16.15 ± 3.08 (s.d.); 74% White, 9% Asian, 8% Black, 6% mixed, 2% other, 1% unreported).

For study 1c, following power analyses looking for an effect size of Cohen's d = 0.2 at 80% power, we aimed to reach a final sample of 800 participants. Starting with 999 participants, we filtered out those who failed attention checks (n = 254) or reported technical issues (n = 11), leading to a final sample of 734 participants (mean age, 28.37 ± 4.68 (s.d.); 52% female; AI condition n = 374; mean years of education, 15.51 ± 2.93 (s.d.); 53% White, 10% Asian, 20% Black, 12% mixed, 5% other).

For study 1d, following power analyses looking for an effect size of Cohen's d = 0.2 at 80% power, we aimed to reach a final sample of 800 participants. Starting with 994 participants who finished the study, we filtered out those who failed attention checks (n = 278), explicitly used a bot to write the story (n = 1), participated twice (n = 2) or reported technical errors (n = 8), leading to a final sample of 705 participants (mean age, 27.63 ± 4.60 (s.d.); 51% female; AI condition n = 357; mean years of education, 15.70 ± 3.33 (s.d.); 56% White, 12% Asian, 25% Black, 5% mixed, 1% other, 1% unreported).

For study 2a, following power analyses looking for an effect size of Cohen's d = 0.2 at 80% power, we aimed to reach a final sample of 800 participants. Starting with 1,003 participants, we filtered out those who failed attention checks (n = 294), reported technical issues (n = 21) or participated twice (n = 1), leading to a final sample of 687 participants (mean age, 28.10 ± 4.64 (s.d.); 48% female; AI condition n = 348; mean years of education, 15.50 ± 2.93 (s.d.); 57% White, 12% Asian, 16% Black, 9% mixed, 4% other, 2% unreported).

For study 2b, following power analyses looking for an effect size of Cohen's d = 0.2 at 80% power, we aimed to reach a final sample of 800 participants. Starting with 979 participants, we filtered out those who failed attention checks (n = 266) or reported technical issues (n = 22), leading to a final sample of 691 participants (mean age, 28.26 ± 4.68 (s.d.); 49% female; AI condition n = 348; mean years of education, 15.43 ± 2.84 (s.d.); 55% White, 6% Asian, 23% Black, 10% mixed, 5% other, 1% unreported).

For study 3, given that we were looking to find complex interactions, we ran a simulation for power analyses. We used 10,000 permutations of a dataset, with an interaction between condition, empathy type and response type with an average effect size of partial η^2 = 0.01. We determined that to find that effect at 80% power we needed at least 200 participants per group, and we aimed for a sample of 1,200

participants. Starting with 1,550 participants who finished the study, we filtered out those who failed attention checks (n = 354), reported technical errors (n = 21) or participated twice (n = 3), leading to a final sample of 1,172 participants (mean age, 28.16 ± 4.48 (s.d.); 54% female; AI condition n = 585; mean years of education, 15.97 ± 3.15 (s.d.); 79% White, 9% Asian, 7% Black, 4% mixed, 1% other, 1% unreported).

For study 4, given that we had no similar comparison to rely on for an effect size, we followed the guidelines proposed by Brysbaert[70] and aimed to recruit 500 participants, 100 in each condition. Among the initial 489 participants, none reported meaningful technical issues during the experiment, so that was the final sample size (mean age, 28.05 ± 4.62 (s.d.); 50% female; n('1 week') = 97, n = 98 in all other groups; mean years of education, 15.96 ± 3.38 (s.d.); 76% White, 8% Asian, 8% Black, 5% mixed, 2% other, 1% unreported).

For study 5, given that we had no similar comparison to rely on for an effect size, we followed the guidelines proposed by Brysbaert[70] and aimed to recruit 500 participants, 100 in each condition. Among the initial 486 participants, none reported meaningful technical issues during the experiment, so that was the final sample size (mean age, 27.91 ± 4.68 (s.d.); 47% female; n('1 week') = 95, n('24 hours') = 97, n = 98 in all other groups; mean years of education, 15.97 ± 3.29 (s.d.); 74% White, 10% Asian, 8% Black, 6% mixed, 1% other).

Studies 1–3

Goal and procedure. The goal of these studies was to test our main hypotheses. First, do participants rate responses they perceive to come from a human source as more empathic than responses from an artificial source? Second, is this difference related to a specific aspect of empathy?

To do this, after obtaining informed consent and informing the participants that they could withdraw at any time, we asked the participants to share an emotional experience they had in the past month. We randomly assigned participants to one of two conditions: the human condition, where they were told they were paired with another participant, and the AI condition, where they were told they were paired with an AI. It should be noted that in study 1a, the participants were first assigned to the conditions and then asked to write down their experience, and in studies 1b, 1c, 1d, 2a, 2b and 3, they were first asked to write down their experience and then assigned to the conditions. In all studies, there was no experimenter involvement as they were run online.

In studies 1a, 1b, 1c, 2a and 3, all participants next waited for 60 seconds and were told that their partner was writing a response in the human condition or that the AI was generating a response in the AI condition. In study 2b the procedure was identical, but the participants waited three minutes for the response. Their stories were sent via Qualtrics web services to the relevant API (OpenAI: GPT 4-0613 for studies 1a, 1b, 1d and 3; model GPT o1-preview-2024-09-12 for studies 2a and 2b; and Fireworks API for model Llama 3.1 405-B for study 1c), with instructions defining an empathic response and requesting an empathic response to the story (see Supplementary Information section 13 for the exact prompts).

The participants then received an AI-generated response that was instructed to be empathic and was specific to the content of their story. In study 3, the participants were randomly assigned to one of three different response types: the AI was instructed to respond in a manner that was mostly cognitively, affectively or motivationally empathic (see Supplementary Information section 13 for the exact prompts).

In study 1d, the response was the first message they received in a four-turn instant-messaging interaction. After waiting five seconds, they were told that they had been matched with another participant or with an AI (their assigned experimental condition). They could then send three more messages while conversing back and forth with the AI, which was portrayed as a human in one condition (with participants seeing an indicator saying 'participant is typing…' between messages) and as an AI in the other (with participants seeing an indicator saying 'AI is generating…'). To maintain credibility, the maximum number of sentences and words in each sentence were randomly generated within a predefined range, to create variance between messages. Sentence limits varied between 2 and 4, with an average of 2.99 ± 0.82 (s.d.). Word limits varied between 7 and 11, with an average of 9.02 ± 1.40 (s.d.) words per sentence. Additionally, after each message was sent, the 'participant typing/AI generating' indicator appeared after 2.5–4.5 seconds, and the reply appeared after a delay of 8–12 seconds (6 seconds with an additional second for each sentence, leading to an average of 8.99 ± 0.82 seconds). During the interaction, the participants were shown a counter with the remaining number of messages they could send.

The AI was given a system prompt detailing how to respond empathically to the conversation, as well as the number of messages left in the conversation, and it was instructed to close the conversation on the last one (see below).

After reading the response in studies 1–3 or finishing the conversation in study 1d, the participants rated the interaction for empathy and positivity resonance. In studies 1b, 1c, 1d, 2a, 2b and 3, we added questions on authenticity, support, positive and negative emotions, and assumed aid from the other source (detailed below).

Although these experimental designs involve deception, we believe investigating the gaps in perceived empathy and support between human and artificial communication is critical and justifies this short deception, given the increasing use of AI in psychology, mental health, education and health care, in both research and practice. Importantly, at the end of the experiment, all participants were debriefed that the responses were generated by AI. They were informed that this was done to measure the differences between emotional reactions when believing the response was written by a human as opposed to generated by AI. The participants were also provided with support resources in their own country, available to them in case they were distressed by the study, and with the researchers' contact information for concerns about the research specifically (no concerns were raised).

Measures in studies 1–3. *Empathy in response inventory*. We constructed a self-report questionnaire to measure the degree to which participants perceived the response as empathic. The questionnaire included 15 items, five for each aspect of empathy: cognitive, affective and motivational. They rated the truth of each item from 0 (not true at all) to 9 (completely true). Each participant saw the questions phrased in congruency with their assigned experimental condition (for example, a participant in the AI condition saw 'I feel the AI correctly understood my emotions', whereas participants in the human condition saw 'I feel the other person correctly understood my emotions'). For a full list of items, see Supplementary Information section 12.

In study 1a, we used all 15 items, presented in a random order. We saw excellent internal consistency for the questionnaire as a whole (Cronbach's α = 0.91) and acceptable-to-good internal consistency for each aspect of empathy (Cronbach's α for cognitive empathy, 0.82; Cronbach's α for affective empathy, 0.75; Cronbach's α for motivational empathy, 0.78). In subsequent versions we removed three items, one in each aspect, that did not correlate well with the other items for that aspect. After cleaning the data for outliers on each item, we ran a confirmatory factor analysis on this 12-item questionnaire to test whether the division into the three aspects of empathy fit the data well, and it showed an acceptable fit over a single-factor model (root mean square error of approximation (RMSEA), 0.11; Akaike information criterion (AIC), 21,844; comparative fit index (CFI), 0.91).

In study 1a, overall empathy scores were the average of the entire 15-item questionnaire, and each aspect was scored by averaging the questions for that aspect. In studies 1b, 1c, 1d, 2a, 2b and 3, we did the same, using the 12-item version, and saw similar levels of consistency (in study 2, overall Cronbach's α = 0.91, cognitive Cronbach's α = 0.84, affective Cronbach's α = 0.78, motivational Cronbach's α = 0.70;

in study 3, overall Cronbach's $\alpha = 0.93$, cognitive Cronbach's $\alpha = 0.85$, affective Cronbach's $\alpha = 0.83$, motivational Cronbach's $\alpha = 0.73$).

Positivity resonance. We used three questions taken from the Positivity Resonance measure developed in ref. 55, which is designed to measure emotional synchrony, warmth and positivity between two individuals. The participants rated from 0 (did not feel at all) to 100 (felt profoundly) the degree to which they felt mutual warmth and concern, a mutual sense of feeling energized and uplifted, and a mutual sense of trust and respect when reading the response. These showed excellent internal consistency throughout our studies (Cronbach's $\alpha = 0.92–0.93$). We averaged these ratings into a final score.

Support questions. The participants rated from 0 (not at all) to 9 (very much) the degree to which they felt the interaction helped them, how much they felt further interactions could help them in the future and how much they felt they would like to keep conversing with their partner. Similar to the questions on empathy, each participant saw a phrasing that was congruent with their assigned condition. We observed high correlations among these questions ($r = 0.77–0.85$, all $P < 0.001$) and excellent internal consistency (Cronbach's $\alpha = 0.93$), and we averaged them into a single score.

Positive and negative emotions. The participants rated from 0 (not at all) to 9 (very much) the degree to which they felt four negative emotions (disturbed, angry, distressed and annoyed) and four positive emotions (comforted, validated, happy and understood) after reading the response. They were also asked to freely state what impressed them about the response and what bothered them about it. The items were presented in a random order. We saw good internal consistency for both positive (Cronbach's $\alpha = 0.90$) and negative (Cronbach's $\alpha = 0.81–0.82$) emotions. We averaged each set of four emotions into a final score.

Additional questions. We asked the participants five additional questions. First, the participants rated from 0 (not at all) to 9 (very much) the degree to which they felt the response they received was authentic. Second, they rated the degree to which they believed AI was involved in a human response, or a human involved in an AI response (termed above 'assumed aid from the other source'). They also rated the way they felt about AI from 0 (very negatively) to 9 (very positively), and separately the frequency with which they use AI personally and professionally from 0 (never) to 9 (all the time).

Analysis. All analyses were done using R[71]. We removed any participants that were more than 2.5 standard deviations away from the mean of the dependent variable for that analysis. We adjusted for multiple comparisons using Bonferroni corrections. All independent categorical variables were effect-coded. To ensure that non-normality of some of the dependent variables did not affect the results, we also conducted non-parametric or permutation tests for each analysis. Additionally, we repeated all analyses without filtering outliers. All main findings were significant across analyses. All additional analyses are reported in the Supplementary Information for each study.

In studies 1a, 1b, 1c, 1d, 2a and 2b, we used *t*-tests to analyse the basic differences between conditions. To examine the more complex interactions between condition and aspect of empathy, we first scaled the empathy scores within each aspect and used linear mixed models with a random intercept for each participant, given that each participant rated all three aspects. In study 3, we used similar analyses, but instead of *t*-tests for the basic difference between conditions, we used multiple regression models to account for the effects of condition, response type and the interaction between them. When using multiple regression or mixed regression models, we analysed them in an analysis of variance to test the significance of each effect and used post hoc contrasts to differentiate the effects.

Studies 4 and 5

Goals and procedure. The goal of studies 4 and 5 was to see how long people are willing to wait to receive a human response (in study 4) or to have their story read by a fellow human (study 5) to differentiate between the desire for human empathy and the desire to feel heard by another person. To do this, after obtaining informed consent, we asked the participants to describe a recent emotional event, as in the previous studies. We then asked them to choose between receiving an immediate AI response, stressing that it would not be reviewed by a human in any way, or a response from a therapist in training, within a time frame that would not include any AI influence. We used a logarithmic scale to decide on relative time frames, and participants were randomly assigned to one of five possible time frames: 2 hours (rounding 10^2 minutes), 24 hours, 1 week, 2 months and 2 years (rounding 10^6 minutes). We then asked them to rate the truth of several statements concerning different possible reasons for their choice. Afterwards, they answered questions on feelings about AI and their use of it, and they responded to a short loneliness scale (detailed below). At the end of the study, the participants were debriefed in the same manner as in studies 1–3 (see above). In both studies, there was no experimenter involvement as they were run online.

Measures in studies 4 and 5. *Statements on reasons for their choice*. The participants were asked to rate from 1 (not true at all) to 9 (very true) the truth of each statement in describing the reason for their choice between a human response and an AI response. All statements were presented in a random order. They rated the degree to which they were hesitant about the response of the alternative option (participants who chose AI were asked if they were hesitant about a human response), their curiosity about the response from their choice and the degree to which they felt their choice could better understand them, share their emotions, care about them and alleviate their loneliness compared with the alternative. They were also asked whether their choice was influenced by a preference for an immediate or a delayed response.

Questions on AI. We asked the participants to rate their feelings about AI and how often they used AI in their professional and personal lives, as stated previously.

Loneliness scale. We asked the participants to answer the three-item loneliness scale[72], a short self-report questionnaire designed to measure levels of loneliness, and it showed excellent internal consistency in our samples (Cronbach's $\alpha = 0.83$ in study 4, 0.81 in study 5).

Analysis. To answer the central question of whether there is a length of time that would predict a specific choice, we used a logistic binomial regression model, with the time frame offered as a predictor. In another exploratory analysis trying to predict the choice of participants, we scaled participants' feeling about AI and added it to the earlier logistic regression model. We did the same with participants' reported loneliness.

To test the difference between the ratings of reasonings in the different groups, we used a series of *t*-tests to compare the ratings of those who chose an AI response to those who chose a human response. We also conducted non-parametric tests replicating these findings (Supplementary Information section 2).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All preprocessed data, excluding participants' personal experiences when they did not provide consent to share them, are available via OSF at https://osf.io/w4hkd/?view_only=52a2324d36bc4c03ad9f1d90ba75ab7b.

Code availability

All analysis files are available via OSF at https://osf.io/w4hkd/?view_only=52a2324d36bc4c03ad9f1d90ba75ab7b.

References

1. Gero, K. I. *AI and the Writer: How Language Models Support Creative Writers* (Columbia Univ., 2023).
2. Joksimovic, S., Ifenthaler, D., Marrone, R., De Laat, M. & Siemens, G. Opportunities of artificial intelligence for supporting complex problem-solving: findings from a scoping review. *Comput. Educ. Artif. Intell.* **4**, 100138 (2023).
3. Wang, L. et al. Document-level machine translation with large language models. In *Proc. Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H. et al.) 16646–16661 (ACL, 2023). https://doi.org/10.18653/v1/2023.emnlp-main.1036
4. Yao, S. et al. Tree of thoughts: deliberate problem solving with large language models. *Adv. Neural Inf. Process. Syst.* **36**, 11809–11822 (2023).
5. Inzlicht, M., Cameron, C. D., D'Cruz, J. & Bloom, P. In praise of empathic AI. *Trends Cogn. Sci.* **28**, 89–91 (2023).
6. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* **5**, 46–57 (2023).
7. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
8. Morris, R. We provided mental health support to about 4,000 people—using GPT-3. Here's what happened. *X* https://twitter.com/RobertRMorris/status/1611450197707464706 (2023).
9. Ong, D. et al. Is discourse role important for emotion recognition in conversation? In *Proc. AAAI Conference on Artificial Intelligence* Vol. 36 11121–11129 (PKP, 2022).
10. Sharma, A. et al. Cognitive reframing of negative thoughts through human-language model interaction. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics* Vol. 1 (eds Rogers, A. et al.) 9977–10000 (ACL, 2023). https://doi.org/10.18653/v1/2023.acl-long.555
11. Lin, I. et al. IMBUE: improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics* Vol. 1 (eds Ku, L.-W. et al.) 810–840 (ACL, 2024). https://doi.org/10.18653/v1/2024.acl-long.47
12. Replika https://replika.com (Luka, 2025).
13. Maples, B., Cerit, M., Vishwanath, A. & Pea, R. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Ment. Health Res.* **3**, 4 (2024).
14. Zaki, J. & Ochsner, K. N. The neuroscience of empathy: progress, pitfalls and promise. *Nat. Neurosci.* **15**, 675–680 (2012).
15. Schurz, M. et al. Toward a hierarchical model of social cognition: a neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychol. Bull.* **147**, 293–327 (2021).
16. Feldman, R. Social behavior as a transdiagnostic marker of resilience. *Annu. Rev. Clin. Psychol.* **17**, 153–180 (2021).
17. Andreychik, M. R. I like that you feel my pain, but I love that you feel my joy: empathy for a partner's negative versus positive emotions independently affect relationship quality. *J. Soc. Pers. Relat.* **36**, 834–854 (2019).
18. Batson, C. D. et al. Empathic joy and the empathy–altruism hypothesis. *J. Pers. Soc. Psychol.* **61**, 413–426 (1991).
19. Pierce, J. R., Kilduff, G. J., Galinsky, A. D. & Sivanathan, N. From glue to gasoline: how competition turns perspective takers unethical. *Psychol. Sci.* **24**, 1986–1994 (2013).
20. Weisz, E. & Zaki, J. in *The Oxford Handbook of Compassion Science* (eds Seppälä, E. M. et al.) 205–218 (Oxford Univ. Press, 2017); https://doi.org/10.1093/oxfordhb/9780190464684.013.16
21. Bartal, I. B.-A., Decety, J. & Mason, P. Empathy and pro-social behavior in rats. *Science* **334**, 1427–1430 (2011).
22. de Waal, F. B. M. & Preston, S. D. Mammalian empathy: behavioural manifestations and neural basis. *Nat. Rev. Neurosci.* **18**, 498–509 (2017).
23. Cameron, C. D. et al. Empathy is hard work: people choose to avoid empathy because of its cognitive costs. *J. Exp. Psychol. Gen.* **148**, 962–976 (2019).
24. Eyal, T., Steffel, M. & Epley, N. Perspective mistaking: accurately understanding the mind of another requires getting perspective, not taking perspective. *J. Pers. Soc. Psychol.* **114**, 547–571 (2018).
25. Choshen-Hillel, S. et al. Physicians prescribe fewer analgesics during night shifts than day shifts. *Proc. Natl Acad. Sci. USA* **119**, e2200047119 (2022).
26. Guadagni, V., Burles, F., Ferrara, M. & Iaria, G. The effects of sleep deprivation on emotional empathy. *J. Sleep Res.* **23**, 657–663 (2014).
27. Seo, H.-Y. et al. Burnout as a mediator in the relationship between work–life balance and empathy in healthcare professionals. *Psychiatry Investig.* **17**, 951–959 (2020).
28. Vévodová, Š., Vévoda, J., Vetešníková, M., Kisvetrová, H. & Chrastina, J. The relationship between burnout syndrome and empathy among nurses in emergency medical services. *Kontakt* **18**, e17–e21 (2016).
29. Wilkinson, H., Whittington, R., Perry, L. & Eames, C. Examining the relationship between burnout and empathy in healthcare professionals: a systematic review. *Burn. Res.* **6**, 18–29 (2017).
30. Ferguson, A. M., Cameron, C. D. & Inzlicht, M. When does empathy feel good? *Curr. Opin. Behav. Sci.* **39**, 125–129 (2021).
31. Tak, A. N. & Gratch, J. GPT-4 emulates average-human emotional cognition from a third-person perspective. Preprint at https://arxiv.org/abs/2408.13718 (2024).
32. Sorin, V. et al. Large language models and empathy: systematic review. *J. Med. Internet Res.* **26**, e52597 (2024).
33. Paiva, A., Leite, I., Boukricha, H. & Wachsmuth, I. Empathy in virtual agents and robots: a survey. *ACM Trans. Interact. Intell. Syst.* **7**, 11:1–11:40 (2017).
34. Wang, Y. et al. A systematic review on affective computing: emotion models, databases, and recent advances. *Inf. Fusion* **83–84**, 19–52 (2022).
35. Gandhi, K. et al. Human-like affective cognition in foundation models. Preprint at https://arxiv.org/abs/2409.11733 (2024).
36. Ovsyannikova, D., de Mello, V. O. & Inzlicht, M. Third-party evaluators perceive AI as more compassionate than expert humans. *Commun. Psychol.* **3**, 4 (2025).
37. Lee, Y. K., Suh, J., Zhan, H., Li, J. J. & Ong, D. C. Large language models produce responses perceived to be empathic. Preprint at https://arxiv.org/abs/2403.18148 (2024).
38. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Towards facilitating empathic conversations in online mental health support: a reinforcement learning approach. In *Proc. Web Conference* 194–205 (ACM, 2021). https://doi.org/10.1145/3442381.3450097
39. Zhan, H. et al. Large language models are capable of offering cognitive reappraisal, if guided. Preprint at https://arxiv.org/abs/2404.01288 (2024).
40. Pataranutaporn, P., Liu, R., Finn, E. & Maes, P. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nat. Mach. Intell.* **5**, 1076–1086 (2023).

41. Yin, Y., Jia, N. & Wakslak, C. J. AI can help people feel heard, but an AI label diminishes this impact. *Proc. Natl Acad. Sci. USA* **121**, e2319112121 (2024).
42. Hohenstein, J. & Jung, M. AI as a moral crumple zone: the effects of AI-mediated communication on attribution and trust. *Comput. Hum. Behav.* **106**, 106190 (2020).
43. Purcell, Z. A., Dong, M., Nussberger, A.-M., Köbis, N. & Jakesch, M. Fears about AI-mediated communication are grounded in different expectations for one's own versus others' use. Preprint at https://arxiv.org/abs/2305.01670 (2023).
44. Hohenstein, J. et al. Artificial intelligence in communication impacts language and social relationships. *Sci. Rep.* **13**, 5487 (2023).
45. Mieczkowski, H., Hancock, J. T., Naaman, M., Jung, M. & Hohenstein, J. AI-mediated communication: language use and interpersonal effects in a referential communication task. *Proc. ACM Hum. Comput. Interact.* **5**, 17:1–17:14 (2021).
46. Glikson, E. & Asscher, O. AI-mediated apology in a multilingual work context: implications for perceived authenticity and willingness to forgive. *Comput. Hum. Behav.* **140**, 107592 (2023).
47. Hancock, J. T., Naaman, M. & Levy, K. AI-mediated communication: definition, research agenda, and ethical considerations. *J. Comput. Mediat. Commun.* **25**, 89–100 (2020).
48. Mohanasundari, S. K. et al. Can artificial intelligence replace the unique nursing role? *Cureus* **15**, e51150 (2023).
49. Montemayor, C., Halpern, J. & Fairweather, A. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI Soc.* **37**, 1353–1359 (2022).
50. Nass, C. & Moon, Y. Machines and mindlessness: social responses to computers. *J. Soc. Issues* **56**, 81–103 (2000).
51. Reeves, B. & Nass, C. I. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places* xiv, 305 (Cambridge Univ. Press, 1996).
52. Shteynberg, G. et al. Does it matter if empathic AI has no empathy? *Nat. Mach. Intell.* **6**, 496–497 (2024).
53. Perry, A. AI will never convey the essence of human empathy. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-023-01675-w (2023).
54. Haugeland, J. Understanding natural language. *J. Philos.* **76**, 619–632 (1979).
55. Major, B. C., Le Nguyen, K. D., Lundberg, K. B. & Fredrickson, B. L. Well-being correlates of perceived positivity resonance: evidence from trait and episode-level assessments. *Pers. Soc. Psychol. Bull.* **44**, 1631–1647 (2018).
56. Forgas, J. P. & Laham, S. M. in *Cognitive Illusions* (ed. Pohl, R. F.) 276–290 (Psychology Press, 2016).
57. Rubin, M., Arnon, H., Huppert, J. D. & Perry, A. Considering the role of human empathy in AI-driven therapy. *JMIR Ment. Health* **11**, e56529 (2024).
58. Lucas, G. M., Gratch, J., King, A. & Morency, L.-P. It's only a computer: virtual humans increase willingness to disclose. *Comput. Hum. Behav.* **37**, 94–100 (2014).
59. Bhattacharya, K., Ghosh, A., Monsivais, D., Dunbar, R. & Kaski, K. Absence makes the heart grow fonder: social compensation when failure to interact risks weakening a relationship. *EPJ Data Sci.* **6**, 1 (2017).
60. Huxhold, O., Fiori, K. L. & Windsor, T. Rethinking social relationships in adulthood: the differential investment of resources model. *Pers. Soc. Psychol. Rev.* **26**, 57–82 (2022).
61. Jakesch, M., Hancock, J. T. & Naaman, M. Human heuristics for AI-generated language are flawed. *Proc. Natl Acad. Sci. USA* **120**, e2208839120 (2023).
62. Haim, G. B. et al. Empathy and clarity in GPT-4-generated emergency department discharge letters. Preprint at *medRxiv* https://doi.org/10.1101/2024.10.07.24315034 (2024).
63. Schork, N. J. Artificial intelligence and personalized medicine. *Cancer Treat. Res.* **178**, 265–283 (2019).
64. Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S. & Torous, J. B. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *Can. J. Psychiatry* **64**, 456–464 (2019).
65. Felzmann, H., Villaronga, E. F., Lutz, C. & Tamò-Larrieux, A. Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Soc.* **6**, 2053951719860542 (2019).
66. Laux, J., Wachter, S. & Mittelstadt, B. Three pathways for standardisation and ethical disclosure by default under the European Union Artificial Intelligence Act. *Comput. Law Secur. Rev.* **53**, 105957 (2024).
67. Leib, M., Köbis, N., Rilke, R. M., Hagens, M. & Irlenbusch, B. Corrupted by algorithms? How AI-generated and human-written advice shape (dis)honesty. *Econ. J.* **134**, 766–784 (2024).
68. Li, J. Z., Herderich, A. & Goldenberg, A. Skill but not effort drive GPT overperformance over humans in cognitive reframing of negative scenario. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/fzvd8 (2024).
69. Brinkmann, L. et al. Machine culture. *Nat. Hum. Behav.* **7**, 1855–1868 (2023).
70. Brysbaert, M. How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *J. Cogn.* **2**, 16 (2019).
71. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2024); http://www.R-project.org/
72. Hughes, M. E., Waite, L. J., Hawkley, L. C. & Cacioppo, J. T. A short scale for measuring loneliness in large surveys: results from two population-based studies. *Res. Aging* **26**, 655–672 (2004).

Acknowledgements

This work was supported in part by grants from the Mind and Life Institute and the Azrieli Israel Center for Addiction and Mental Health to A.P., and a fellowship from the Azrieli Israel Center for Addiction and Mental Health to M.R. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M.R., J.L., F.Z., A.G. and A.P. were involved in the experimental design and project planning. M.R., D.C.O., A.G. and A.P. contributed to the data analyses. M.R., A.G. and A.P. wrote the paper. All authors reviewed and edited the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02247-w.

Correspondence and requests for materials should be addressed to Matan Rubin or Anat Perry.

Peer review information *Nature Human Behaviour* thanks Corina Pelau and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Matan Rubin, Anat Perry

Last updated by author(s): Jan 24, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Participants were recruited using Prolific. Data collection was done through Qualtrics online survey software version 11-23 for study 1a, version 12-23 for study 1b, version 10-24 for studies 1c, 2a, and 2b, version 11-24 for study 1d, and version 01-24 for studies 3, 4 and 5. In these studies we utilized the web services function to access GPT API to access GPT models 4, 4o, and o1, and Fireworks API to access Llama 3.1 - 405B.
Data analysis	Data analysis was conducted using R Version 4.3.3. Analyses codes are available in OSF depository: https://osf.io/w4hkd/?view_only=52a2324d36bc4c03ad9f1d90ba75ab7b

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Processed data is available here: https://osf.io/w4hkd/?view_only=52a2324d36bc4c03ad9f1d90ba75ab7b

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We did not limit participation at all based on sex or gender. We did not collect any data regarding biological sex, only on gender identity, which was collected as part of a demographic questionnaire at the end of the study. We did not include gender in any of our analyses, but only used it to report basic demographic information in our methods section.
Reporting on race, ethnicity, or other socially relevant groupings	We did not limit participation at all based on such variables. The only socially relevant variables we collected were total years in education and participants' native tongue. Both of these were collected as part of a demographic questionnaire at the end of the study. We also received participants' ethnicity as reported by them on Prolific. We did not include any of these in our analyses, nor use them in any way.
Population characteristics	All participants were from the UK. Their characteristics are as follows: Study 1a: Mean age = 28.42 ± 4.51 (SD), 53% female. Mean years of education = 15.78 ± 3.10 (SD), 77% White, 9% Asian, 9% Black, 4% mixed, 1% other. Study 1b: Mean age = 28.06 ± 4.57 (SD), 53% female. Mean years of education = 16.15 ± 3.08 (SD), 74% White, 9% Asian, 8% Black, 6% mixed, 2% other, 1% unreported). Study 1c: Mean age = 28.37 ± 4.68 (SD), 52% female. Mean years of education = 15.51 ± 2.93 (SD), 53% White, 10% Asian, 20% Black, 12% mixed, 5% other Study 1d: Mean age = 27.63 ± 4.60 (SD), 51% female. Mean years of education = 15.70 ± 3.33 (SD), 56% White, 12% Asian, 25% Black, 5% mixed, 1% other, 1% unreported. Study 2a: Mean age = 28.10 ± 4.64 (SD), 48% female. Mean years of education = 15.50 ± 2.93 (SD), 57% White, 12% Asian, 16% Black, 9% mixed, 4% other, 2% unreported. Study 2b: Mean age = 28.26 ± 4.68 (SD), 49% female. Mean years of education = 15.43 ± 2.84 (SD), 55% White, 6% Asian, 23% Black, 10% mixed, 5% other, 1% unreported. Study 3: Mean age = 28.16 ± 4.48 (SD), 54% female. Mean years of education = 15.97 ± 3.15 (SD), 79% White, 9% Asian, 7% Black, 4% mixed, 1% other, 1% unreported. Study 4: Mean age = 28.05 ± 4.62 (SD), 50% female. Mean years of education = 15.96, ± 3.38 (SD), 76% White, 8% Asian, 8% Black, 5% mixed, 2% other, 1% unreported. Study 5: Mean age = 27.91 ± 4.68 (SD), 47% female. Mean years of education = 15.97 ± 3.29 (SD), 74% White, 10% Asian, 8% Black, 6% mixed, 1% other.
Recruitment	All participants were recruited via Prolific and received pay on a rate of 9 GBP per hour. The study did not mention AI on the prolific platform, but did mention sharing an event from the participants' life. As such, the sample may be biased in that the population on prolific may differ than the general population broadly, and specifically there is a possibility of a self-selection bias in that these participants may have specifically wanted to share something. However, they were not asked to describe negative or difficult events (and many did not), or events they wanted advice on, thus generalizing our results to any sort of personal interaction which involves sharing between humans and AI.
Ethics oversight	Ethics committee of the social sciences faculty of the Hebrew University of Jerusalem

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We ran a series of quantitative studies to test the effects of the perceived source of empathy as artifical or human on the level of empathy perceived, and the value participants regard human empathy as opposed to artificial empathy.
Research sample	The research samples were non-representative, gender-balanced, random samples of prolific workers from the UK, aged 18-35. We aimed for a general sample of the population, without many filtering criteria, to try and gather a wider perspective on the issue. We did filter by age to avoid major differences is technological literacy. Demographic information follows: Study 1a: Mean age = 28.42 ± 4.51 (SD), 53% female. Mean years of education = 15.78 ± 3.10 (SD), 77% White, 9% Asian, 9% Black, 4% mixed, 1% other. Study 1b: Mean age = 28.06 ± 4.57 (SD), 53% female. Mean years of education = 16.15 ± 3.08 (SD), 74% White, 9% Asian, 8% Black, 6% mixed, 2% other, 1% unreported). Study 1c: Mean age = 28.37 ± 4.68 (SD), 52% female. Mean years of education = 15.51 ± 2.93 (SD), 53% White, 10% Asian, 20% Black, 12% mixed, 5% other Study 1d: Mean age = 27.63 ± 4.60 (SD), 51% female. Mean years of education = 15.70 ± 3.33 (SD), 56% White, 12% Asian, 25% Black, 5% mixed, 1% other, 1% unreported. Study 2a: Mean age = 28.10 ± 4.64 (SD), 48% female. Mean years of education = 15.50 ± 2.93 (SD), 57% White, 12% Asian, 16% Black, 9% mixed, 4% other, 2% unreported. Study 2b: Mean age = 28.26 ± 4.68 (SD), 49% female. Mean years of education = 15.43 ± 2.84 (SD), 55% White, 6% Asian, 23% Black, 10% mixed, 5% other, 1% unreported. Study 3: Mean age = 28.16 ± 4.48 (SD), 54% female. Mean years of education = 15.97 ± 3.15 (SD), 79% White, 9% Asian, 7% Black, 4% mixed, 1% other, 1% unreported. Study 4: Mean age = 28.05 ± 4.62 (SD), 50% female. Mean years of education = 15.96, ± 3.38 (SD), 76% White, 8% Asian, 8% Black, 5% mixed, 2% other, 1% unreported. Study 5: Mean age = 27.91 ± 4.68 (SD), 47% female. Mean years of education = 15.97 ± 3.29 (SD), 74% White, 10% Asian, 8% Black, 6% mixed, 1% other.
Sampling strategy	Sampling procedure was random. In study 1a, as we had no previous reference for similar effects, we started with power analyses for an effect size of cohen's d 0.2 at 80% power, using R. Results showed we would need 394 participants per group, and accounting for dropout, we recruited 998. We had more dropout than expected, but reached a close sample size in initial recruitment. In study 1b we relied on the larger effect sizes we saw in study 1, and following a power analysis for that effect size (cohen's d = 0.36), determined a total sample size of 600 participants was necessary. Accounting for dropout, we aimed to recruit 800. Following a small technical error, we recruited a total of 827, leading to a final sample size of 593 participants after dropouts. Study 1c: Following power analyses looking for an effect size of Cohen's d = 0.2 at 80% power, we aimed to reach a final sample of 800 participants. We recruited a total of 999 participants, leading to a final sample of 734. Study 1d: Following power analyses looking for an effect size of Cohen's d = 0.2 at 80% power, we aimed to reach a final sample of 800 participants. We recruited a total of 994 participants, leading to a final sample of 705. Study 2a: Following power analyses looking for an effect size of Cohen's d = 0.2 at 80% power, we aimed to reach a final sample of 800 participants. We recruited a total of 1003 participants, leading to a final sample of 687. Study 2b: Following power analyses looking for an effect size of Cohen's d = 0.2 at 80% power, we aimed to reach a final sample of 800 participants. We recruited a total of 979 participants, leading to a final sample of 691. In study 3 w aimed to find complicated interaction effects, and a simulated interaction power analysis showed we would require 1200 participants to find a small interaction effect at 0.8 power. We recruited 1550 to account for droupouts, based on previous dropout rates, leading to a final sample size of 1172 participants. In studies 4 &5 we also had no frame of reference, so following Brysbaert (2019), we used 100 participants per condition, and aimed to recruit 500 participants in each study. The final sample size were 489 participants in study 4, and 486 in study 5.
Data collection	All studies were done online via Qualtrics, with no possible intervention of an experimenter. Participants were offered to participate on prolific, and after providing informed consent, were directed to share an emotional event that happened to them recently. They then received a response presented as AI-generated or human-authored in studies (1a-3), or chose between an AI and human response or confirmation (in studies 4-5, respectively). In studies 1a-3 they then rated the empathy and emotional experience they had reading the response.
Timing	Study 1a: Nov 20th 2023 Study 1b: Dec 28th 2023 - Jan 2nd 2024 Study 1c: Oct 31st - Nov 5th 2024 Study 1d: Nov 15-19th 2024 Study 2a: Oct 23-29th 2024 Study 2b: Oct 30-31st 2024 Study 3: Jan 10-19th 2024 Study 4: Jan 25th - 30th 2024 Study 5: Jan 29th - Feb 4th, 2024
Data exclusions	For all analyses, we removed participants who rated the relevant dependent variable to a degree further than 2.5. General empathy: In Study 1a we removed 31 participants as outliers. In Study 1b, we removed 18 participants as outliers. In Study 1c, we removed 21 participants as outliers. In Study 1d, we removed 28 participants as outliers. In Study 2a, no participants were removed as outliers.

In Study 2b, we removed 25 participants as outliers.
In Study 3, we removed 49 participants as outliers.

Aspects of empathy:
In Study 1a we removed 7 participants as outliers.
In Study 1b, we removed 4 participants as outliers.
In Study 1c, we removed 7 participants as outliers.
In Study 1d, we removed 6 participants as outliers.
In Study 2a, no participants were removed as outliers.
In Study 2b, we removed 7 participants as outliers.
In Study 3, we removed 9 participants as outliers.
Positivity resonance:
In Study 1a we removed 29 participants as outliers.
In Study 1b, we removed 21 participants as outliers.
In Study 1c, we removed 32 participants as outliers.
In Study 1d, we removed 18 participants as outliers.
In Study 2a, no participants were removed as outliers.
In Study 2b, we removed 16 participants as outliers.
In Study 3, we removed 22 participants as outliers.

Positive emotions:
In Study 1b we removed 20 participants as outliers.
In Study 1c, we removed 37 participants as outliers.
In Study 1d, we removed 23 participants as outliers.
In Study 2a, no participants were removed as outliers.
In Study 2b, we removed 26 participants as outliers.
In Study 3, we removed 46 participants as outliers.

Negative emotions
In Study 1b we removed 17 participants as outliers.
In Study 1c, we removed 34 participants as outliers.
In Study 1d, we removed 27 participants as outliers.
In Study 2a, we removed 32 participants as outliers.
In Study 2b, we removed 25 participants as outliers.
In Study 3, we removed 54 participants as outliers.

Authenticity
In Study 1b we removed 14 participants as outliers.
In Study 1c, we removed 38 participants as outliers.
In Study 1d, no participants were removed as outliers.
In Study 2a, no participants were removed as outliers.
In Study 2b, we removed 34 participants as outliers.
In Study 3, we removed 55 participants as outliers.
In no analyses were there any outliers for support.

Non-participation

Study 1a: Starting with 998 participants, we filtered out those who failed attention checks (n = 263) or reported technical issues (n = 10), leading to a final sample of 725 participants (mean age = 28.42, SD = 4.51, 53% female, AI condition n = 371).
Study 1b: Starting with 827 participants, we filtered out those who failed attention checks (n = 190) or reported technical issues (n = 44), leading to a final sample of 593 participants (mean age = 28.06, SD = 4.57, 53% female, AI condition n = 296).
Study 1c: Starting with 999 participants, we filtered out those who failed attention checks (n = 254) or reported technical issues (n = 11), leading to a final sample of 734 participants
Study 1d: Starting with 994 participants who finished the study, we filtered out those who failed attention checks (n = 278), explicitly used a bot to write the story (n = 1), participated twice (n = 2), or reported technical errors (n = 8), leading to a final sample of 705 participants.
Study 2a: Starting with 1003 participants, we filtered out those who failed attention checks (n = 294), reported technical issues (n = 21), or participated twice (n = 1), leading to a final sample of 687 participants
Study 2b: Starting with 979 participants, we filtered out those who failed attention checks (n = 266) or reported technical issues (n = 22), leading to a final sample of 691 participants
Study 3: SStarting with 1,550 participants who finished the study, we filtered out those who failed attention checks (n = 354), reported technical errors (n = 21), or participated twice (n = 3), leading to a final sample of 1,172 participants.
In studies 4-5 no meaningful technical errors were reported, and there were no attention checks, so no participants were removed or dropped out, and original sample sizes were kept.

Randomization

Participants were randomly assigned to the AI source or the Human source condition in studies 1-3, and to a specific time frame in studies 4-5.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.