nature machine intelligence

Article

https://doi.org/10.1038/s42256-022-00593-2

Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support

Received: 5 April 2022

Accepted: 14 November 2022

Check for updates

Ashish Sharma[1], Inna W. Lin[1], Adam S. Miner[2,3], David C. Atkins[4] & Tim Althoff[1]

Advances in artificial intelligence (AI) are enabling systems that augment and collaborate with humans to perform simple, mechanistic tasks such as scheduling meetings and grammar-checking text. However, such human–AI collaboration poses challenges for more complex tasks, such as carrying out empathic conversations, due to the difficulties that AI systems face in navigating complex human emotions and the open-ended nature of these tasks. Here we focus on peer-to-peer mental health support, a setting in which empathy is critical for success, and examine how AI can collaborate with humans to facilitate peer empathy during textual, online supportive conversations. We develop HAILEY, an AI-in-the-loop agent that provides just-in-time feedback to help participants who provide support (peer supporters) respond more empathically to those seeking help (support seekers). We evaluate HAILEY in a non-clinical randomized controlled trial with real-world peer supporters on TalkLife ($N = 300$), a large online peer-to-peer support platform. We show that our human–AI collaboration approach leads to a 19.6% increase in conversational empathy between peers overall. Furthermore, we find a larger, 38.9% increase in empathy within the subsample of peer supporters who self-identify as experiencing difficulty providing support. We systematically analyse the human–AI collaboration patterns and find that peer supporters are able to use the AI feedback both directly and indirectly without becoming overly reliant on AI while reporting improved self-efficacy post-feedback. Our findings demonstrate the potential of feedback-driven, AI-in-the-loop writing systems to empower humans in open-ended, social and high-stakes tasks such as empathic conversations.

As artificial intelligence (AI) technologies continue to advance, AI systems have started to augment and collaborate with humans in application domains ranging from e-commerce to health care[1–9]. In many, and especially in high-stakes settings, such human–AI collaboration has proven more robust and effective than completely replacing humans with AI[10,11]. However, such collaboration faces dual challenges of developing human-centred AI models to assist humans and designing human-facing interfaces for humans to interact with the AI[12–17].

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [2]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. [3]Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA. [4]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA. e-mail: althoff@cs.washington.edu

For AI-assisted writing, for instance, we must build AI models that generate actionable writing suggestions and simultaneously design human-facing systems that help people see, understand and act on those suggestions in a just-in-time fashion[17–23]. Though initial systems have been proposed for tasks such as story writing[18] and graphic designing[24], it remains challenging to develop human–AI collaboration for a wide range of open-ended, social and high-stakes tasks, as opposed to simple, mechanistic tasks such as scheduling meetings, checking spelling and grammar and booking flights and restaurants.

In this paper, we focus on text-based, peer-to-peer mental health support and investigate how AI systems can collaborate with humans to help facilitate the expression of empathy in textual supportive conversations. Empathy is the ability to understand and relate to the emotions and experiences of others and to effectively communicate that understanding[25]. Empathic support is one of the critical factors that contribute to successful conversations in mental health support, showing strong correlations with symptom improvement[26] and the formation of alliance and rapport[25,27–29]. While online peer-to-peer platforms such as TalkLife (talklife.com) and Reddit (reddit.com) enable such supportive conversations between support seekers (people who seek support) and peer supporters (people who provide support) in non-clinical contexts, highly empathic conversations are rare on these platforms[29]. Peer supporters are typically untrained in expressing complex and open-ended skills such as empathy[30–33] and may lack the required expertise. With an estimated 400 million people suffering from mental health disorders worldwide[34], combined with a pervasive lack of qualified mental health professionals[35,36], these platforms have pioneered avenues for seeking social support and discussing mental health issues for millions of people[37]. However, the challenge lies in improving conversational quality by encouraging untrained peer supporters to adopt complicated and nuanced skills such as empathic writing.

As shown in prior work[38,39], untrained peer supporters report difficulties in writing supportive, empathic responses to support seekers. Without deliberate training or specific feedback, this difficulty persists over time[29,40,41] and may even lead to a gradual decrease in supporters' effectiveness due to factors such as empathy fatigue[42–44]. Furthermore, current efforts to improve empathy (for example, in-person empathy training) do not scale to the millions of peer supporters providing support online. Thus, empowering peer supporters with automated, actionable, just-in-time feedback and training, for example, through human–AI collaboration systems, can help them express higher levels of empathy and, as a result, improve the overall effectiveness of these platforms[29,45–47].

To this end, we develop and evaluate a human–AI collaboration approach for helping untrained peer supporters write more empathic responses in online, text-based peer-to-peer support. We propose Human–AI coLlaboration approach for EmpathY (HAILEY), an AI-in-the-loop agent that offers just-in-time suggestions to express empathy more effectively in conversations (Fig. 1b,c). We design HAILEY to be collaborative, actionable and mobile friendly (Methods).

Unlike the AI-only task of empathic dialogue generation (generating empathic responses from scratch)[48–50], HAILEY adopts a collaborative design that edits existing human responses to make them more empathic[47]. This design reflects the high-stakes setting of mental health, where AI is probably best used to augment rather than replace human skills[46,51]. Furthermore, while current AI-in-the-loop systems are often restricted in the extent to which they can guide humans (for example, simple classification methods that tell users to be empathic when they are not)[52–55], we ensure actionability by guiding peer supporters with concrete steps they may take to respond with more empathy. HAILEY is designed to suggest the insertion of new empathic sentences or replacement of existing low-empathy sentences with more empathic counterparts (Fig. 1c). For complex, hard-to-learn skills such as empathy, this enables just-in-time suggestions on not just 'what' to improve but on 'how' to improve it.

We consider the general setting of text-based, asynchronous conversations between a support seeker and a peer supporter (Fig. 2). In these conversations, the support seeker authors a post seeking mental health support (for example, 'My job is becoming more and more stressful with each passing day.') to which the peer supporter writes a supportive response (for example, 'Don't worry! I'm there for you.'). In this context, we support the peer supporters by providing just-in-time AI feedback to improve the empathy of their responses. To do so, HAILEY prompts the peer supporter through a pop-up ('Would you like some help with your response?') placed above the response text box. On clicking this prompt, HAILEY shows just-in-time AI feedback consisting of 'Insert' (for example, insert 'Have you tried talking to your boss?' at the end of the response) and 'Replace' (for example, replace 'Don't worry!' with 'It must be a real struggle!') suggestions based on the original seeker post and the current peer supporter response. The peer supporter can incorporate these suggestions by directly clicking on the appropriate 'Insert' or 'Replace' button, by further editing them and/or by deriving ideas from the suggestions to indirectly use in their response. These suggestions are generated using PARTNER (emPAthic RewriTing in meNtal hEalth suppoRt), a deep reinforcement learning model that learns to take sentence-level edits as actions to increase the expressed level of empathy while maintaining conversational quality (Methods)[47,56].

To evaluate HAILEY, we conducted a randomized controlled trial in a non-clinical, ecologically informed setting with peer supporters as participants ($N = 300$; Supplementary Table S1[57]), recruited from a large peer-to-peer support platform (TalkLife, talklife.com). Our study was performed outside the TalkLife platform to ensure platform users' safety but adopted an interface similar to TalkLife's chat feature (Fig. 1 and Methods). We employed a between-subjects study design, where each participant was randomly assigned to one of two conditions: human + AI (treatment, with feedback) or human only (control, without feedback).

While peer supporters do not typically receive empathy training from these platforms, we provided participants in both the human + AI (treatment) and human only (control) groups with basic training on empathy, which included empathy definitions, frameworks and examples, just before the main study procedure of writing supportive, empathic responses (Supplementary Fig. S1). This let us conservatively estimate the effect of just-in-time feedback beyond traditional, offline feedback or training (Discussion). During the study, each participant was asked to write supportive, empathic responses to a unique set of ten existing seeker posts (one at a time) that were sourced at random from a subset of TalkLife posts. We filtered out harm-related content, such as suicidal ideation or self-harm, to ensure participant safety (Methods and Discussion). While writing responses, participants in the human + AI (treatment) group received feedback via HAILEY (Fig. 1b,c). Participants in the human-only (control) group, on the other hand, wrote responses but received no feedback, reflecting the current status quo on online peer-to-peer support platforms (Fig. 1a). After completing responses to the ten posts, participants were asked to assess HAILEY by answering questions about the challenges they experienced while writing responses and the effectiveness of our approach.

Our primary hypothesis was that human–AI collaboration would lead to more empathic responses, that is, that responses in the human + AI (treatment) group would show higher empathy than those in the human-only (control) group. We evaluated this hypothesis using both human and automatic evaluation, which helped us capture platform users' perceptions and provided a theory-based assessment of empathy in the collected responses, respectively (Methods). Note that, due to the sensitive mental health context and for reasons of safety, our evaluation of empathy was only based on empathy that was expressed in responses and not the empathy that might have been perceived by the support seeker of the original seeker post[58]. Psychotherapy research indicates a strong correlation between expressed empathy and positive

https://doi.org/10.1038/s42256-022-00593-2

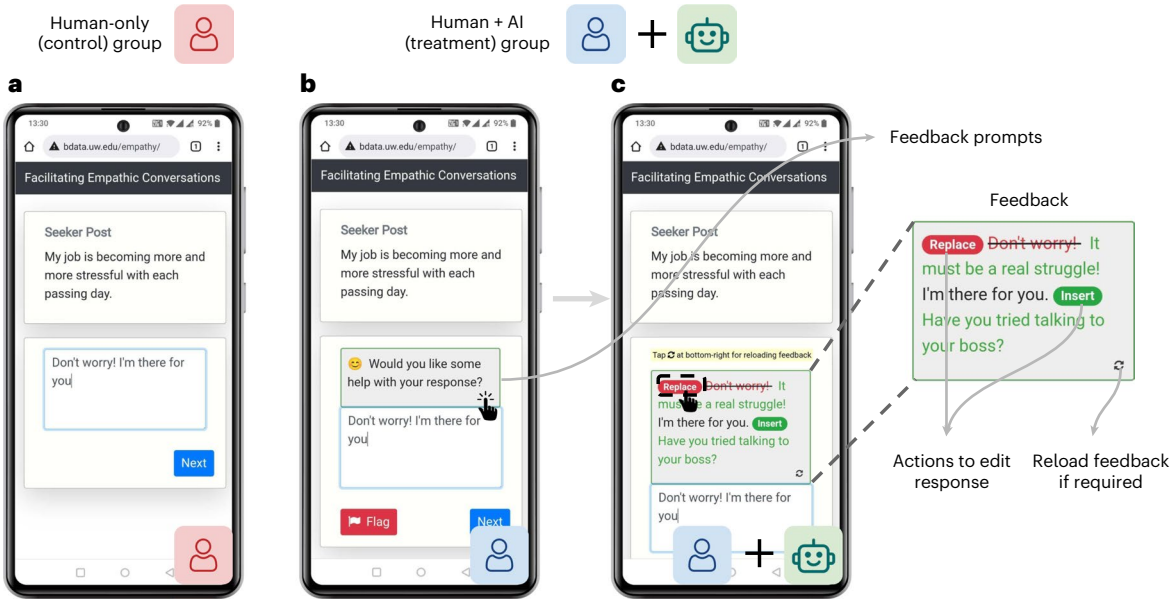

Fig. 1 | A randomized controlled trial with 300 TalkLife peer supporters as participants. We randomly divided participants into human only (control) and human + AI (treatment) groups and asked them to write supportive, empathic responses to seeker posts without feedback and with feedback, respectively. To identify whether just-in-time human–AI collaboration helped increase expressed empathy beyond potential (but rare) traditional training methods, participants in both groups received initial empathy training before starting the study (Methods and Supplementary Fig. S1). **a**, Without AI, human peer supporters are presented with an empty chatbox to author their response (the current status quo). As peer supporters are typically untrained on best practices in therapy (such as empathy) they rarely conduct highly empathic conversations. **b**, Our feedback agent (HAILEY) prompts peer supporters for providing just-in-time AI feedback as they write their responses. **c**, HAILEY then suggests changes that can be made to the response to make it more empathic. These suggestions include new sentences that can be inserted and options for replacing current sentences with more empathic counterparts. Participants can accept these suggestions by clicking on the 'Insert' and 'Replace' buttons and continue editing the response, or get more feedback if needed.

therapeutic outcomes and commonly uses it as a credible alternative[25] (Methods and Discussion).

We conducted multiple post hoc evaluations to assess whether the participants who self-reported challenges in writing supportive responses could benefit more from our system, to investigate the differences in how participants collaborated with the AI and to assess the participants' perceptions of our approach.

Results

Increase in expressed empathy due to human–AI collaboration

Our primary finding is that providing just-in-time AI feedback to participants leads to more empathic responses (Fig. 2). Specifically, through human evaluation from an independent set of TalkLife users (Methods), we found that the human + AI responses were rated as being more empathic than the human-only responses 46.8% of the time and were rated equivalent in empathy to human-only responses 15.7% of the time. On the other hand, human-only responses were preferred only 37.4% of the time ($P = 3.4 \times 10^{-5}$, $t = 4.15$, d.f. = 2,998, two-sided Student's t-test; Fig. 2a). In addition, by automatically estimating empathy levels of responses using a previously validated empathy classification model on a scale from 0 to 6 (Methods), we found that the human + AI approach led to 19.6% higher empathic responses compared with the human-only approach (1.77 versus 1.48; Cohen's $d = 0.24$, $P = 5.1 \times 10^{-8}$, $t = 5.46$; d.f. = 2,998, two-sided Student's t-test; Fig. 2b).

Higher gains for those who report peer support challenges

Prior work has shown that online peer supporters find it extremely challenging to write supportive and empathic responses[29,38,39]. Some participants have little to no prior experience with peer support (for example, if they are new to the platform; $N = 95/300$; Methods). Even as the participants gain more experience, in the absence of explicit training or feedback, the challenge of writing supportive responses persists over time and may even lead to a gradual decrease in empathy levels due to factors such as empathy fatigue[40–44], as also observed during the course of our 30-min study (Supplementary Fig. S6). Therefore, it is particularly important to better assist the many participants who struggle with writing responses.

For the subsample of participants who self-reported challenges in writing responses at the end of our study ($N = 91/300$; Methods), a post hoc analysis revealed significantly higher empathy gains when using the human–AI collaboration approach. For such participants, we found an absolute 4.5% stronger preference for the human + AI responses (49.1% versus 44.6%; $P = 5.6 \times 10^{-5}$, $t = 4.05$, d.f. = 718, two-sided Student's t-test; Fig. 2c) and a 27.0% higher increase in expressed empathy when using the human + AI approach (38.8% versus 11.8%; $P = 5.5 \times 10^{-9}$, $t = 5.90$; d.f. = 1,818, two-sided Student's t-test; Fig. 2d) compared with participants who did not report any challenges. For the subsample of participants who self-reported no previous experience with online peer support at the start of our study ($N = 95/300$; 37 of these participants also self-reported challenges), we found a 8.1% stronger preference for the human + AI responses (51.8% versus 43.7%; $P = 1.4 \times 10^{-6}$, $t = 4.84$, d.f. = 758, two-sided Student's t-test;) and a 21.2% higher increase in expressed empathy when using the human + AI approach (33.7% versus 12.5%; $P = 5.5 \times 10^{-6}$, $t = 4.58$, d.f. = 1,898, two-sided Student's t-test; Supplementary Fig. S11d) compared with participants who reported experience with online peer support.

Patterns of human–AI collaboration

Collaboration between humans and AI can take many forms, but specific formulations of human–AI collaboration remain poorly defined and challenging to measure[11]. Investigating how humans collaborate with our AI can help us better understand the system's use-cases and better inform design decisions. Here, we analysed the collaboration patterns of participants both over the course of the study as well as during a single response instance. We leveraged this analysis to derive

https://doi.org/10.1038/s42256-022-00593-2

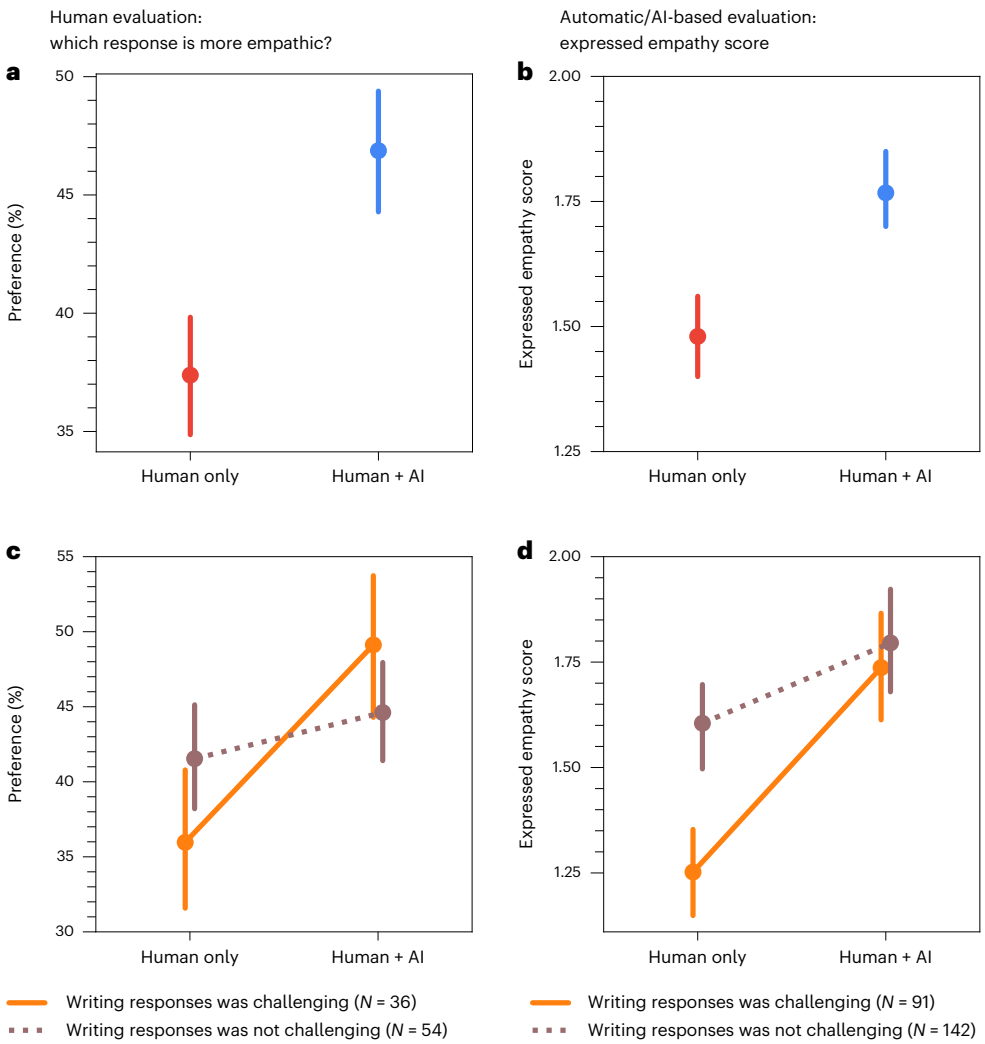

Fig. 2 | Randomized controlled trial demonstrates that Human–AI collaboration enables more empathic conversations. a, Human evaluation from an independent set of TalkLife users showed that the human + AI responses ($N = 139$) were strictly preferred 46.9% of the time relative to a 37.4% strict preference for the human-only responses ($N = 161$). **b**, Through automatic evaluation using an AI-based expressed empathy score[29], we found that the human + AI responses ($N = 139$) had 19.6% higher empathy than the human-only responses ($N = 161$; 1.77 versus 1.48; Cohen's $d = 0.24$, $P = 5.1 \times 10^{-8}$, $t = 5.46$, d.f. = 2,998, two-sided Student's t-test). **c**, For the participants who reported challenges in writing responses after the study, we found a stronger preference for the human + AI responses versus human-only responses (49.1% versus 34.0%), compared with participants who did not report challenges (44.6% versus 41.5%). **d**, For participants who reported challenges in writing responses after the study, we found a higher improvement in expressed empathy scores of the human + AI responses versus human-only responses (38.9%; 1.74 versus 1.25; Cohen's $d = 0.43$), compared with participants who did not report challenges (11.9%; 1.79 versus 1.60; Cohen's $d = 0.15$). In **c** and **d**, the sample size varied to ensure comparable conditions (Methods). The point estimates represent the mean. The error bars represent bootstrapped 95% confidence intervals.

a hierarchical taxonomy of human–AI collaboration patterns based on how often the AI was consulted during the study and how AI suggestions were used (Fig. 3a and Methods).

Our analysis revealed several categories of collaboration. For example, some participants chose to always rely on the AI feedback, whereas others only utilized it as a source of inspiration and rewrote it in their own style. Based on the number of posts in the study for which AI was consulted (out of the ten posts for each participant), we found that participants consulted AI either always (15.5%), often (56.0%), once (6.0%) or never (22.4%). Very few participants always consulted and used the AI (2.6%), indicating that they did not rely excessively on AI feedback. A substantial number of participants also chose to never consult the AI (22.4%). Such participants, however, also expressed the least empathy in their responses (1.13 on average out of 6; Fig. 3b), suggesting that consulting the AI could have been beneficial.

Furthermore, based on how AI suggestions were used, we found that participants used the suggestions either directly (64.6%), indirectly (18.5%) or not at all (16.9%). As expected given our system's design, the most common way of usage was direct, which entailed clicking on the suggested actions to incorporate them into the response. In contrast, participants who indirectly used AI (Methods) drew ideas from the suggested feedback and rewrote it in their own words in the final response. Some participants, however, chose not to use suggestions at all (16.9%). A review of these instances by the researchers, as well as the subjective feedback from participants, suggested that reasons included the feedback not being helpful, the feedback not being personalized or their response already being empathic and leaving little room for improvement. Finally, multiple types of feedback are possible for the same combination of seeker post and original response, and some participants (16.9%) used our reload functionality (Methods) to read through these multiple suggestions before they selected a final response.

https://doi.org/10.1038/s42256-022-00593-2

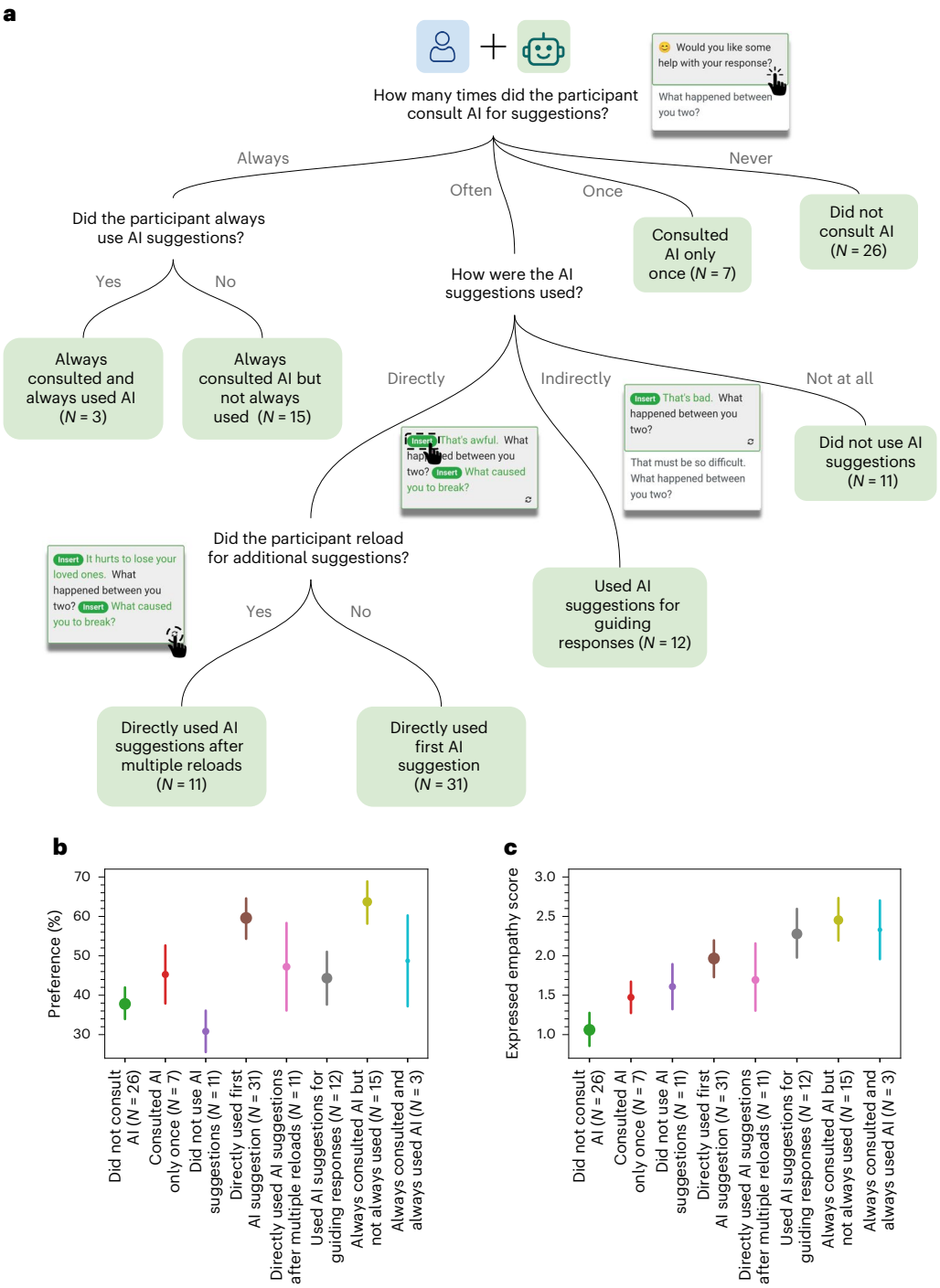

Fig. 3 | The derived hierarchical taxonomy of human–AI collaboration categories. a, We clustered the interaction patterns of human + AI (treatment) participants based on how often the AI was consulted during the study and how the AI suggestions were used ($N = 116/139$). We excluded participants who belonged to multiple clusters ($N = 23/139$). Very few participants always consulted and used AI (2.6%), indicating that participants did not rely excessively on AI feedback. Participants could use AI feedback directly through suggested actions (64.6%) or indirectly by drawing ideas from the suggested feedback and rewriting it in their own words in the final response (18.5%). **b**, Empathy increased when participants consulted and used AI more frequently, with those who did not consult AI (22.4%) or did not use AI (9.5%) having significantly lower preference over human-only responses ($N = 37$; $P = 6.4 \times 10^{-6}$, two-sided Student's *t*-test). **c**, Participants who did not consult AI had the lowest empathy levels based on our automatic evaluation (1.13 on average out of 6). The area of the points is proportional to the number of participants in the respective human–AI collaboration categories. The point estimates represent the mean. The error bars represent bootstrapped 95% confidence intervals.

In general, participants who consulted and used AI more often expressed higher empathy, though this pattern was more pronounced when evaluated through our automatic expressed empathy score (Fig. 3c) than through human evaluation (Fig. 3b).

Increase in participants' confidence to provide support

At the end of our study, we collected study participants' perceptions about the usefulness and actionability of the feedback and their intention to adopt the system. We observed that 63.3% of participants found

the feedback they received helpful, 60.4% found it actionable and 77.7% of participants wanted this type of feedback system to be deployed on TalkLife or other similar peer-to-peer support platforms (Supplementary Fig. S3), indicating the overall effectiveness of our approach. We also found that 69.8% of participants self-reported feeling more confident at providing support after our study. This indicates the potential value of our system for training and increased self-efficacy (Supplementary Fig. S3).

Discussion

Our work demonstrates how humans and AI might collaborate on open-ended, social and high-stakes tasks such as conducting empathic conversations. Empathy is complex and nuanced[30–33] and is thus more challenging for AI than many other human–AI collaboration tasks such as scheduling meetings, therapy appointments and checking grammar in text. We show how the joint effects of humans and AI can be leveraged to help peer supporters, especially those who have difficulty providing support, converse more empathically with those seeking mental health support.

Our study has implications for addressing barriers to mental health care, where existing resources and interventions are insufficient to meet the current and emerging need. According to a World Health Organization report, over 400 million people globally suffer from a mental health disorder, with approximately 300 million suffering from depression[34]. Overall, mental illness and related behavioural health problems contribute 13% to the global burden of disease, more than both cardiovascular diseases and cancer[59]. Although psychotherapy and social support[60] can be effective treatments, many vulnerable individuals have limited access to therapy and counselling[35,36]. For example, most countries have less than one psychiatrist per 100,000 individuals, indicating widespread shortages of workforce and inadequate in-person treatment options[61].

A scalable approach to improving access to mental health support globally is by connecting support seekers to peer supporters using online platforms such as TalkLife (talklife.com), YourDost (yourdost.com) or Mental Health Subreddits (reddit.com)[37]. However, a key challenge in doing so lies in enabling effective and high-quality conversations between untrained peer supporters and those in need, at scale. We show that human–AI collaboration can considerably increase empathy in peer supporter responses, a core component of effective and quality support that ensures improved feelings of understanding and acceptance[25,27–29]. While fully replacing humans with AI for empathic care has previously drawn scepticism from psychotherapists[31,32], our results suggest that it is feasible to empower untrained peer supporters with appropriate AI-assisted technologies in relatively lower-risk settings, such as peer-to-peer support[35,45,46,62,63].

Our findings also point to potential secondary gain for peer supporters in terms of (1) increased self-efficacy, as indicated by 69.8% of participants feeling more confident in providing support after the study, and (2) gained experience and expertise by multiple example learning when using the reload functionality to scroll through multiple types of responses for the same seeker post. This has implications for helping untrained peer supporters beyond providing them just-in-time feedback. One criticism of AI is that it may steal or dampen opportunities for training more clinicians and workforce[31,32]. We show that human–AI collaboration can actually enhance, rather than diminish, these training opportunities. This is also reflected in the subjective feedback from participants (Methods), with several participants reporting different types of learning after interacting with the AI (for example, one participant wrote, 'I realized that sometimes I directly jump on to suggestions rather than being empathic first. I will have to work on it.', while another wrote, 'Feedback in general is helpful. It promotes improvement and growth.').

Further, we find that participants not only directly accept suggestions but also draw higher-level inspiration from the suggested feedback (for example, a participant wrote 'Sometimes it just gave me direction on what [should] be said and I was able to word it my way. Other times it helped add things that made it sound more empathic...'), akin to having access to a therapist's internal brainstorming, which participants can use to rewrite responses in their own style.

In our study, many participants ($N = 91$) reported challenges in writing responses (for example, several participants reported not knowing what to say: 'I sometimes have a hard time knowing what to say.'), which is characteristic of the average user on online peer-to-peer support platforms[29,38,39]. We demonstrate a significantly larger improvement in empathy for these users, suggesting that we can provide significant assistance in writing more empathic responses, thereby improving the empathy awareness and expression of the typical platform user[29,47]. Through qualitative analysis of such participants' subjective feedback on our study, we find that HAILEY can guide someone who is unsure about what to say (for example, a participant wrote, 'Feedback gave me a point to start my own response when I didn't know how to start.') and can help them frame better responses (for example, one participant wrote, 'Sometimes I didn't really knew [sic] how to form the sentence but the feedback helped me out with how I should incorporate the words.', while another wrote, 'Sometimes I do not sound how I want to and so this feedback has helped me on sounding more friendly...'; Methods). One concern is that AI, if used in practice, may cause harm to multiple stakeholders from support seekers to care providers and peer supporters, through inappropriate interventions, role confusion and data sharing concerns[31,32,64]. According to our findings, however, the individuals struggling to do a good job are the ones who benefit the most, which forms an important use-case of AI in health care.

The reported difference in empathy between the treatment and control groups conservatively estimates the impact of our AI-in-the-loop feedback system due to (1) additional initial empathy training to both human + AI and human-only groups (Supplementary Fig. S1) and (2) a potential selection effect that may have attracted TalkLife users who care more about supporting others (Supplementary Fig. S7). In practice, training of peer supporters is very rare, and the effect of training typically diminishes over time[42,43]. We included this training to understand whether just-in-time AI feedback is helpful beyond traditional training methods. Moreover, the human-only responses in our study had 34.5% higher expressed empathy than existing human-only responses to the corresponding seeker posts on the TalkLife platform (1.11 versus 1.48; $P \ll 10^{-5}$, two-sided Student's t-test; Supplementary Fig. S7), reflecting the effects of additional training as well as a potential selection effect. We show here that human–AI collaboration improves empathy expression even for participants who already express empathy more often. Practical gains for the average user of the TalkLife platform could be even higher than the intentionally conservative estimates presented here.

Safety, privacy and ethics

Developing computational methods for intervention in high-stakes settings such as mental health care involves ethical considerations related to safety, privacy and bias[13,65–67]. There is a risk that, in attempting to help, AI may have the opposite effect on the potentially vulnerable support seeker or peer supporter[64]. The present study included several measures to reduce such risks and unintended consequences. First, our collaborative, AI-in-the-loop writing approach ensured that the primary conversation remains between two humans, with AI offering feedback only when it appears useful, and allowing the human supporter to accept or reject it. Providing such human agency is safer than relying solely on AI, especially in a high-stakes mental health context[46]. Moreover, using only AI results in a loss of authenticity in the responses. Hence, our human–AI collaboration approach leads to responses with high empathy as well as high authenticity (Supplementary Fig. S2).

Second, our approach intentionally assists only peer supporters, not support seekers in crisis, since they are likely to be at a lower risk

and more receptive to the feedback. Third, we filtered posts related to suicidal ideation and self-harm by using pre-defined unsafe regular expressions (for example, '.*(commit suicide).*' and '.*(cut).*'). Such posts did not enter our feedback pipeline, but instead we recommended escalating them to therapists. We applied the same filtering to every generated feedback, as well, to try and ensure that HAILEY did not suggest unsafe text as responses. Fourth, such automated filtering may not be perfect. Therefore, we included a mechanism to flag inappropriate/unsafe posts and feedback by providing our participants with an explicit 'Flag' button (Supplementary Fig. S31). In our study, 1.6% posts (out of 1,390 in the treatment group) and 2.9% feedback instances (out of 1,939 requests) were flagged as inappropriate or unsafe. While the majority of them were concerned with unclear seeker posts or irrelevant feedback, we found six cases (0.2%) that warranted further attention. One of these cases involved the post containing intentionally misspelled self-harm content (for example, 'c u t' with spaces between letters in order to circumvent safety filters). Another related to feedback containing a self-harm-related term. Three addressed the post or feedback containing a swear word that may not directly be a safety concern (for example, 'You are so f**king adorable'), and one contained toxic/offensive feedback ('It's a bad face').

Future iterations of our system could address these issues by leveraging more robust filtering methods and toxicity/hate speech classifiers (for example, Perspective API (perspectiveapi.com)). Several platforms, including TalkLife, already have systems in place to prevent triggering content from being shown, which can be integrated into our system on deployment. Finally, we removed all personally identifiable information (user and platform identifiers) from the TalkLife data set prior to training the AI model.

Limitations

While our study results reveal the promise of human–AI collaboration in open-ended and even high-stakes settings, the study is not without limitations. Some of our participants indicated that empathy may not always be the most helpful way to respond (for example, when support seekers are looking for concrete actions). However, as demonstrated repeatedly in the clinical psychology literature[25,27–29], empathy is a critical, foundational approach to all evidence-based mental health support, plays an important role in building alliance and relationship between people and is highly correlated with symptom improvement. It has consistently proven to be an important aspect of responding, but support seekers may sometimes benefit from additional responses involving different interventions (for example, concrete problem solving or motivational interviewing[68]). Future work should investigate when such additional responses are helpful or necessary.

Some participants may have been apprehensive about using our system, as indicated by the fact that many participants did not consult or use it ($N = 37$). Qualitatively analysing the subjective feedback from these participants suggested that this might be due to feedback communicating incorrect assumptions about the preferences, experience and background of participants (for example, assuming that a participant is dealing with the same issues as the support seeker: 'Not sure this can be avoided, but the feedback would consistently assume I've been through the same thing.'). Future work should personalize prompts and feedback to individual participants. This could include personalizing the content and the frequency of the prompt as well as personalizing the type of feedback that is shown from multiple possible feedback options.

Our assessment includes validated yet automated and imperfect measures. Specifically, our evaluation of empathy is based only on empathy that was expressed in responses, not empathy that might have been perceived by the support seeker[58]. In sensitive contexts such as ours, however, obtaining perceived empathy ratings from support seekers is challenging and involves ethical risks (Safety, privacy and ethics section). We attempted to reduce the gap between expressed and perceived empathy in our human evaluation by recruiting participants from TalkLife who may be seeking support on the platform (Methods). Nevertheless, studying the effects of human–AI collaboration on perceived empathy in conversations is a vital future research direction. However, note that psychotherapy research indicates a strong correlation between expressed empathy and positive therapeutic outcomes and commonly uses it as a credible alternative[25,27–29].

Furthermore, we acknowledge that a variety of social and cultural factors might affect the dynamics of the support and the expression of empathy[69–71]. As such, our human–AI collaboration approach must be adapted and evaluated in various socio-cultural contexts, including underrepresented communities and minorities. While conducting randomized controlled trials on specific communities and investigating heterogeneous treatment effects across demographic groups is beyond the scope of our work, our study was deployed globally and included participants of various gender identities, ethnicities, ages and countries (Methods and Supplementary Figs. S10 and S11). However, this is a critical area of research, and ensuring equitable access to culturally sensitive empathic support requires further investigation.

Our study evaluated a single human–AI collaboration interface design, and there could have been other potential interface designs, as well. Additionally, as a secondary exploration, we analysed a classification-based interface design, which provided participants with the option to request automatic expressed empathy scores[29] for their responses (Supplementary Fig. S5). We assigned this secondary classification-based AI treatment to 10% of the incoming participants at random ($N = 30$). Due to conflicting human and automatic evaluation results, we observed that the effects of this secondary treatment on the empathy of participants were ambiguous (Supplementary Fig. S4a,b). However, the design was perceived as being less actionable than our primary rewriting-based interface (Supplementary Fig. S4c). This poses questions regarding what types of design are optimal and how best to provide feedback.

Finally, we recruited participants from a single platform (TalkLife) and only for providing empathic support in the English language. We further note that this study focuses on empathy expression in peer support and does not investigate long-term clinical outcomes.

Conclusion

We developed and evaluated HAILEY, a human–AI collaboration system that led to a 19.6% increase in empathy in peer-to-peer conversations overall (Cohen's $d = 0.24$) and a 38.9% increase in empathy for mental health supporters who experience difficulty in writing responses (Cohen's $d = 0.43$) in a randomized controlled trial on a large peer-to-peer mental health platform. Our findings demonstrate the potential of feedback-driven, AI-in-the-loop writing systems to empower online peer supporters to improve the quality of their responses without increasing the risk of harmful responses.

Methods

Study design

We employed a between-subjects study design in which each participant was randomly assigned to one of human + AI (treatment, $N = 139$) or human-only (control, $N = 161$) conditions. Participants in both groups were asked to write supportive, empathic responses to a unique set of ten existing seeker posts (one at a time), sourced at random from a subset of TalkLife posts. The human + AI (treatment) group participants were given the option of receiving feedback through prompts as they typed their responses. Participants in the human-only (control) group, in contrast, wrote responses with no option for feedback.

TalkLife platform. Founded in 2012, TalkLife (talklife.com/about) is the largest global peer-to-peer support platform for mental health. It enables people in distress to interact with other peers on the platform. Users typically access this platform through a smartphone application,

though a web interface is available as well. The interactions typically occur through conversational threads, which is the focus of our study. A conversational thread on TalkLife is characterized by a user initially authoring a post seeking support (for example, 'My job is becoming more and more stressful with each passing day.'). The post then receives responses from the peers on the platform, sometimes leading to back-and-forth conversations between the users.

Participant recruitment. We worked with TalkLife to recruit participants directly from their platform. Because users on such platforms are typically untrained in best practices of providing mental health support, their work offers a natural place to deploy feedback systems such as ours. To recruit participants, we advertised our study on TalkLife. Recruitment started in April 2021 and continued until September 2021. The study was approved by the University of Washington's Institutional Review Board (determined to be exempt; IRB ID STUDY00012706).

Power analysis. We used a power analysis to estimate the number of participants required for our study. For an effect size of 0.1 difference in empathy, a power analysis with a significance level of 0.05, powered at 80%, indicated that we required 1,500 samples of (seeker post, response post) pairs each for the treatment and control groups. To meet the required sample size, we collected ten samples per participant and therefore recruited 300 participants in total (with the goal of 150 participants per condition), for a total of 1,500 samples each.

Data set of seeker posts. We obtained a unique set of 1,500 seeker posts, sampled at random with consent from the TalkLife platform, in the observation period from May 2012 to June 2020. Prior to sampling, we filtered posts related to (1) critical settings of suicidal ideation and self-harm, using pre-defined unsafe regular expressions (for example, '.*(commit suicide).*' and '.*(cut).*'), to ensure participant safety (Discussion), and (2) common social media interactions not related to mental health (for example, 'Happy mother's day') using a standard text classifier based on BERT (Bidirectional Encoder Representations from Transformers)[72], trained on a manually annotated data set of ~3k posts with answers to the question 'Is the seeker talking about a mental health related issue or situation in his/her post?' (~85% accuracy)[47]. We randomly divided these 1,500 posts into 150 subsets of 10 posts each. We used the same 150 subsets for both treatment and control conditions for consistent context for both groups of participants.

Participant demographics. In our study, 54.3% of the participants identified as female, 36.7% as male and 7.3% as non-binary, while the remaining 1.7% preferred not to report their gender. The average age of participants was 26.3 years (s.d. 9.5 years). Of the participants, 45.7% identified as white, 20.3% as Asian, 10.7% as Hispanic or Latino, 10.3% as Black or African American, 0.7% as Pacific Islander or Hawaiian and 0.3% as American Indian or Alaska Native, while the remaining 12.0% preferred not to report their race/ethnicity. Of the participants, 62.3% were from the United States, 13.7% were from India, 2.3% were from the United Kingdom and 2.3% were from Germany, while the remaining 19.3% were from 36 different countries (spanning six of seven continents, excluding Antarctica). Moreover, 31.7% of the participants reported having no experience with peer-to-peer support despite having been recruited from the TalkLife platform, 26.3% as having less than 1 year of experience, and 42.0% as having ≥1 year of experience with peer-to-peer support.

RCT group assignment. On clicking the advertised pop-up used for recruitment, a TalkLife user was randomly assigned to one of the human + AI (treatment) or human-only (control) conditions for the study duration.

Study workflow. We divided our study into four phases:

- Phase I: pre-intervention survey. First, both control and treatment group participants were asked the same set of survey questions describing their demographics, background and experience with peer-to-peer support (Supplementary Figs. S19 and S20).
- Phase II: empathy training and instructions. Next, to address whether participants held similar understandings of empathy, both groups received the same initial empathy training, which included empathy definitions, frameworks and examples based on psychology theory, before starting the main study procedure of writing empathic responses (Supplementary Fig. S1). Participants were also shown instructions on using our study interface in this phase (Supplementary Figs. S21–S28).
- Phase III: write supportive, empathic responses. Participants then started the main study procedure and wrote responses to one of the 150 subsets of 10 existing seeker posts (one post at a time). For each post, participants in both the groups were prompted 'Write a supportive, empathic response here'. The human + AI (treatment) group participants were given the option of receiving feedback through prompts as they typed their responses (Supplementary Fig. S30). Participants in the human-only (control) group wrote responses without any option for feedback (Supplementary Fig. S29).
- Phase IV: post-intervention survey. After completing the ten posts, participants in both groups were asked to assess the study by answering questions about the difficulty they faced while writing responses, the helpfulness and actionability of the feedback, their self-efficacy after the study and their intent to adopt the system (Supplementary Figs. S32–S34).

If participants dropped out of the study before completing it, their data were removed from our analyses. Participants took 20.6 min on average to complete the study and wrote responses with 25.9 words on average (Supplementary Fig. S9[73]. U.S. citizens and permanent U.S. residents were compensated with a U.S. $5 Amazon gift card. Furthermore, the top two participants in the human evaluation (Evaluation section) received an additional U.S. $25 Amazon gift card. Based on local regulations, we were unable to pay non-U.S. participants. This was explicitly highlighted in the participant consent form on the first landing page of our study (Supplementary Fig. S18 and S35).

Design goals

HAILEY is designed with a collaborative AI-in-the-loop approach, to provide actionable feedback and to be mobile friendly.

Collaborative AI-in-the-loop design. In the high-stakes setting of mental health support, AI is best used to augment rather than replace human skill and knowledge[46,51]. Current natural-language processing technology (including language models, conversational AI methods and chatbots) continue to pose risks related to toxicity, safety and bias, which can be life-threatening in contexts of suicidal ideation and self-harm[64,74–76]. To mitigate these risks, researchers have called for human–AI collaboration methods, where the primary communication remains between two humans with an AI system in the loop to assist humans in improving their conversation[46,51]. In HAILEY, humans remain at the centre of the interaction, receive suggestions from our AI in the loop and retain full control over which suggestions to use in their responses (for example, by selectively choosing the most appropriate 'Insert' or 'Replace' suggestions and editing them as needed).

Actionable feedback. Current AI-in-the-loop systems are often limited to addressing 'what' (rather than 'how') participants should improve[52–55]. For such a goal, it is generally acceptable to design simple interfaces that prompt participants to leverage strategies for successful

supportive conversations (for example, prompting 'you may want to empathize with the user') without any instructions on how to concretely apply those strategies. However, for complex, hard-to-learn constructs such as empathy[25,30], there is a need to address the more actionable goal of steps that participants should take to improve. HAILEY, designed to be actionable, suggests concrete actions (for example, sentences to insert or replace) that participants may take to make their current response more empathic.

Mobile-friendly design. Online conversations and communication are increasingly mobile based. This is also true for peer-to-peer support platforms, which generally provide their services through a smartphone application. Therefore, a mobile-friendly design is critical for the adoption of conversational assistive agents such as ours. However, the challenge here relates to the complex nature of the feedback and the smaller, lower-resolution screen on a mobile device as compared with a desktop. We therefore designed a compact, minimal interface that works equally well on desktop and mobile platforms. We created a conversational experience based on the mobile interface of peer-to-peer support platforms that was design minimal, used responsive prompts that adjusted in form based on screen sizes, placed AI feedback compactly above the response text box for easy access and provided action buttons that were easy for mobile users to click on.

Feedback workflow

Through HAILEY, we showed prompts to participants that they could click on to receive feedback. Our feedback, driven by a previously validated empathic rewriting model, consists of actions that users can take to improve the empathy of their responses (Supplementary Fig. S30).

Prompts to trigger feedback. We showed the prompt 'Would you like some help with your response?' to participants, which was placed above the response text box (Fig. 1b). Participants could at any point click on the prompt to receive feedback on their current response (including when it is still empty). When this prompt is clicked, HAILEY acts on the seeker post and the current response to suggest changes that will make the response more empathic. Our suggestions consisted of 'Insert' and 'Replace' operations generated through empathic rewriting of the response.

Generating feedback through empathic rewriting. The goal of empathic rewriting, originally proposed in ref. 47, is to transform low-empathy to higher-empathy text. The authors proposed PARTNER, a deep reinforcement learning model that learns to take sentence-level edits as actions to increase the expressed level of empathy while maintaining conversational quality. PARTNER's learning policy is based on a transformer language model (adapted from GPT-2 (Generative Pre-trained Transformer 2)[77]), which performs the dual task of generating candidate empathic sentences and adding those sentences at appropriate positions. PARTNER-generated rewritings increase empathy by 1.6 (on the 6-point empathy scale), which is >35% more than all state-of-the-art baseline methods, and are judged more empathic over 65% of the time than baselines by human annotators. Here, we build on PARTNER by further improving training data quality through additional filtering, supporting multiple generations for the real-world use-case of multiple types of feedback for the same post, and evaluating a broader range of hyperparameter choices. Source code of PARTNER was taken from ref. 56.

Displaying feedback as suggested actions. We map the rewritings generated by our optimized version of PARTNER to suggestions to 'Insert' and 'Replace' sentences. These suggestions are then shown as actions to edit the response. To incorporate the suggested changes, the participant clicks on the respective 'Insert' or 'Replace' buttons. Continuing our example from Fig. 1, given the seeker post 'My job is becoming more and more stressful with each passing day.' and the original response 'Don't worry! I'm there for you.', PARTNER takes two insert actions: replace 'Don't worry!' with 'It must be a real struggle!' and Insert 'Have you tried talking to your boss?' at the end of the response. These actions are shown as feedback to the participant. See Supplementary Fig. S8 for more qualitative examples.

Reload feedback if required. For the same combination of seeker post and original response, multiple feedback suggestions are possible. In the Fig. 1 example, instead of suggesting the insert 'Have you tried talking to your boss?', we could also propose inserting 'I know how difficult things can be at work.' These feedback variations can be sampled from our model and, if the initial sampled feedback does not meet participant needs, iterated upon to help participants find better-suited feedback. HAILEY provides an option to reload feedback, allowing participants to navigate through different feedback and suggestions if necessary.

Evaluation

Empathy measurement. We evaluated empathy using both human and automated methods. For our human evaluation, we recruited an independent set of participants from the TalkLife platform and asked them to compare responses written with feedback versus those written without feedback given the same seeker post (Supplementary Figs. S35–S37). We found that the participant annotations had a Cohen's kappa score of 0.55 ($N = 150$ pair of responses; note that Cohen's kappa controls for agreement by chance). We found this score to be comparable to the inter-annotator agreement for complex therapeutic constructs annotations[29,78]. When analysing strata of participants based on challenges in writing responses (Fig. 1c), we considered only those seeker post instances for which the respective human-only and human + AI participants both indicated writing as challenging or not challenging. Since our human evaluation involves comparing human-only and human + AI responses, this ensures that participants in each strata belong to only one of the challenging or not challenging category.

Though our human evaluation captures platform users' perceptions of empathy in responses, it is unlikely to measure empathy from the perspective of psychology theory given the limited training of TalkLife users. Therefore, we conducted a second, complementary evaluation by applying the theory-based empathy classification model proposed by Sharma et al.[29], which assigns a score between 0 and 6 to each response and has been validated and used in prior work[47,79–81]. Note that this approach evaluates empathy expressed in responses and not the empathy perceived by support seekers of the original seeker post (Discussion).

Hierarchical taxonomy of human–AI collaboration patterns

We wanted to understand how different participants collaborated with HAILEY. To derive collaboration patterns at the participant level, we aggregated and clustered post-level interactions for each participant over the ten posts in our study. First, we identified three dimensions of interest that were based on the design and features of HAILEY as well as by qualitatively analysing the interaction data: (1) the number of posts in the study for which the AI was consulted, (2) the way in which AI suggestions were used (direct versus indirect versus not at all) and (3) whether the participant looked for additional suggestions for a single post (using the reload functionality).

Direct use of AI was defined as directly accepting the AI's suggestions by clicking on the respective 'Insert' or 'Replace' button. Indirect use of AI, in contrast, was defined as making changes to the response by drawing ideas from the suggested edits. We operationalized indirect use as a cosine similarity of more than 95% between the BERT-based embeddings[72] of the final changes to the response by the participant and the edits suggested by the AI. Next, we used k-means to cluster the interaction data of all participants on the above dimensions ($k = 20$

based on the elbow method[82]). We manually analysed the distribution of the 20 inferred clusters, merged similar clusters, discarded the clusters that were noisy (for example, too small or having no consistent interaction behaviour) and organized the remaining 8 clusters in a top-down approach to derive the hierarchical taxonomy of human–AI collaboration patterns (Fig. 3a and Results). Finally, for the collaboration patterns with simple rule-based definitions (for example, participants who never consulted AI), we manually corrected the automatically inferred cluster boundaries to make the patterns more precise, for example, by keeping only the participants who had never consulted AI in that cluster.

We note that conditioning on the collaboration patterns, as in Fig. 3b,c, may introduce selection effects, as the type of collaboration was not randomly assigned. For example, participants who never used feedback suggestions may have been less engaged with the study and task overall.

Reporting summary

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

Data availability

Data used for training the empathy classification model used for automatic evaluation are available at https://github.com/behavioral-data/Empathy-Mental-Health[29,83]. Data used for training PARTNER and the data collected in our randomized controlled trial are available on request from the corresponding author with a clear justification and a license agreement from TalkLife.

Code availability

Source code of the empathy classification model used for automatic evaluation is available at https://github.com/behavioral-data/Empathy-Mental-Health[29,83]. Source code of PARTNER is available at https://github.com/behavioral-data/PARTNER[47,56]. Code used for designing the interface of HAILEY is available at https://github.com/behavioral-data/Human-AI-Collaboration-Empathy[84]. Code used for the analysis of the study data is available on request from the corresponding author. For the most recent project outcomes and resources, please visit https://bdata.uw.edu/empathy.

References

1. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
2. Hosny, A. & Aerts, H. J. Artificial intelligence for global health. *Science* **366**, 955–956 (2019).
3. Patel, B. N. et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digit. Med.* **2**, 111 (2019).
4. Tschandl, P. et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
5. Cai, C. J., Winter, S., Steiner, D., Wilcox, L. & Terry, M. 'Hello AI': uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making. *Proc. ACM Hum.-Comput. Interact.* **3**, 1–24 (2019).
6. Suh, M., Youngblom, E., Terry, M. & Cai, C. J. AI as social glue: uncovering the roles of deep generative AI during social music composition. In *CHI Conference on Human Factors in Computing Systems*, 1–11 (Association for Computing Machinery, 2021).
7. Wen, T.-H. et al. A network-based end-to-end trainable task-oriented dialogue system. In *European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 438–449 (Association for Computational Linguistics, 2017).
8. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
9. Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
10. Verghese, A., Shah, N. H. & Harrington, R. A. What this computer needs is a physician: humanism and artificial intelligence. *J. Am. Med. Assoc.* **319**, 19–20 (2018).
11. Bansal, G. et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *CHI Conference on Human Factors in Computing Systems*, 1–16 (Association for Computing Machinery, 2021).
12. Yang, Q., Steinfeld, A., Rosé, C. & Zimmerman, J. Re-examining whether, why, and how human–AI interaction is uniquely difficult to design. In *CHI Conference on Human Factors in Computing Systems*, 1–13 (Association for Computing Machinery, 2020).
13. Li, R. C., Asch, S. M. & Shah, N. H. Developing a delivery science for artificial intelligence in healthcare. *npj Digit. Med.* **3**, 107 (2020).
14. Gillies, M. et al. Human-centred machine learning. In *CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 3558–3565 (Association for Computing Machinery, 2016).
15. Amershi, S. et al. Guidelines for human–AI interaction. In *CHI Conference on Human Factors in Computing Systems*, 1–13 (Association for Computing Machinery, 2019).
16. Norman, D. A. How might people interact with agents. *Commun. ACM* **37**, 68–71 (1994).
17. Hirsch, T., Merced, K., Narayanan, S., Imel, Z. E. & Atkins, D. C. Designing contestability: interaction design, machine learning, and mental health. *Des Interact Syst Conf* **2017**, 95–99 (2017).
18. Clark, E., Ross, A. S., Tan, C., Ji, Y. & Smith, N. A. Creative writing with a machine in the loop: case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*, 329–340 (Association for Computing Machinery, 2018).
19. Roemmele, M. & Gordon, A. S. Automated assistance for creative writing with an RNN language model. In *23rd Intl Conference on Intelligent User Interfaces Companion*, 1–2 (Association for Computing Machinery, 2018).
20. Lee, M., Liang, P. & Yang, Q. Coauthor: designing a human–AI collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*, 1–19 (Association for Computing Machinery, 2022).
21. Paraphrasing tool. QuillBot https://quillbot.com/ (2022).
22. Buschek, D., Zürn, M. & Eiband, M. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native English writers. In *CHI Conference on Human Factors in Computing Systems*, 1–13 (Association for Computing Machinery, 2021).
23. Gero, K. I., Liu, V. & Chilton, L. B. Sparks: inspiration for science writing using language models. In Designing Interactive Systems Conference, 1002–1019 (2022).
24. Chilton, L. B., Petridis, S. & Agrawala, M. Visiblends: a flexible workflow for visual blends. In *CHI Conference on Human Factors in Computing Systems*, 1–14 (Association for Computing Machinery, 2019).
25. Elliott, R., Bohart, A. C., Watson, J. C. & Greenberg, L. S. Empathy. *Psychotherapy* **48**, 43–49 (2011).
26. Elliott, R., Bohart, A. C., Watson, J. C. & Murphy, D. Therapist empathy and client outcome: an updated meta-analysis. *Psychotherapy* **55**, 399–410 (2018).
27. Bohart, A. C., Elliott, R., Greenberg, L. S. & Watson, J. C. in *Psychotherapy Relationships That Work: Therapist Contributions and Responsiveness to Patients* (ed. Norcross, J. C.) **Vol. 452**, 89–108 (Oxford Univ. Press, 2002).
28. Watson, J. C., Goldman, R. N. & Warner, M. S. *Client-Centered and Experiential Psychotherapy in the 21st Century: Advances in Theory, Research, and Practice* (PCCS Books, 2002).

29. Sharma, A., Miner, A. S., Atkins, D. C. & Althoff, T. A computational approach to understanding empathy expressed in text-based mental health support. In *Conference on Empirical Methods in Natural Language Processing*, 5263–5276 (Association for Computational Linguistics, 2020).
30. Davis, M. H. A. et al. A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology* **10**, 85–103 (1980).
31. Blease, C., Locher, C., Leon-Carlyle, M. & Doraiswamy, M. Artificial intelligence and the future of psychiatry: qualitative findings from a global physician survey. *Digit. Health* **6**, 2055207620 968355 (2020).
32. Doraiswamy, P. M., Blease, C. & Bodner, K. Artificial intelligence and the future of psychiatry: Insights from a global physician survey. *Artif. Intell. Med.* **102**, 101753 (2020).
33. Riess, H. The science of empathy. *J. Patient Exp.* **4**, 74–77 (2017).
34. Mental disorders. *World Health Organization* https://www.who.int/news-room/fact-sheets/detail/mental-disorders (2022).
35. Kazdin, A. E. & Blase, S. L. Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspect. Psychol. Sci.* **6**, 21–37 (2011).
36. Olfson, M. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Aff.* **35**, 983–990 (2016).
37. Naslund, J. A., Aschbrenner, K. A., Marsch, L. A. & Bartels, S. J. The future of mental health care: peer-to-peer support and social media. *Epidemiol. Psychiatr. Sci.* **25**, 113–122 (2016).
38. Kemp, V. & Henderson, A. R. Challenges faced by mental health peer support workers: peer support from the peer supporter's point of view. *Psychiatr. Rehabil. J.* **35**, 337–340 (2012).
39. Mahlke, C. I., Krämer, U. M., Becker, T. & Bock, T. Peer support in mental health services. *Curr. Opin. Psychiatry* **27**, 276–281 (2014).
40. Schwalbe, C. S., Oh, H. Y. & Zweben, A. Sustaining motivational interviewing: a meta-analysis of training studies. *Addiction* **109**, 1287–1294 (2014).
41. Goldberg, S. B. et al. Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *J. Couns. Psychol.* **63**, 1–11 (2016).
42. Nunes, P., Williams, S., Sa, B. & Stevenson, K. A study of empathy decline in students from five health disciplines during their first year of training. *J. Int. Assoc. Med. Sci. Educ.* **2**, 12–17 (2011).
43. Hojat, M. et al. The devil is in the third year: a longitudinal study of erosion of empathy in medical school. *Acad. Med.* **84**, 1182–1191 (2009).
44. Stebnicki, M. A. Empathy fatigue: healing the mind, body, and spirit of professional counselors. *Am. J. Psychiatr. Rehabil.* **10**, 317–338 (2007).
45. Imel, Z. E., Steyvers, M. & Atkins, D. C. Computational psychotherapy research: scaling up the evaluation of patient–provider interactions. *Psychotherapy* **52**, 19–30 (2015).
46. Miner, A. S. et al. Key considerations for incorporating conversational AI in psychotherapy. *Front. Psychiatry* **10**, 746 (2019).
47. 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. of the Web Conference,* 194–205 (Association for Computing Machinery, 2021).
48. Lin, Z., Madotto, A., Shin, J., Xu, P. & Fung, P. MoEL: mixture of empathetic listeners. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing,* 121–132 (Association for Computational Linguistics, 2019).
49. Majumder, N. et al. Mime: mimicking emotions for empathetic response generation. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 8968–8979 (Association for Computational Linguistics, 2020).
50. Rashkin, H., Smith, E. M., Li, M. & Boureau, Y.-L. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *Annual Meeting of the Association for Computational Linguistics*, 5370–5381 (Association for Computational Linguistics, 2019).
51. Chen, J. H. & Asch, S. M. Machine learning and prediction in medicine – beyond the peak of inflated expectations. *N. Engl. J. Med.* **376**, 2507–2509 (2017).
52. Tanana, M. J., Soma, C. S., Srikumar, V. et al. Development and evaluation of ClientBot: patient-like conversational agent to train basic counseling skills. *J. Med. Internet Res.* **21**, e12529 (2019).
53. Peng, Z., Guo, Q., Tsang, K. W. & Ma, X. Exploring the effects of technological writing assistance for support providers in online mental health community. In *CHI Conference on Human Factors in Computing Systems*, 1–15 (Association for Computing Machinery, 2020).
54. Hui, J. S., Gergle, D. & Gerber, E. M. IntroAssist: a tool to support writing introductory help requests. In *CHI Conference on Human Factors in Computing Systems*, 1–13 (Association for Computing Machinery, 2018).
55. Kelly, R., Gooch, D. & Watts, L. 'It's more like a letter': an exploration of mediated conversational effort in message builder. *Proc. ACM Hum. Comput. Interact.* **2**, 1–18 (2018).
56. Sharma, A. behavioral-data/partner: code for the WWW 2021 paper on empathic rewriting. *Zenodo* https://doi.org/10.5281/ZENODO.7053967 (2022).
57. Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. & Shah, N. H. Minimar (minimum information for medical AI reporting): developing reporting standards for artificial intelligence in health care. J. Am. *Med. Inf. Assoc.* **27**, 2011–2015 (2020).
58. Barrett-Lennard, G. T. The empathy cycle: refinement of a nuclear concept. *J. Couns. Psychol.* **28**, 91–100 (1981).
59. Collins, P. Y. Grand challenges in global mental health. *Nature* **475**, 27–30 (2011).
60. Kaplan, B. H., Cassel, J. C. & Gore, S. Social support and health. *Med. Care* **15**, 47–58 (1977).
61. Rathod, S. et al. Mental health service provision in low- and middle-Income countries. *Health Serv. Insights* **10**, 1178632917694350 (2017).
62. Lee, E. E. et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **6**, 856–864 (2021).
63. Vaidyam, A. N., Linggonegoro, D. & Torous, J. Changes to the psychiatric chatbot landscape: a systematic review of conversational agents in serious mental illness. *Can. J. Psychiatry* **66**, 339–348 (2021).
64. Richardson, J. P. et al. Patient apprehensions about the use of artificial intelligence in healthcare. *npj Digit. Med* **4**, 140 (2021).
65. Collings, S. & Niederkrotenthaler, T. Suicide prevention and emergent media: surfing the opportunity. *Crisis* **33**, 1–4 (2012).
66. Luxton, D. D., June, J. D. & Fairall, J. M. Social media and suicide: a public health perspective. *Am. J. Public Health* **102**, S195–200 (2012).
67. Martinez-Martin, N. & Kreitmair, K. Ethical issues for direct-to-consumer digital psychotherapy apps: addressing accountability, data protection, and consent. *JMIR Ment. Health* **5**, e32 (2018).
68. Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C. & Srikumar, V. A comparison of natural language processing methods for automated coding of motivational interviewing. *J. Subst. Abuse Treat.* **65**, 43–50 (2016).

69. De Choudhury, M., Sharma, S. S., Logar, T. et al. Gender and cross-cultural differences in social media disclosures of mental illness. In *ACM Conference on Computer Supported Cooperative Work and Social Computing*, 353–369 (Association for Computing Machinery, 2017).
70. Cauce, A. M. et al. Cultural and contextual influences in mental health help seeking: a focus on ethnic minority youth. *J. Consult. Clin. Psychol.* **70**, 44–55 (2002).
71. Satcher, D. *Mental Health: Culture, Race, and Ethnicity—A Supplement to Mental Health: a Report of the Surgeon General* (U.S. Department of Health and Human Services, 2001).
72. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **Vol. 1**, 4171–4186 (Association for Computational Linguistics, 2019).
73. Li, J., Galley, M., Brockett, C., Gao, J. & Dolan, W. B. A diversity-promoting objective function for neural conversationmodels. In *NAACL-HLT* (2016).
74. Wolf, M. J., Miller, K. & Grodzinsky, F. S. Why we should have seen that coming: comments on microsoft's tay 'experiment,' and wider implications. *ACM SIGCAS Comput. Soc.* **47**, 54–64 (2017).
75. Bolukbasi, T., Chang, K.-W., Zou, J. Y. et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 29 (2016).
76. Daws, R. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. *AI News* https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/ (2020).
77. Radford, A. et al. Language models are unsupervised multitask learners. *CloudFront* https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (2022).
78. Lee, F.-T., Hull, D., Levine, J. et al. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proc. of the 6th Workshop on Computational Linguistics and Clinical Psychology*, 12–23 (Association for Computational Linguistics, 2019).
79. Zheng, C., Liu, Y., Chen, W. et al. CoMAE: a multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics*, 813–824 (Association for Computational Linguistics, 2021).
80. Wambsganss, T., Niklaus, C., Söllner, M. et al. Supporting cognitive and emotional empathic writing of students. In *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 4063–4077 (Association for Computational Linguistics, 2021).
81. Majumder, N. et al. Exemplars-guided empathetic response generation controlled by the elements of human communication. *IEEE Access* **10**, 77176–77190 (2022).
82. Elbow method (clustering). *Wikipedia* https://en.wikipedia.org/wiki/Elbow_method_(clustering) (2022).
83. Sharma, A. Behavioral-data/empathy-mental-health: code for the EMNLP 2020 paper on empathy. *Zenodo* https://doi.org/10.5281/ZENODO.7061732 (2022).
84. Sharma, A. Behavioral-data/human–AI-collaboration-empathy: code for HAILEY. *Zenodo* https://doi.org/10.5281/ZENODO.7295902 (2022).

Acknowledgements

We thank TalkLife and J. Druitt for supporting this work, for advertising the study on their platform and for providing us access to a TalkLife data set. We also thank members of the UW Behavioral Data Science Group, Microsoft AI for Accessibility team and D.S. Weld for their suggestions and feedback. T.A., A.S. and I.W.L. were supported in part by NSF grant IIS-1901386, NSF CAREER IIS-2142794, NSF grant CNS-2025022, NIH grant R01MH125179, Bill & Melinda Gates Foundation (INV-004841), the Office of Naval Research (#N00014-21-1-2154), a Microsoft AI for Accessibility grant and a Garvey Institute Innovation grant. A.S.M. was supported by grants from the National Institutes of Health, National Center for Advancing Translational Science, Clinical and Translational Science Award (KL2TR001083 and UL1TR001085) and the Stanford Human-Centered AI Institute. D.C.A. was supported by NIH career development award K02 AA023814.

Author contributions

A.S., I.W.L., A.S.M., D.C.A. and T.A. were involved with the design of HAILEY and the formulation of the study. A.S. and I.W.L. conducted the study. All authors interpreted the data, drafted the manuscript and made significant intellectual contributions to the manuscript.

Competing interests

D.C.A. is a co-founder with equity stake in a technology company, Lyssn.io, focused on tools to support training, supervision and quality assurance of psychotherapy and counselling. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00593-2.

Correspondence and requests for materials should be addressed to Tim Althoff.

Peer review information *Nature Machine Intelligence* thanks Ryan Kelly and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature research

Corresponding author(s): Tim Althoff

Last updated by author(s): Apr 28, 2022

Reporting Summary

Nature Research 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 Research 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	TalkLife data was sourced with a license agreement and consent from the TalkLife platform through Microsoft SQL Server 2019. Study data was collected through a randomized controlled trial conducted on an independent platform with participants from TalkLife through HTML 5, Python 3.7.3 and Django 3.2. The code used for designing this platform is available at https://github.com/behavioral-data/Human-AI-Collaboration-Empathy (DOI: 10.5281/ZENODO.7295902).
Data analysis	Data analysis was conducted in Python 3.7.3 using standard data analysis libraries such as numpy 1.21.6, pandas 1.3.5, and seaborn 0.11.2. Code used for this analysis is available on request from the corresponding author

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 Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Data used for training the empathy classification model used for automatic evaluation is available at https://github.com/behavioral-data/Empathy-Mental-Health. Due to a license agreement with TalkLife, the data used for training Partner and the data collected in our randomized controlled trial cannot be shared publicly. They are only available on request from the corresponding author with a clear justification and a license agreement from TalkLife. For more info, the TalkLife team can be contacted at research@talklife.co

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 employed a quantitative, between-subjects study design in which each participant was randomly assigned to one of Human + AI (treatment; N=139) or Human Only (control; N=161) conditions. Participants in both groups were asked to write supportive, empathic responses to a unique set of 10 existing seeker posts (one at a time), sourced at random from a subset of TalkLife posts.The Human + AI (treatment) group participants were given the option of receiving feedback through prompts as they typed their responses. Participants in the Human Only (control) group, in contrast, wrote responses with no option for feedback.
Research sample	We obtained a unique set of 1500 seeker posts, sampled at random with consent from the TalkLife platform. This sample was broadly reflective of the platform's user population in terms of age and self-reported gender (Mean Age = 26.7; %Female = 56.8%; %Male= 24.8%; %Other=18.4%).
Sampling strategy	We used a power analysis to estimate the number of participants required for our study. For an effect size of 0.1 difference in empathy, a power analysis with a significance level of 0.05, powered at 80%, indicated that we required 1,500 samples of (seeker post, response post) pairs each for treatment and control groups.
Data collection	We worked with a large peer-to-peer support platform, TalkLife. To recruit participants, we advertised our study on TalkLife. On clicking the advertised pop-up used for recruitment, a TalkLife user was randomly assigned to one of the Human + AI (treatment) or Human Only (control) conditions for the study duration. The participant was then directed to an independent platform designed by the researchers where they performed the study. Participants did not interact with the researchers during the study.
Timing	April 12, 2021 to September 13, 2021
Data exclusions	We excluded participants who belonged to multiple Human-AI collaboration clusters in our collaboration patterns analysis. No data was excluded in the other analyses.
Non-participation	A total of 5987 users opened our advertisement on the TalkLife platform. Out of these, 652 users proceeded to the Phase I of pre-intervention survey with 352 users eventually dropping out.
Randomization	Each participant was randomly assigned to one of Human + AI (treatment) or Human Only (control) conditions.

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
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	In our study, 54.33% of the participants identified as female, 36.67% as male, 7.33% as non-binary, and the remaining 1.67% preferred not to report their gender. The average age of participants was 26.34 years (std = 9.50). 45.67% of the participants identified as White, 20.33% as Asians, 10.67% as Hispanic or Latino, 10.33% as Black or African American, 0.67% as Pacific Islander or Hawaiian, 0.33% as American Indian or Alaska Native, and the remaining 12.00% preferred not to report their

	race/ethnicity. 62.33% of the participants were from the United States, 13.67% were from India, 2.33% were from United Kingdom, 2.33% were from Germany, and the remaining 19.33% were from 36 different countries (spanning six of seven continents excluding Antarctica). Moreover, 31.67% of the participants reported having no experience with peer-to-peer support despite having been recruited from the TalkLife platform, 26.33% as having less than one year of experience, and 42.00% reported having greater than or equal to one year of experience with peer-to-peer support.
Recruitment	We worked with a large peer-to-peer support platform, TalkLife. To recruit participants, we advertised our study on their platform. This advertisement was shown to TalkLife users after they submitted a response on the platform, with an aim of targeting active peer supporters. A potential self-selection bias may have attracted TalkLife users who care more about supporting others. We found that this self-selection bias potentially results in conservative estimates of the effects of our feedback intervention.
Ethics oversight	University of Washington Institutional Review Board

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