Human-AI Collaboration Enables More Empathic Conversations in Text-based Peer-to-Peer Mental Health Support

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

[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
[*]althoff@cs.washington.edu

Abstract

Advances in artificial intelligence (AI) are enabling systems that augment and collaborate with humans to perform simple, mechanistic tasks like 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 difficulties of AI systems 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 analyze 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.

Introduction

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 healthcare[1–9]. In many and especially in high-stakes settings, such Human-AI collaboration has proven more robust and effective than totally replacing humans with AI[10,11]. However, the collaboration faces dual challenges of developing human-centered AI models to assist humans and designing human-facing interfaces for humans to interact with the AI[12–17]. 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 just-in-time[17–23]. Though initial systems have been proposed for tasks like 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, like

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 contributes 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 like 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 like 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 like empathic writing.

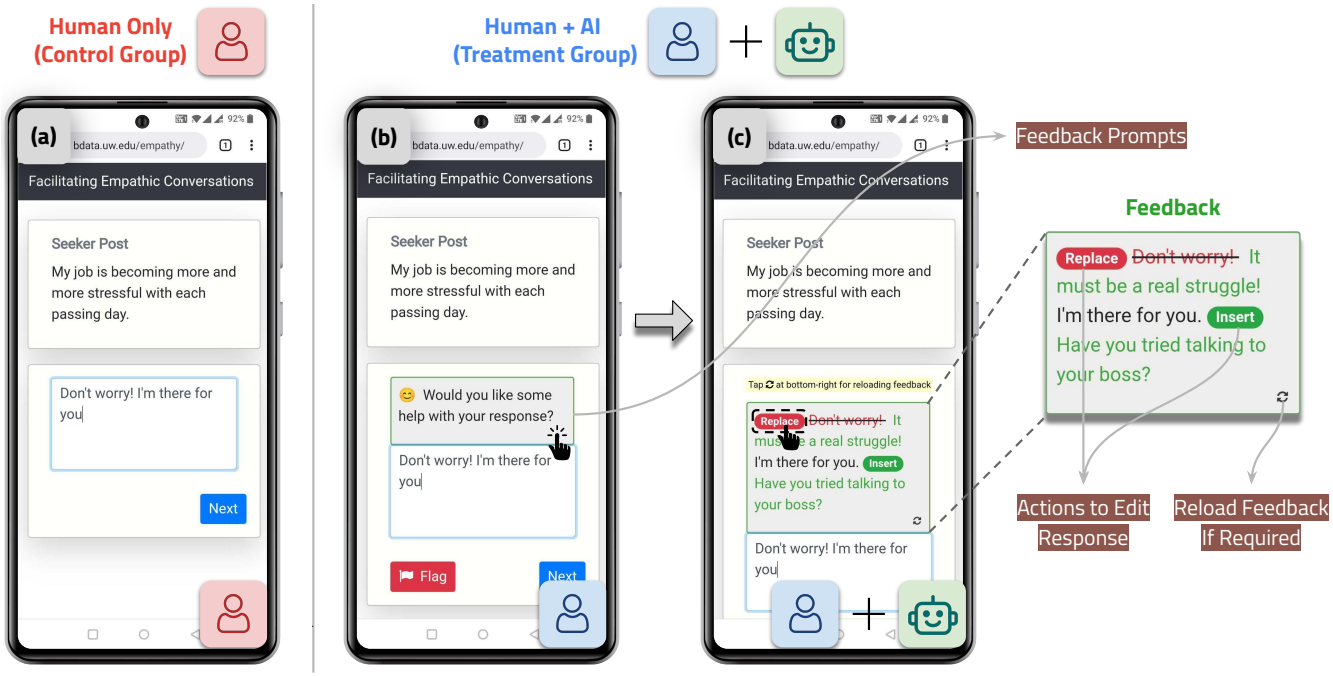
As shown in prior work[38,39], untrained peer supporters report difficulties in writing supportive, empathic responses to support seekers. Without deliberate training or specific feedback, the difficulty persists over time[29,40,41] and may even lead to a gradual decrease in supporters' effectiveness due to factors such as empathy fatigue[42–44]. Furthermore, current efforts to improve empathy (e.g., 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, such as 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 HAILEY (Human-AI coLlaboration approach for EmpathY), an AI-in-the-loop agent that offers just-in-time suggestions to express empathy more effectively in conversations (Figure 1b, 1c). 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 likely 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 (e.g., 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 their more empathic counterparts (Figure 1c). For complex, hard-to-learn skills like 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 (Figure 1). In these conversations, the support seeker authors a post seeking mental health support (e.g., "*My job is becoming more and more stressful with each passing day.*") to which the peer supporter writes a supportive response (e.g., "*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

Figure 1. We performed 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; Supplementary Figure 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 their 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.



AI feedback consisting of *Insert* (e.g., Insert "*Have you tried talking to your boss?*" at the end of the response) and *Replace* (e.g., 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 buttons, by further editing them, and/or by deriving ideas from the suggestions to indirectly use in their response. These suggestions are generated using PARTNER, a deep reinforcement learning model that learns to take sentence-level edits as actions in order increase 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), 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 (Figure 1; 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 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 Figure 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 10 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; Discussion). While writing responses, participants in the Human + AI (treatment) group received feedback via HAILEY (Figure 1b, 1c). 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 (Figure 1a). After completing responses to the 10 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, i.e., responses in the Human + AI (treatment) group would show higher empathy than the Human Only (control) group responses. 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[57]. Psychotherapy research indicates a strong correlation between expressed empathy and positive therapeutic outcomes and commonly uses it as a credible alternative[25] (Methods; 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 (Figure 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 ($p < 0.01$; $t = 5.88$; $D_f = 2998$; Two-sided Student's t-test; Figure 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 to the Human Only approach (1.77 vs. 1.48; Cohen's $d = 0.24$; $p < 10^{-5}$; $t = 5.46$; $D_f = 2998$; Two-sided Student's t-test; Figure 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 (e.g., 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-minute study (Supplementary Figure S6). Therefore, it is particularly important to better assist the many participants who struggle with writing responses.

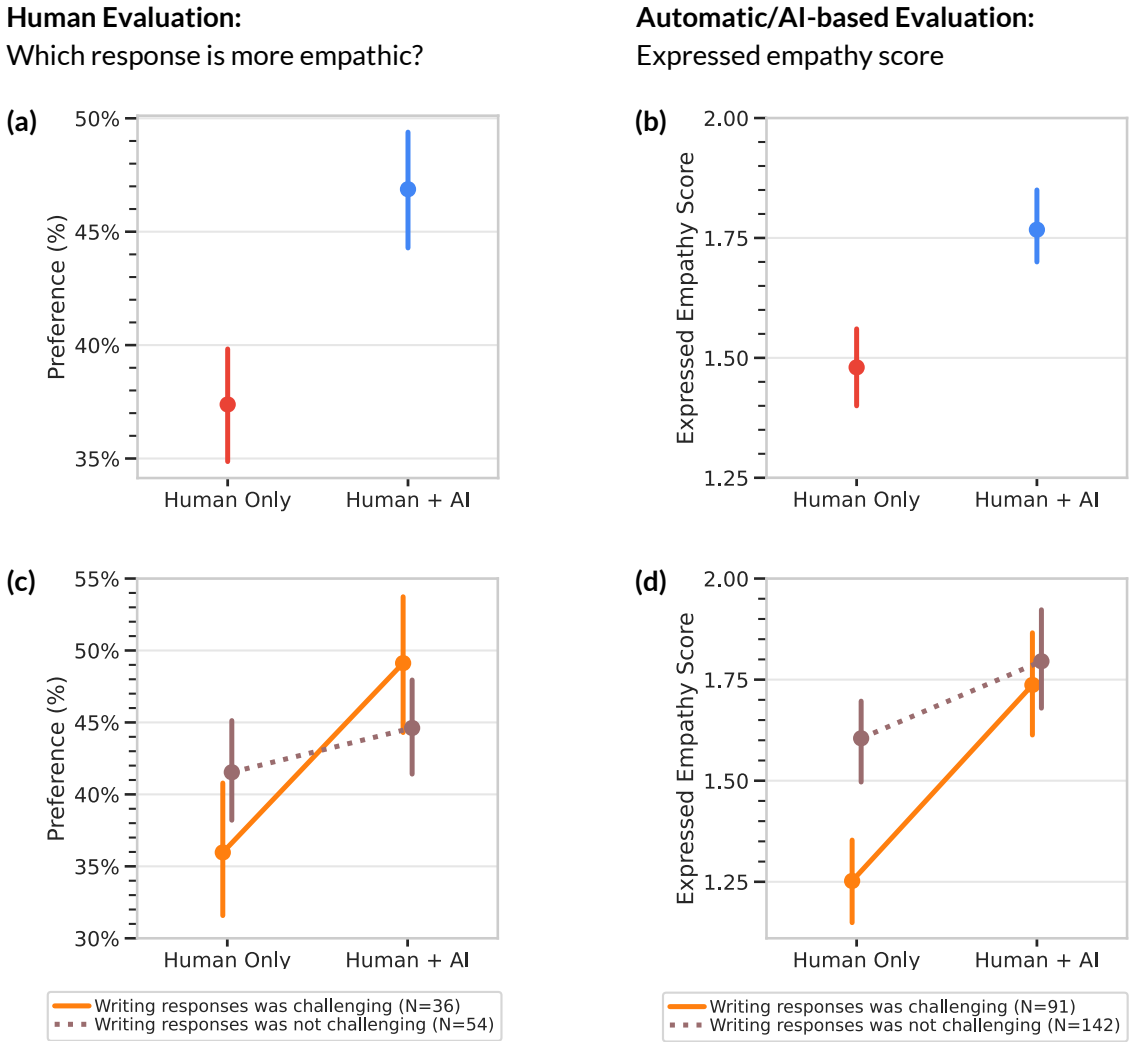
For the subsample of participants who self-reported challenges in writing responses at the end of our study (N=91/300; Methods), a post hoc analysis revealed significantly higher empathy gains using the Human-AI collaboration approach. For such participants, we found an absolute 4.5% stronger preference for the Human + AI responses (49.1% vs. 44.6%; $p < 0.01$; $t = 4.05$; $D_f = 718$; Two-sided Student's t-test; Figure 2c) and a 27.0% higher increase in expressed empathy using the Human + AI approach (38.8% vs. 11.8%; $p < 10^{-5}$; $t = 5.90$; $D_f = 1818$; Two-sided Student's t-test; Figure 2d) compared to 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% vs. 43.7%; $p < 0.01$; $t = 4.42$; $D_f = 758$; Two-sided Student's t-test;) and a 21.2% higher increase in expressed empathy using the Human + AI approach (33.7% vs. 12.5%; $p < 10^{-5}$; $t = 4.58$; $D_f = 1898$; Two-sided Student's t-test; Supplementary Figure S11d) compared to participants who reported experience with online peer support.

Key Human-AI Collaboration Patterns

The 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 inform better design decisions. Here, we analyzed 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 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 (Figure 3a; 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 10 posts

Figure 2. **(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 vs. 1.48; Cohen's d=0.24; p=$5.1 * e^{-8}$; t = 5.46; $D_f = 2998$; 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 vs. Human Only responses (49.1% vs. 34.0%), compared to participants who did not report challenges (44.6% vs. 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 vs. Human Only responses (38.9%; 1.74 vs. 1.25; Cohen's d=0.43), compared to participants who did not report challenges (11.9%; 1.79 vs. 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 and the error bars represent bootstrapped 95% confidence intervals.



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; Figure 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 in 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.

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 (Figure 3c) than through human evaluation (Figure 3b).

Positive Perceptions Of Participants

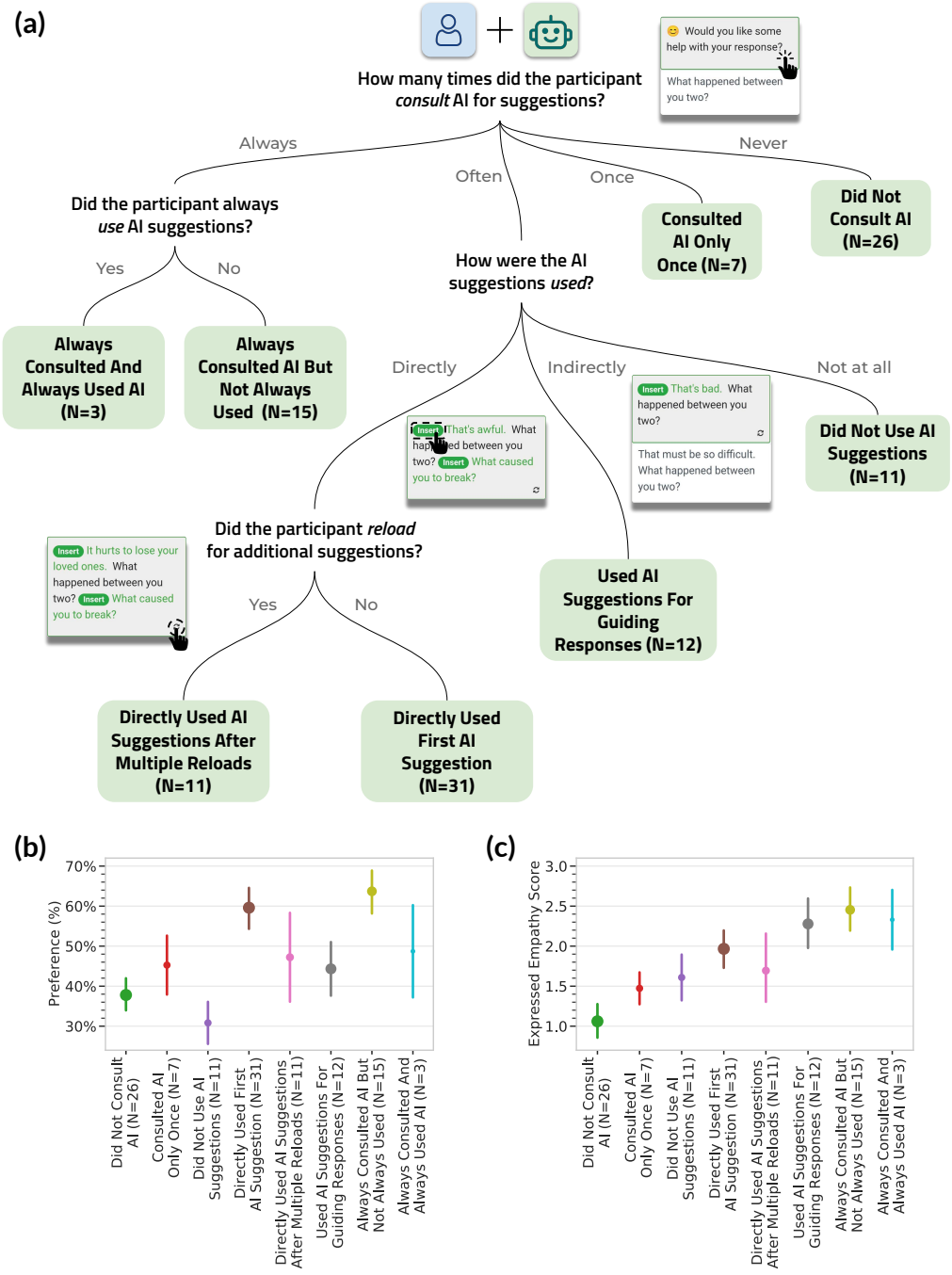
At the end of our study, we collected study participants' perceptions about the usefulness and actionability of the feedback and their intention to adopt the system. We observed that 63.3% of participants found the feedback they received helpful, 60.4% found it actionable, and 77.7% of participants wanted this type of feedback system to be deployed on TalkLife or other similar peer-to-peer support platforms (Supplementary Figure 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 Figure 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 WHO 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 behavioral health problems contribute 13% to the global burden of disease, more than both cardiovascular diseases and cancer[58]. Although psychotherapy and social support[59] can be effective treatments, many vulnerable individuals have limited access to therapy and counseling[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[60].

Figure 3. We derived a 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; $p = 6.4 * e^{-6}$; 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 and the error bars represent bootstrapped 95% confidence intervals.



A scalable approach to improving access to mental health support globally is by connecting support seekers and peer supporters using online platforms like TalkLife (talklife.com), YourDost (yourdost.com) or Mental Health Subreddits (reddit.com)[37] to those with mental health issues. 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 skepticism 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[35,45,46,61,62].

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 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 (e.g., 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 (e.g., 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 (e.g., 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 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 (e.g., 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 (e.g., 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[31,32,63]. 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 healthcare.

The reported differences in empathy between 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 Figure S1), and (2) a potential selection effect that may have attracted TalkLife users who care more about supporting others (Supplementary Figure 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 vs. 1.48; $p \ll 10^{-5}$; Two-sided Student's t-test; Supplementary Figure 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[13,64–66]. There is a risk that in attempting to help, AI may have the opposite effect on the potentially vulnerable support seeker or peer supporter[63]. 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 loss of authenticity in responses; hence, our Human-AI collaboration approach leads to responses with high empathy as well as high authenticity (Supplementary Figure 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 (e.g., ".*(commit suicide).*", ".*(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 Figure S31). In our study, 1.6% posts (out of 1390 in the treatment group) and 2.9% feedback instances (out of 1939 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 (e.g., "*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 (e.g., "*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 (e.g., 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 dataset 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 (e.g., 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 (e.g., concrete problem solving, motivational interviewing[67]). 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 analyzing 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 (e.g., 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[57]. In sensitive contexts like ours, however, obtaining perceived empathy ratings from support seekers is challenging and involves ethical risks (Safety, Privacy, and Ethics). 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[68–70]. 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; Supplementary Figure S10, 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 analyzed a classification-based interface design, which provided participants with the option to request automatic expressed empathy scores[29] for their responses (Supplementary Figure 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 empathy of participants were ambiguous (Supplementary Figure S4a, S4b); however, the design was perceived as being less actionable than our primary rewriting-based interface (Supplementary Figure S4c). This poses questions on 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 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.

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 (e.g., *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 like 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 treatment and control groups. To meet the required sample size, we collected 10 samples per participant and therefore recruited from 300 participants in total (with the goal of 150 participants per condition), for a total of 1,500 samples each.

Dataset of Seeker Posts. We obtained a unique set of 1500 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 (e.g., ".*(commit suicide).*", ".*(cut).*"), to ensure participant safety (Discussion), and (2) common social media interactions not related to mental health (e.g., "*Happy mother's day*") using a standard BERT-based text classifier[71], trained on a manually annotated dataset 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 1500 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, 7.3% as non-binary, and the remaining 1.7% preferred not to report their gender. The average age of participants was 26.3 years (std = 9.5). 45.7% of the participants identified as White, 20.3% as Asians, 10.7% as Hispanic or Latino, 10.3% as Black or African American, 0.7% as Pacific Islander or Hawaiian, 0.3% as American Indian or Alaska Native, and the remaining 12.0% preferred not to report their race/ethnicity. 62.3% of the participants were from the United States, 13.7% were from India, 2.3% were from United Kingdom, 2.3% were from Germany, and 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 one year of experience, and 42.0% reported having greater than or equal to one 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 Figure S19, 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 Figure S1). Participants were also shown instructions on using our study interface in this phase (Supplementary Figure S21, S22, S23, S24, S25, S26, S27, 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 Figure S30). Participants in the Human Only (control) group wrote responses without any option for feedback (Supplementary Figure S29).
- **Phase IV: Post-Intervention Survey.** After completing the 10 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 the intent to adopt the system (Supplementary Figure S32, S33, S34).

If participants dropped out of the study before completing it, their data was removed from our analyses. Participants took 20.6 minutes on average to complete the study. US citizens and permanent US residents were compensated with a 5 USD Amazon gift card. Furthermore, the top-2 participants in the human evaluation (Evaluation) received an additional 25 USD Amazon gift card. Based on local regulations, we were unable to pay non-US participants. This was explicitly highlighted in the participant consent form on the first landing page of our study (Supplementary Figure S18, S35).

Design Goals

HAILEY is designed (1) with a collaborative "AI-in-the-loop" approach, (2) to provide actionable feedback, and (3) 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[63,72–74]. To mitigate these risks, researchers have called for Human-AI collaboration methods, where 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 center of the interaction, receive suggestions from our AI "in-the-loop," and retain full control over which suggestions to use in their responses (e.g., 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 (e.g., 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 to take for participants to improve. HAILEY, designed to be actionable, suggests concrete actions (e.g., 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 friendliness design is critical for the adoption of conversational assistive agents like 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 to 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 Figure 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 (Figure 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 Sharma et al.[47], is to transform low empathy text to higher empathy. The authors proposed PARTNER, a deep reinforcement learning model that learns to take sentence-level edits as actions in order 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[75]), 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 Sharma et al.[56].

Showing Feedback as 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 Figure 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 Figure 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 Figure 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 to those written without feedback given the same seeker post (Supplementary Figure S35, S36, S37). We found that the participant annotations have 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[29,76]. When analyzing strata of participants based on challenges in writing responses (Figure 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 challenging or not challenging categories.

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[47,77–79]. 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 10 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 analyzing 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 vs. indirect vs. 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 buttons. 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[71] 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[80]). We manually analyzed the distribution of the 20 inferred clusters, merged similar clusters, discarded the clusters that were noisy (e.g., too small or having no consistent interaction behavior), and organized the remaining 8 clusters in a top-down approach to derive the hierarchical taxonomy of Human-AI collaboration patterns (Figure 3a; Results). Finally, for the collaboration patterns with simple rule-based definitions (e.g., participants who never consulted AI), we manually corrected the automatically inferred cluster boundaries to make the patterns more precise, e.g., by keeping only the participants who had never consulted AI in that cluster.

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

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[29,81]. 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[29,81]. 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[82]. 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.

Acknowledgements

We would like to thank TalkLife and Jamie Druitt for supporting this work, for advertising the study on their platform, and for providing us access to a TalkLife dataset. We also thank members of the UW Behavioral Data Science Group, Microsoft AI for Accessibility team, and Daniel 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.

Author Contributions

A.S., I.W.L., A.S.M., D.C.A., and T.A. were involved with the design of HAILEYand 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 Statement

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 counseling. The remaining authors declare no competing interests.

Figure Captions

Figure 1. We performed 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; Supplementary Figure 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 their 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.

Figure 2. **(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 vs. 1.48; Cohen's d=0.24; p=$5.1 * 10^{-8}$; t = 5.46; $D_f = 2998$; 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 vs. Human Only responses (49.1% vs. 34.0%), compared to participants who did not report challenges (44.6% vs. 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 vs. Human Only responses (38.9%; 1.74 vs. 1.25; Cohen's d=0.43), compared to participants who did not report challenges (11.9%; 1.79 vs. 1.60; Cohen's d=0.15). In **c** and **d**, the sample size varied to ensure comparable conditions (Methods). Data are presented as mean values $\pm$ 1.96$*$SEM (bootstrapped 95% confidence intervals).

Figure 3. We derived a 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; $p = 6.4 * 10^{-6}$; 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. Data are presented as mean values $\pm$ 1.96$*$SEM (bootstrapped 95% confidence intervals).

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** (2019).
6. Suh, M. 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* (2021).
7. Wen, T.-H. *et al.* A network-based End-to-End trainable task-oriented dialogue system. In *EACL* (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. *JAMA* **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* (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* (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* (2016).

15. Amershi, S. *et al.* Guidelines for human-ai interaction. In *CHI* (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. *DIS (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 *IUI* (2018).

19. Roemmele, M. & Gordon, A. S. Automated assistance for creative writing with an RNN language model. In *IUI Companion* (2018).

20. Lee, M., Liang, P. & Yang, Q. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI* (2022).

21. Paraphrasing tool. https://quillbot.com/. Accessed: 2022-1-25.

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* (2021).

23. Gero, K. I., Liu, V. & Chilton, L. B. Sparks: Inspiration for science writing using language models. *arXiv preprint arXiv:2110.07640* (2021).

24. Chilton, L. B., Petridis, S. & Agrawala, M. Visiblends: A flexible workflow for visual blends. In *CHI* (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. Empathy. In Norcross, J. C. (ed.) *Psychotherapy relationships that work: Therapist contributions and responsiveness to patients , (pp*, vol. 452, 89–108 (Oxford University Press, xii, New York, NY, US, 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 *EMNLP* (2020).

30. Davis, M. H. A. *et al.* A multidimensional approach to individual differences in empathy. *J. Pers. Soc. Psychol.* (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. Heal.* **6**, 2055207620968355 (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. https://www.who.int/news-room/fact-sheets/detail/mental-disorders. Accessed: 2022-1-25.

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. *Heal. 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. rehabilitation journal* **35**, 337 (2012).

39. Mahlke, C. I., Krämer, U. M., Becker, T. & Bock, T. Peer support in mental health services. *Curr. opinion 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 *WWW/TheWebConf* (2021).

48. Lin, Z., Madotto, A., Shin, J., Xu, P. & Fung, P. MoEL: Mixture of empathetic listeners. In *EMNLP* (2019).

49. Majumder, N. *et al.* Mime: Mimicking emotions for empathetic response generation. In *EMNLP* (2020).

50. Rashkin, H., Smith, E. M., Li, M. & Boureau, Y.-L. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL* (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., Atkins, D. C. & Imel, Z. E. 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* (2020).

54. Hui, J. S., Gergle, D. & Gerber, E. M. Introassist: A tool to support writing introductory help requests. In *CHI* (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** (2018).

56. Sharma, A. behavioral-data/partner: Code for the www 2021 paper on empathic rewriting, DOI: 10.5281/ZENODO.7053967 (2022).

57. Barrett-Lennard, G. T. The empathy cycle: Refinement of a nuclear concept. *J. Couns. Psychol.* **28**, 91–100 (1981).

58. Collins, P. Y. *et al.* Grand challenges in global mental health. *Nature* **475**, 27–30 (2011).

59. Kaplan, B. H., Cassel, J. C. & Gore, S. Social support and health. *Med. Care* **15**, 47–58 (1977).

60. Rathod, S. *et al.* Mental health service provision in low- and Middle-Income countries. *Heal. Serv Insights* **10**, 1178632917694350 (2017).

61. 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).

62. Vaidyam, A. N., Linggonegoro, D. & Torous, J. Changes to the psychiatric chatbot landscape: A systematic review of conversational agents in serious mental illness: Changements du paysage psychiatrique des chatbots: une revue systématique des agents conversationnels dans la maladie mentale sérieuse. *Can. J. Psychiatry* **66**, 339–348 (2021).

63. Richardson, J. P. *et al.* Patient apprehensions about the use of artificial intelligence in healthcare. *NPJ Digit. Med* **4**, 140 (2021).

64. Collings, S. & Niederkrotenthaler, T. Suicide prevention and emergent media: surfing the opportunity. *Crisis* **33**, 1–4 (2012).

65. Luxton, D. D., June, J. D. & Fairall, J. M. Social media and suicide: a public health perspective. *Am. J. Public Heal.* **102 Suppl 2**, S195–200 (2012).

66. Martinez-Martin, N. & Kreitmair, K. Ethical issues for Direct-to-Consumer digital psychotherapy apps: Addressing accountability, data protection, and consent. *JMIR Ment Heal.* **5**, e32 (2018).

67. 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. Abus. Treat.* **65**, 43–50 (2016).

68. De Choudhury, M., Sharma, S. S., Logar, T., Eekhout, W. & Nielsen, R. C. Gender and cross-cultural differences in social media disclosures of mental illness. In *CSCW* (2017).

69. Cauce, A. M. *et al.* Cultural and contextual influences in mental health help seeking: a focus on ethnic minority youth. *J. consulting clinical psychology* **70**, 44 (2002).

70. Satcher, D. *Mental health: Culture, race, and ethnicity—A supplement to mental health: A report of the surgeon general* (US Department of Health and Human Services, 2001).

71. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT* (2019).

72. 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).

73. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv. neural information processing systems* **29** (2016).

74. Daws, R. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/ (2020). Accessed: 2022-1-25.

75. Radford, A. *et al.* Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2022-1-25.

76. Lee, F.-T., Hull, D., Levine, J., Ray, B. & McKeown, K. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 12–23 (2019).

77. Zheng, C., Liu, Y., Chen, W., Leng, Y. & Huang, M. Comae: A multi-factor hierarchical framework for empathetic response generation. In *ACL Findings* (2021).

78. Wambsganss, T., Niklaus, C., Söllner, M., Handschuh, S. & Leimeister, J. M. Supporting cognitive and emotional empathic writing of students. In *ACL-IJCNLP* (2021).

79. Majumder, N. *et al.* Exemplars-guided empathetic response generation controlled by the elements of human communication. *arXiv preprint arXiv:2106.11791* (2021).

80. Wikipedia contributors. Elbow method (clustering). https://en.wikipedia.org/wiki/Elbow_method_(clustering) (2022). Accessed: 2022-2-28.

81. Sharma, A. behavioral-data/empathy-mental-health: Code for the emnlp 2020 paper on empathy, DOI: 10.5281/ZENODO.7061732 (2022).

82. Sharma, A. behavioral-data/human-ai-collaboration-empathy: Code for hailey, DOI: 10.5281/ZENODO.7295902 (2022).

83. 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. Informatics Assoc.* **27**, 2011–2015 (2020).

84. Li, J., Galley, M., Brockett, C., Gao, J. & Dolan, W. B. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT* (2016).

Supplementary Materials

List of supplementary materials

Table S1
Figures S1 to S39

Table S1. Description of our randomized controlled trial (RCT) study population, setting and model, following reporting standards for artificial intelligence in health care from Hernandez-Boussard et al.[83]

Feature	Description
Study population and setting	
- Population:	300 TalkLife users; 161 in Human Only (control); 139 in Human + AI (treatment).
- Study setting:	Non-clinical, online platform outside of TalkLife to ensure platform users' safety, through an interface similar to TalkLife's chat feature.
- Data collected in RCT:	Participants responded to 10*300=3000 seeker posts (1500 unique seeker posts duplicated across control and treatment), generating 3000 responses (1610 in control, 1390 in treatment). An independent set of 50 participants rated 1390 pairs of control and treatment responses on empathy preference.
- Cohort selection:	Participants were sent a recruitment request after they submitted a response on the TalkLife platform, with an aim of targeting active peer supporters. Participants were excluded if they dropped out of the study before completion.
- Registration:	We did not pre-register on ClinicalTrials.gov because our study was conducted in a non-clinical setting.
Participant demographic characteristics	
- Age:	Mean=26.3 years; Std=9.5 years
- Gender:	Female: 54.3%; Male: 36.7%; Non-binary: 7.3%; Preferred not to say: 1.7%
- Race/Ethnicity:	White: 45.7%; Asian: 20.3%; Hispanic or Latino: 10.7%; Black or African American: 10.3%; Pacific Islander or Hawaiian: 0.7%; American Indian or Alaska Native: 0.3%; Preferred not to say: 12.0%
HAILEY's modeling components	
- Model output:	Empathic rewriting of the response post
- Target user:	Peer supporter (users who provide peer-to-peer support to the support seeker)
- Data splitting:	Training: 3.2M; Test: 0.1M; Validation: 0.1M (seeker post, response post) pairs
- Gold standard:	180 empathic rewritings from human therapy experts used for evaluation of the original PARTNER model[47].
- Model task:	Text generation
- Model architecture:	Deep reinforcement learning with a transformer based language model as its policy.
- Optimization:	Based on reward functions to increase empathy in posts and maintain text fluency, sentence coherence, context specificity, and diversity.
- Internal validation:	Automatic and human evaluation on hold-out test set.
- External validation:	The empathy scale used in HAILEY and PARTNER[47] has previously been shown to correlate with "likes" from the support seeker and the forming of relationships b/w support seekers and peer supporters[29], consistent with empathy theory[25,27,28]. Our present randomized controlled trial represents an external evaluation of the rewriting modeling components of HAILEY and PARTNER[47].
- Transparency:	Data is available from TalkLife through a Data License Agreement; code is available via GitHub (github.com/behavioral-data/PARTNER).

Figure S1. Empathy training used in our study. Participants in both the Human + AI (treatment) and Human Only (control) groups received the same training. The training included the empathy definition, a framework of common ways of expressing empathy in responses, and examples of empathic responses. This ensures that participants were working under similar understandings of empathy. In practice, such training is very rare and the effect of training typically diminishes over time. The identified difference in empathy between treatment and control groups in our study therefore conservatively estimates the impact of our AI-in-the-loop feedback system, and not baseline differences in empathy definitions. The effect in practice may be larger than the intentionally conservative estimates produced here, as such training is uncommon on current mental health platforms.

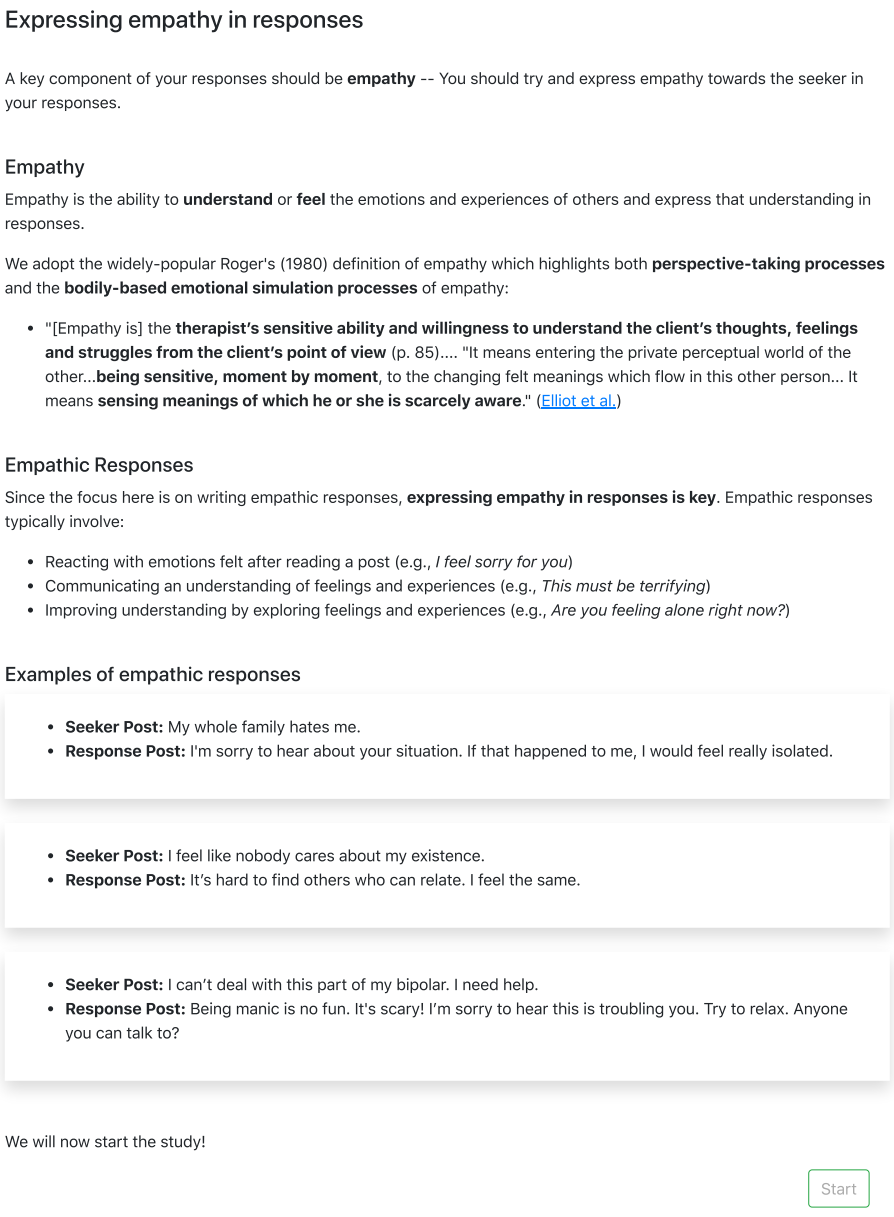
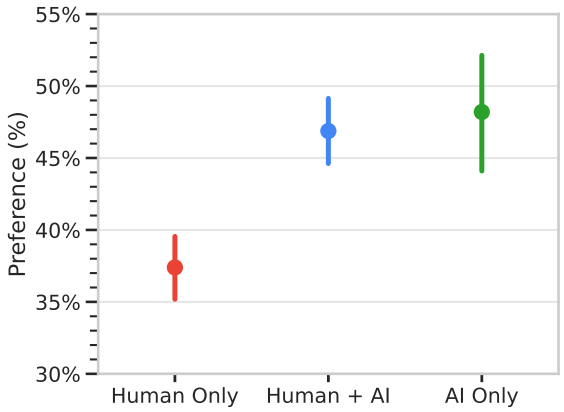
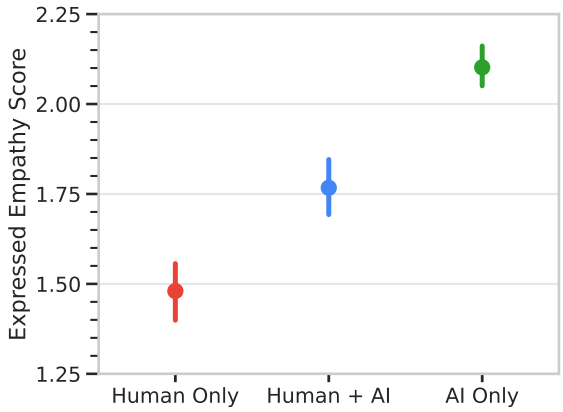


Figure S2. Comparison of Human Only (control) and Human + AI (treatment) responses with AI Only responses (generated directly from PARTNER[47], the deep reinforcement learning model for empathic rewriting, used as a foundation for HAILEY (Methods)). **(a)** Through human evaluation from an independent set of TalkLife users, we found that AI Only responses have a similar preference as the Human + AI responses (48.2% vs. 46.9%; N=161; p=0.23; Two-sided Student's t-test) but a higher preference than the Human Only responses (48.2% vs. 37.4%; N=139; $p=3.3*10^{-5}$; Two-sided Student's t-test). **(b)** Automatic estimation of empathy, on the contrary, suggested that AI Only responses have a higher expressed empathy score compared to Human + AI responses (2.10 vs. 1.77; N=139; Cohen's d=0.28; $p=5.3*10^{-13}$; Two-sided Student's t-test). Importantly however, note that the AI Only responses were optimized on the same scoring function that we use to automatically estimate empathy, which likely explains the high scores of the AI Only approach. **(c)** However, while the authenticity of Human Only and Human + AI responses was comparable (69.6% vs. 65.4%; N=139; p=0.01; Two-sided Student's t-test), the authenticity of AI Only responses was significantly lower (36.5% vs. 65.4%; N=161; $p=3.7*10^{-8}$; Two-sided Student's t-test). This highlights the key issue of authenticity with using AI Only, alongside safety, privacy, bias and other unintended consequences in the high-risk setting of mental health. To summarize, we find that Human + AI is the only approach that leads to both high empathy and high authenticity. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

(a) Human Evaluation: Which response is more empathic?



(b) Automatic/AI-based Evaluation: Expressed empathy score



(c) Authenticity: Is the response human-written or computer-generated?

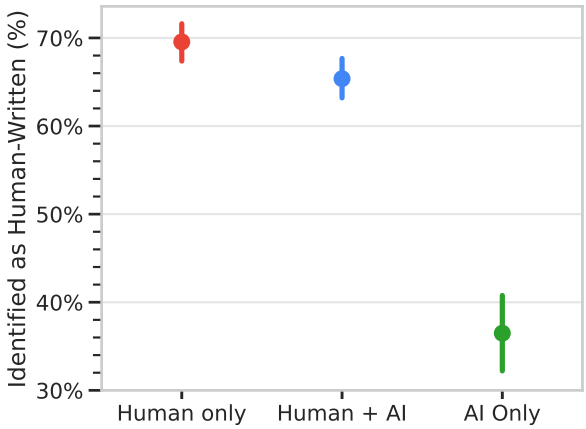


Figure S3. Perceptions of Human + AI (treatment) group participants as reported in phase IV (post-intervention survey). We observed that more than 63.3% of participants found the current feedback helpful, 60.4% found it actionable and 69.8% of participants self-reported feeling more confident at providing support after our study. Also, 77.7% of participants wanted this type of feedback system to be deployed on TalkLife or other similar peer-to-peer support platforms, indicating potential opportunities for deployment in real-world.

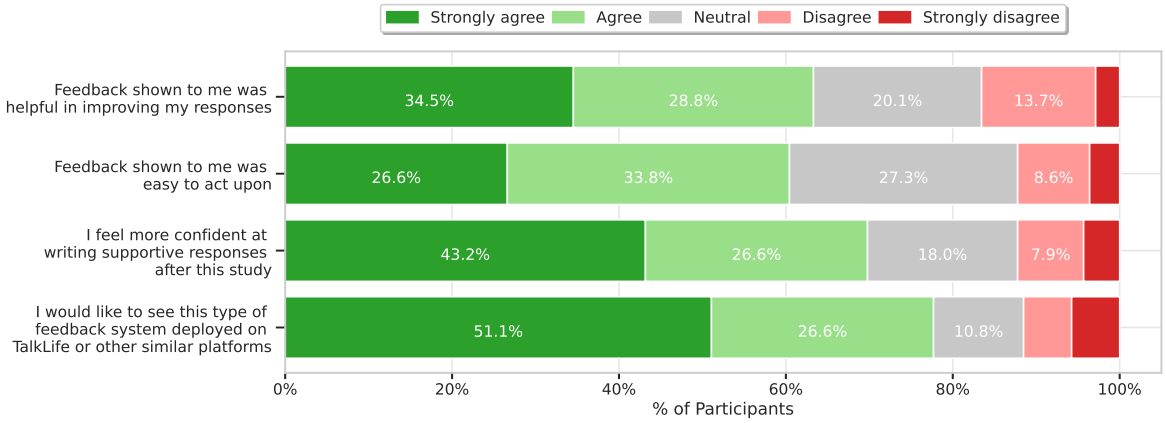
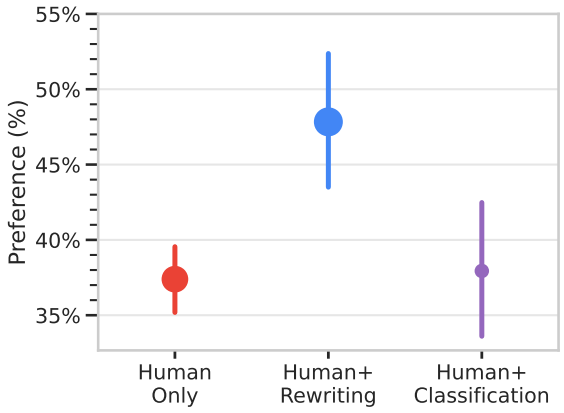
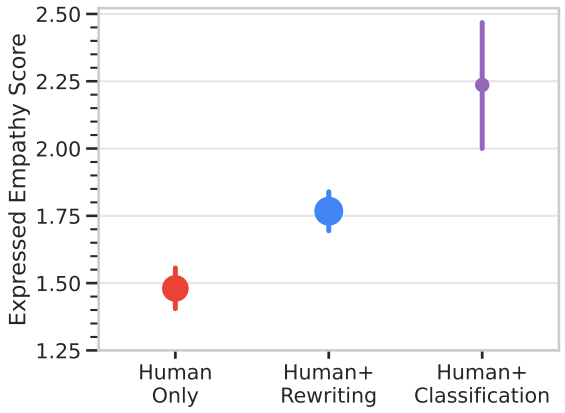


Figure S4. Comparison of our rewriting-based AI treatment with a secondary classification-based AI treatment. A classification-based AI treatment provided participants with an option to request empathy classification scores for their responses, as opposed to the more granular feedback consisting of concrete suggestions to edit responses in our primary rewriting-based approach (Supplementary Figure S5). Our hypothesis was that such a treatment should be less actionable and is likely to lead to less empathic responses than the rewriting-based treatment. In our study, we assigned a secondary classification-based treatment to 10% of the incoming participants at random (N=30). **(a)** Through human evaluation from an independent set of TalkLife users, we found that the Human + Classification responses have a significantly lower preference than the Human + Rewriting responses (37.9% vs. 47.8%; N=30; p=0.002; Two-sided Student's t-test). **(b)** Automatic estimation of empathy, on the contrary, suggested that the Human + Classification responses have a higher expressed empathy score compared to Human + Rewriting responses (2.24 vs. 1.77; N=30; Cohen's d=0.37; $p=4.7*10^{-6}$; Two-sided Student's t-test). As the same score is also exposed to participants just-in-time in the classification-based treatment, it may have led participants to be put particular emphasis on a high expressed empathy score, which participants in the rewriting-based treatment feedback didn't have direct access to. **(c)** We found that less participants in the classification-based treatment group agree on deploying the system on TalkLife than the rewriting-based treatment (63.3% vs. 77.7%; N=30; p=0.0998; Two-sided Student's t-test; Supplementary Figure S3). Also, we observed that more participants in the classification-based treatment disagree on its actionability than participants in the rewriting-based treatment, but the difference may not be statistically significant due to the limited power (23.3% vs. 12.2%; N=30; p=0.1154; Two-sided Student's t-test). The area of the points in the plots is proportional to the number of participants in the respective control/treatment conditions. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

(a) Human Evaluation: Which response is more empathic?



(b) Automatic/AI-based Evaluation: Expressed empathy score



(c) Study participants' perceptions (classification-based treatment group)

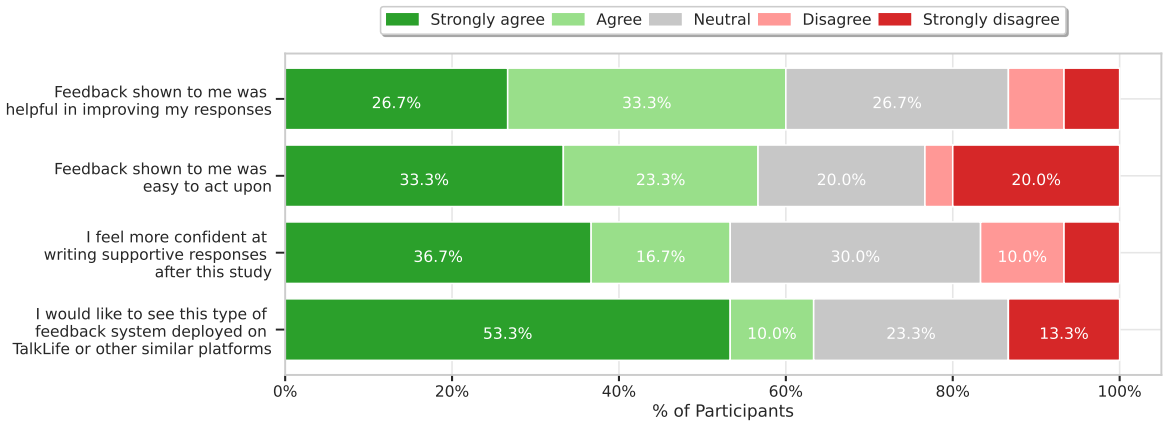


Figure S5. Interface of our classification-based AI treatment (Supplementary Figure S4). **(a)** Participant is asked to write a supportive, empathic response and given an option to receive feedback. **(b)** Participant starts writing the response. **(c)** Participant clicks on the "Get Feedback" button to request classification-based feedback. The feedback consists of classification scores on three empathy communication mechanisms – Emotional Reactions, Interpretations, and Explorations[29]. **(d)** Participant edits the response based on the classification scores, often improving on the communication mechanisms with low scores and requests "More Feedback" if needed.

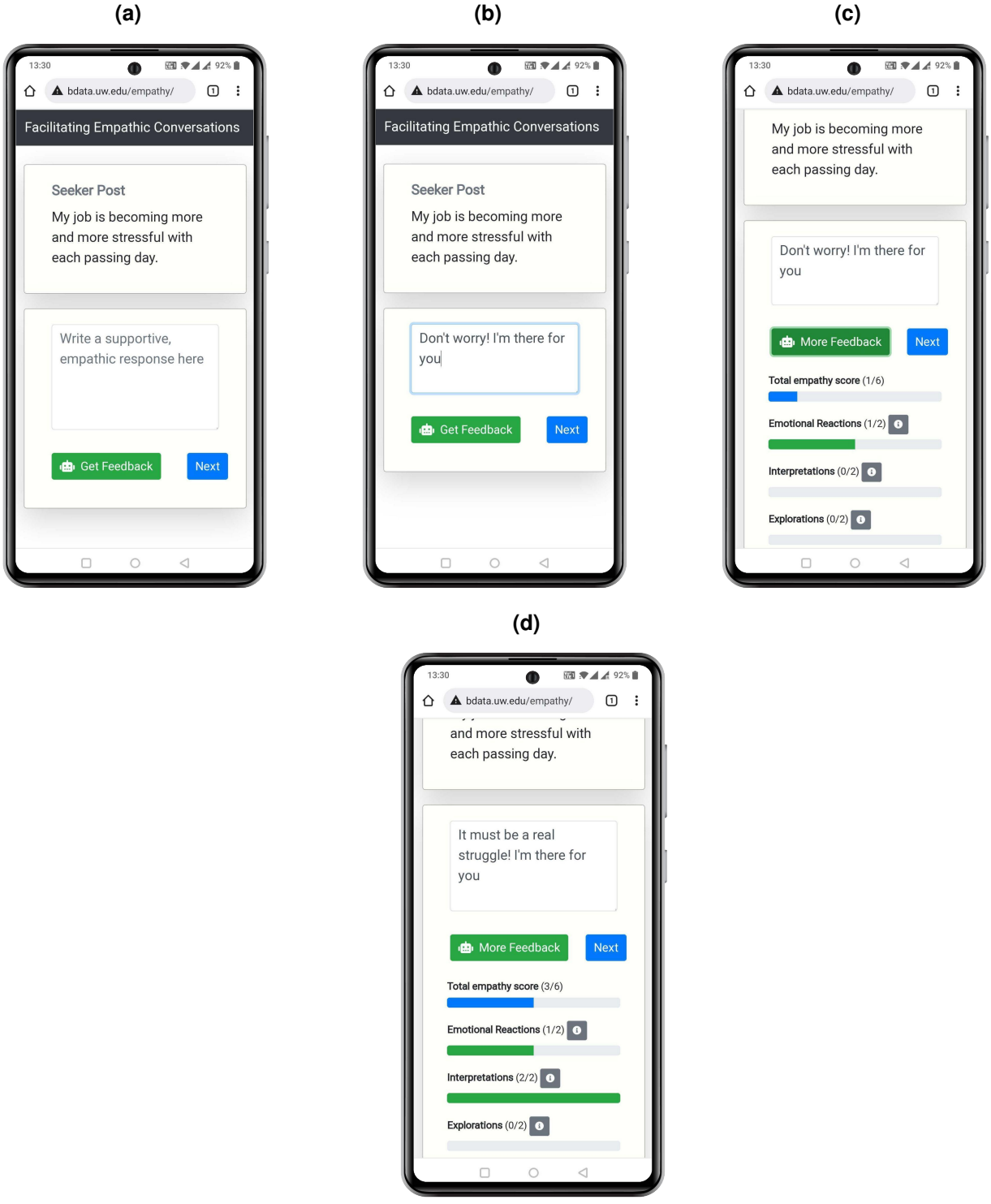


Figure S6. Both Human Only (control) and Human + AI (treatment) group participants showed a significant drop in empathy levels in the last 5 responses of our study. With Human + AI, however, we observed a significantly lower drop in empathy (5.3% vs. 26.0%; N=139; p=0.0062; Two-sided Student's t-test). This indicates the effectiveness of just-in-time AI feedback in alleviating challenges like empathy fatigue, associated with providing mental health support. The empathy differences between Human Only (N=161) and Human + AI (N=139) responses are statistically significant for both first 5 and last 5 responses ($p=1.1 * 10^{-8}$; Two-sided Student's t-test). The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

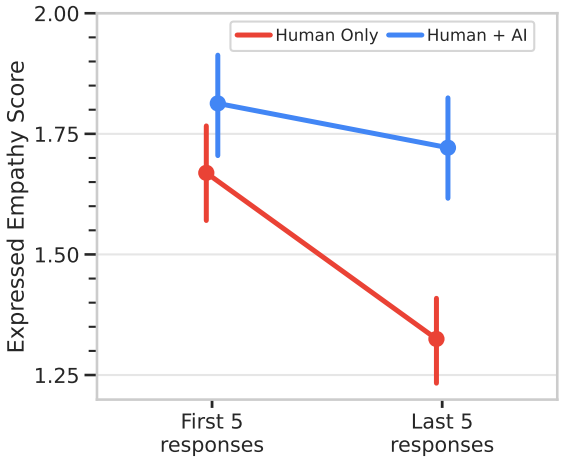
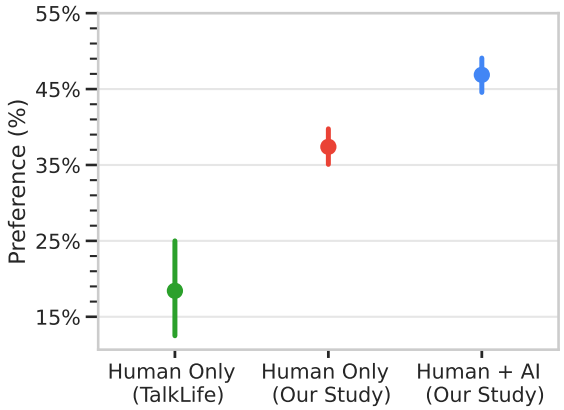


Figure S7. Comparison of existing Human Only responses on TalkLife with Human Only and Human + AI responses in our study. Human Only responses on TalkLife had significantly lower preference for empathy (18.4% vs. 37.4% vs. 46.9%; N=139; $p=1.4*10^{-6}$; Two-sided Student's t-test) and significantly lower expressed empathy score (1.11 vs. 1.48 vs. 1.77; $p=2.8*10^{-46}$; Two-sided Student's t-test). This difference might be attributed to the additional initial empathy training provided to participants, as well as a potential selection effect in our study that may have attracted Talklife users who particularly care about expressing empathy in supporting others. As our study shows that Human-AI collaboration improves empathy expression even for those participants who already express empathy more often, practical gains for the average user of the Talklife platform could be even higher. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

(a) Human Evaluation: Which response is more empathic?



(b) Automatic/AI-based Evaluation: Expressed empathy score

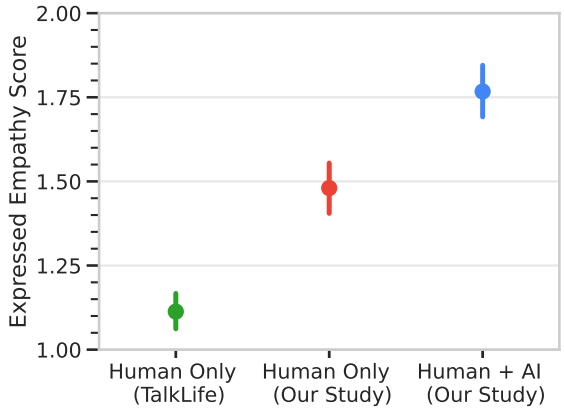


Figure S8. Qualitative examples of just-in-time AI feedback provided to participants by HAILEY. In **(b)** and **(c)**, the original peer supporter response was empty. Seeker posts in these examples have been paraphrased for anonymization.

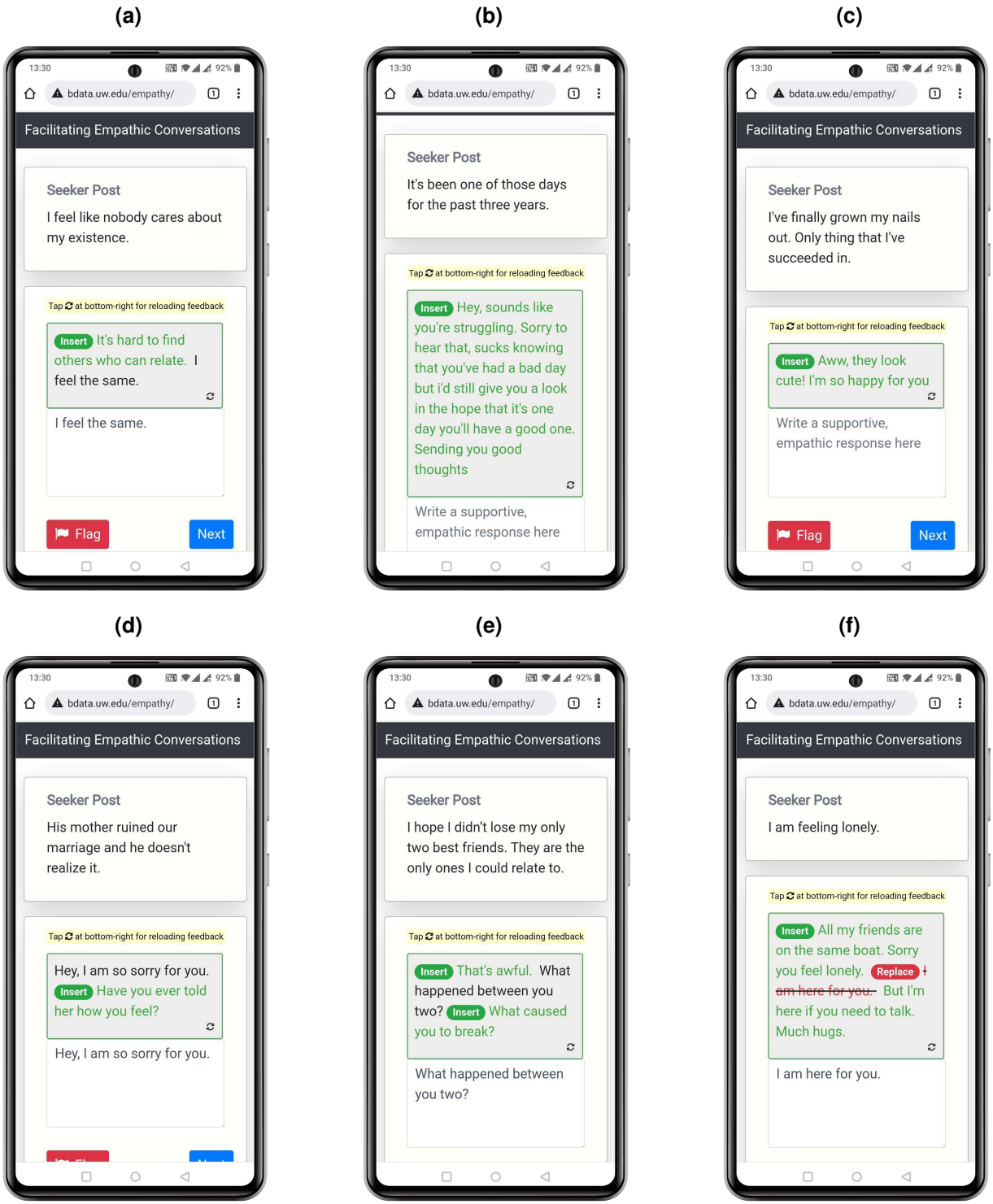
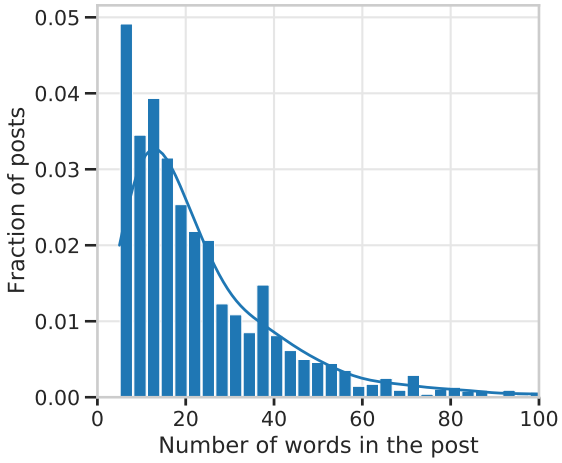


Figure S9. The distribution of post and response lengths. The seeker posts in our dataset had a mean length of 25.9 words, a standard deviation of 25.3 words and a median of 18.0 words. The response posts collected in our study had a mean length of 25.9 words, a standard deviation of 34.6 words and a median of 19.0 words. Also, Human + AI responses (mean = 22.4 words; std = 34.6; median = 19.0) were 28.9% shorter in length compared to Human Only responses on average (mean = 32.1 words; std = 44.4; median = 21.0; $p < 0.001$; Two-sided Student's t-test). In addition, we found that Human + AI responses had 5.2% higher diversity than the Human Only responses based on the Distinct-1 metric[84], which computes the number of distinct unigrams divided by the total number of tokens (0.146 vs. 0.139; $p = 0.019$; Two-sided Wilcoxon signed-rank test).

(a) Seeker Post



(b) Response Post

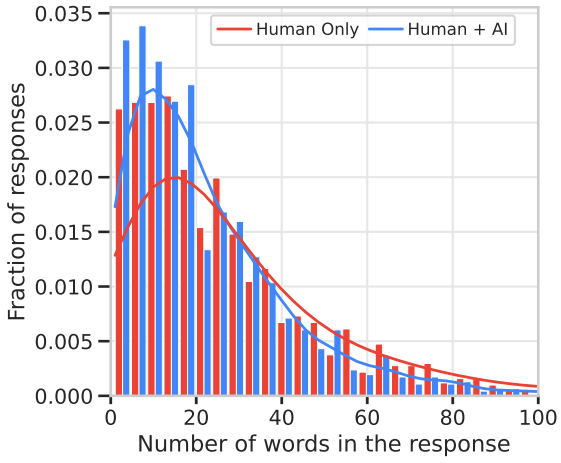
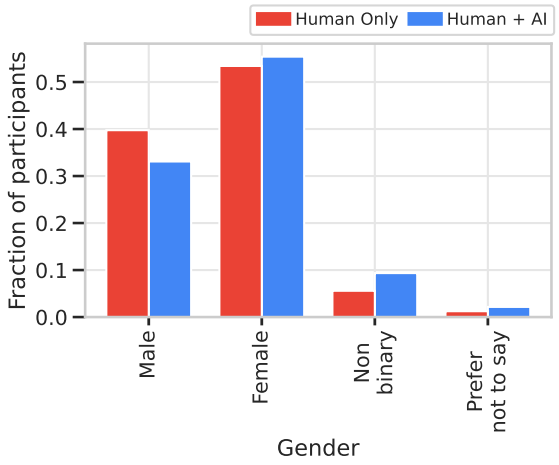
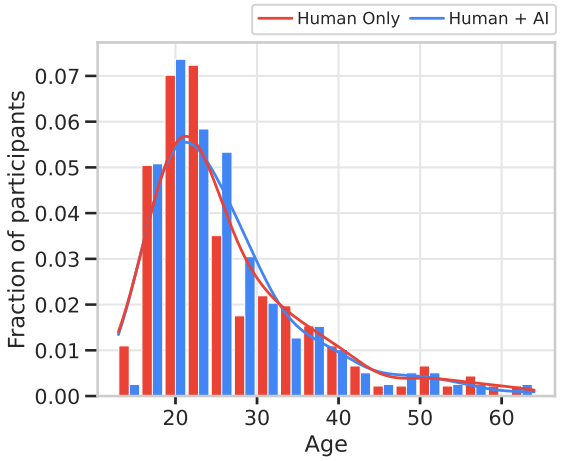


Figure S10. Background and demographics of participants in Human Only (control) and Human + AI (treatment) groups, as reported in phase I (pre-intervention survey).

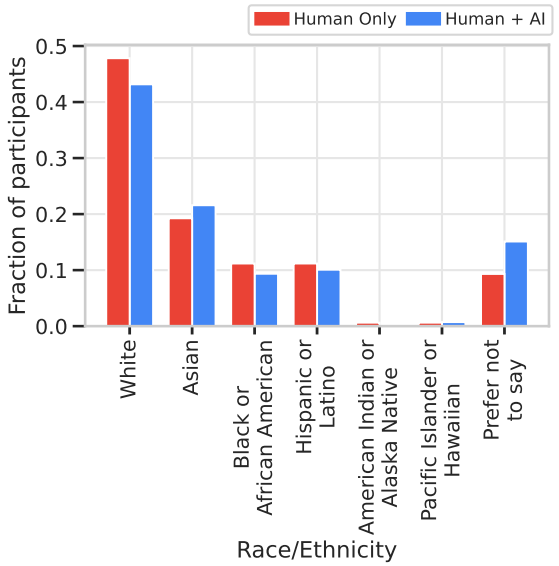
(a) Gender



(b) Age



(c) Race/Ethnicity



(d) Experience with online peer support

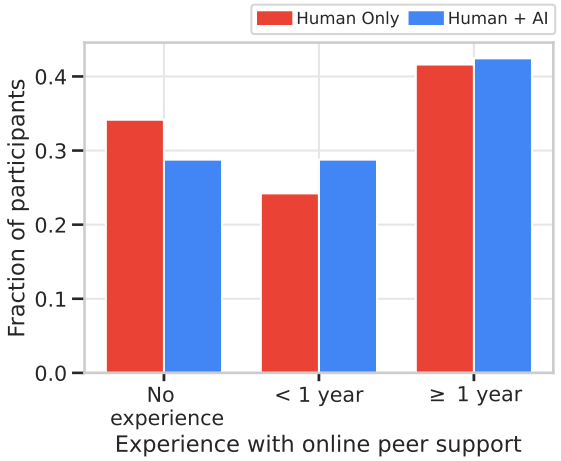
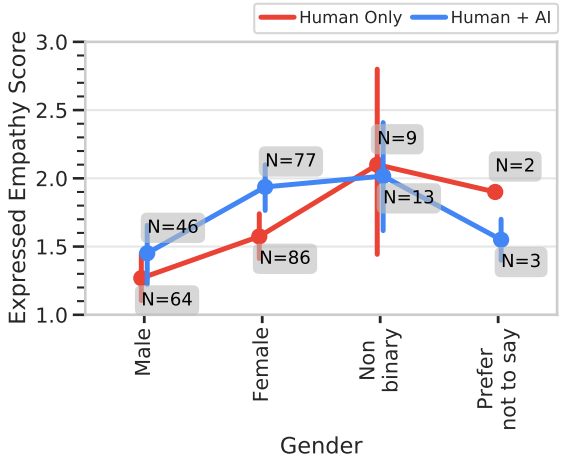
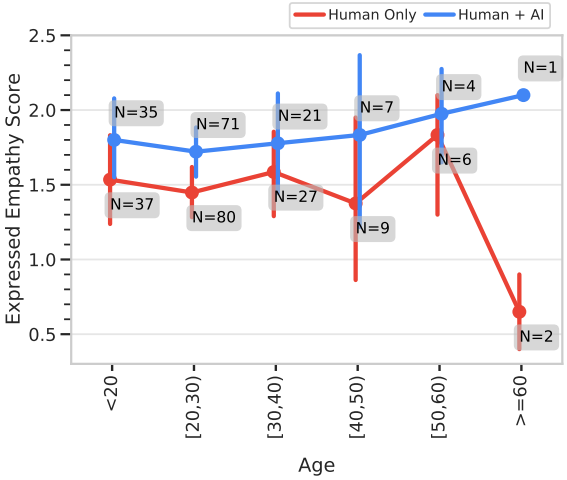


Figure S11. Differences between expressed empathy scores of participants in Human Only (control) and Human + AI (treatment) groups, stratified by demographics of participants and their prior experience with online peer support. The area of the points is proportional to the number of participants in the respective categories. The point estimates represent the mean and the error bars represent bootstrapped 95% confidence intervals.

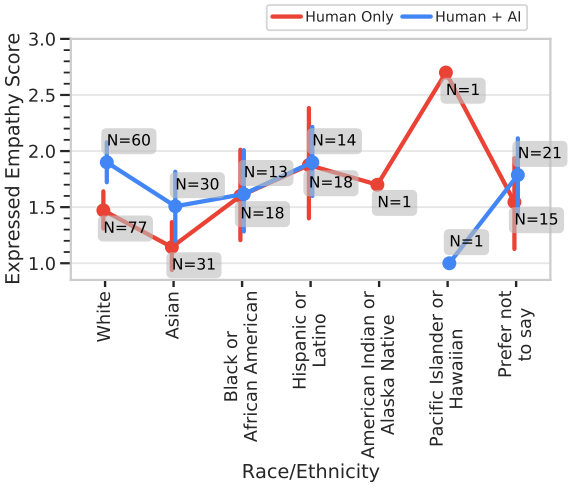
(a) Gender



(b) Age



(c) Race/Ethnicity



(d) Prior experience with online peer support

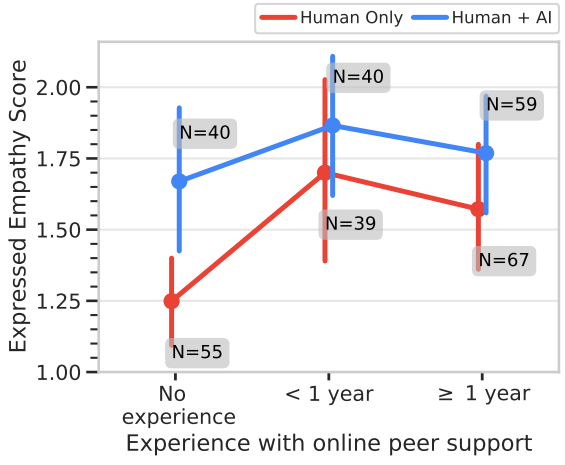
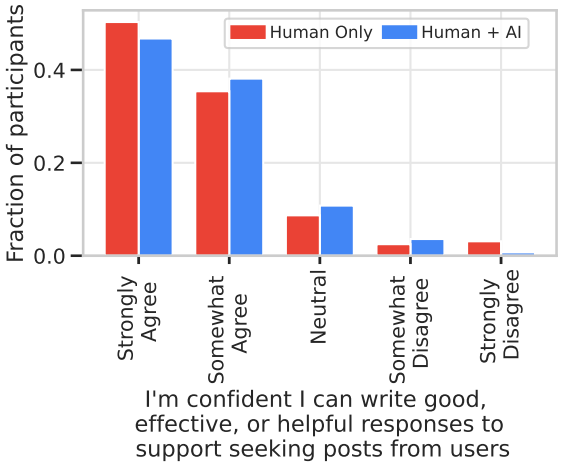
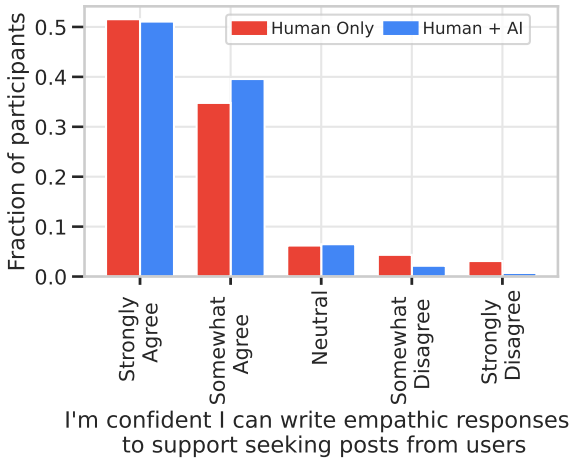


Figure S12. Perceptions of participants in Human Only (control) and Human + AI (treatment) groups, as reported in phase I (pre-intervention survey).

(a) Self-efficacy in writing good, effective, or helpful responses



(b) Self-efficacy in writing empathic responses



(c) Could feedback be helpful?

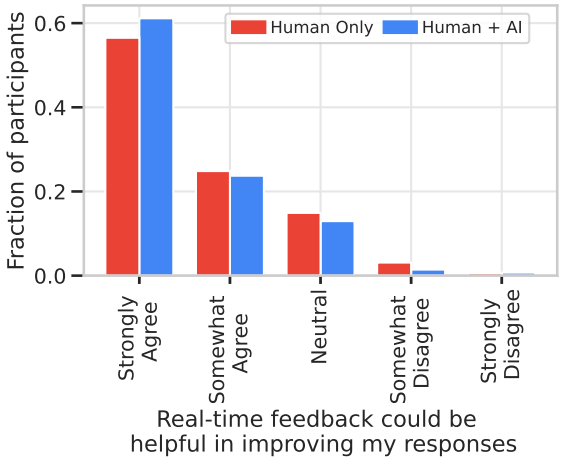


Figure S13. Distribution of participants in Human Only (control) and Human + AI (treatment) groups who report writing responses as challenging or stressful, as reported in phase IV (post-intervention survey).

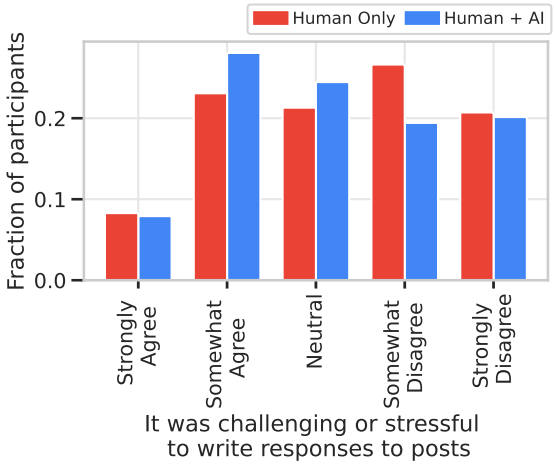


Figure S14. Distribution of participants in the Human Only (control) group who indicate that feedback could have improved responses, as reported in phase IV (post-intervention survey).

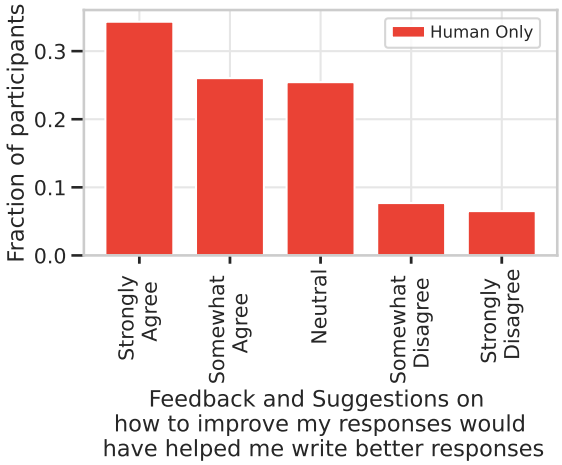
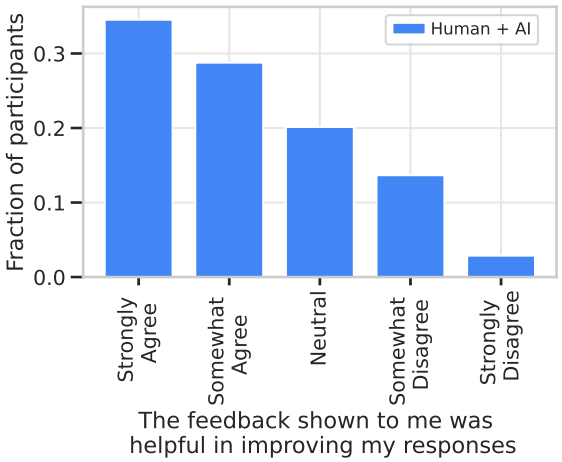
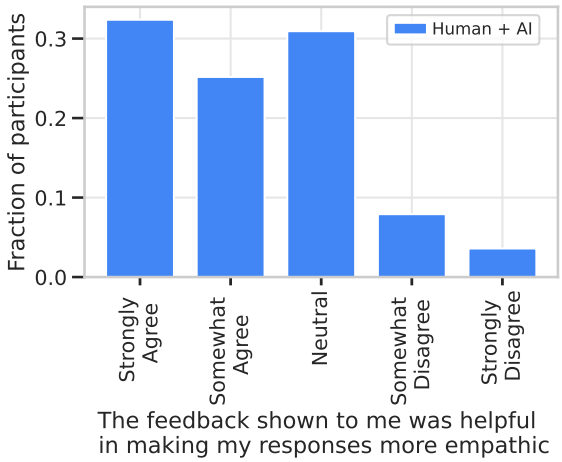


Figure S15. Perceptions of participants in the Human + AI (treatment) group, as reported in phase IV (post-intervention survey).

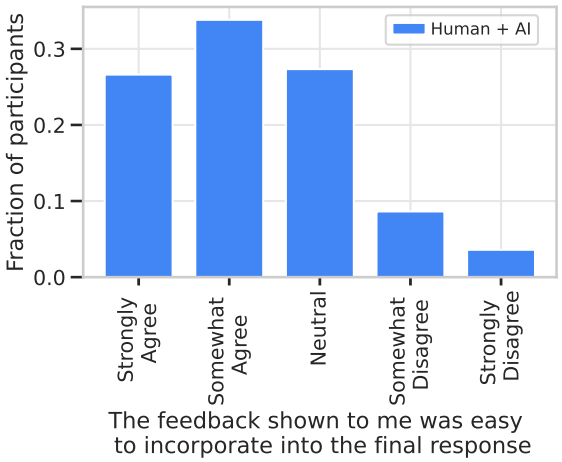
(a) Overall helpfulness



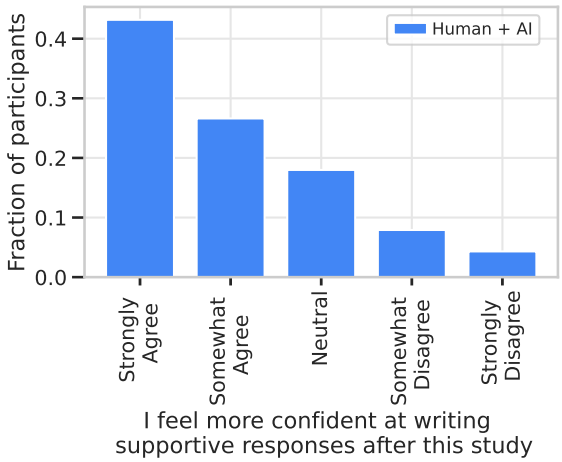
(b) Helpfulness for empathy



(c) Actionability



(d) Self-efficacy



(e) Intention to adopt

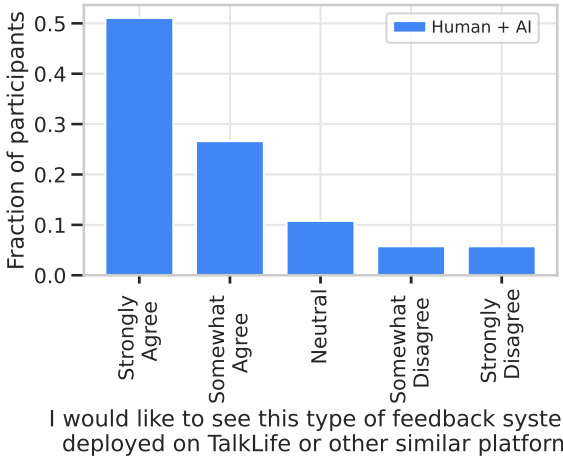
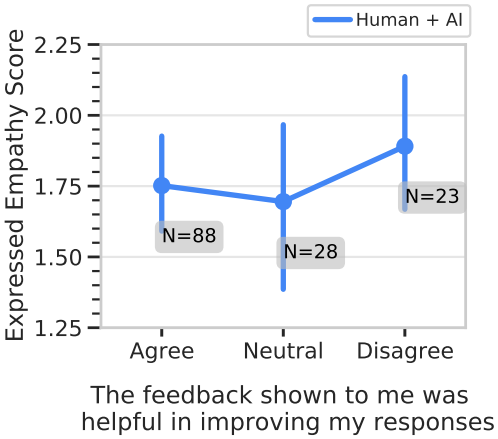
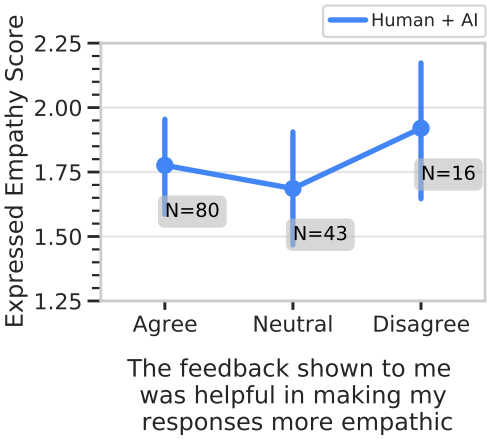


Figure S16. Expressed empathy levels of responses with perceptions of Human + AI (treatment) group participants, as reported in phase IV (post-intervention survey). The area of the points is proportional to the number of participants with respective perceptions. Error bars indicate bootstrapped 95% confidence intervals.

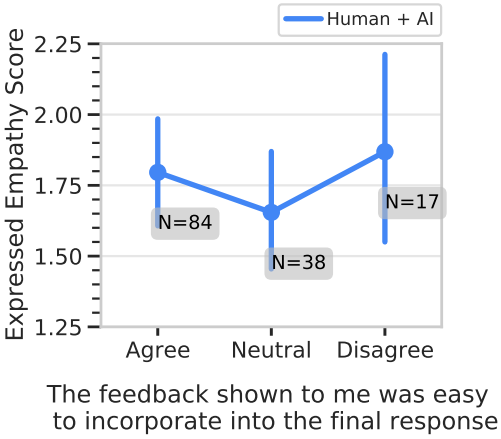
(a) Overall helpfulness



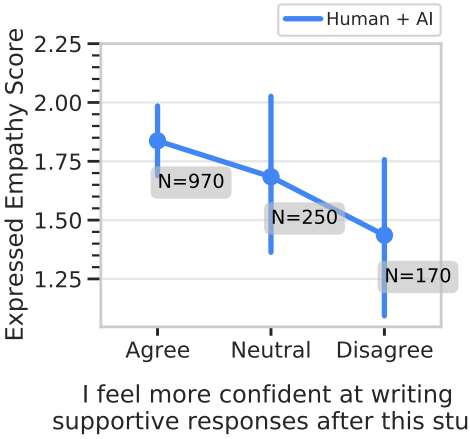
(b) Helpfulness for empathy



(c) Actionability



(d) Self-efficacy



(e) Intention to adopt

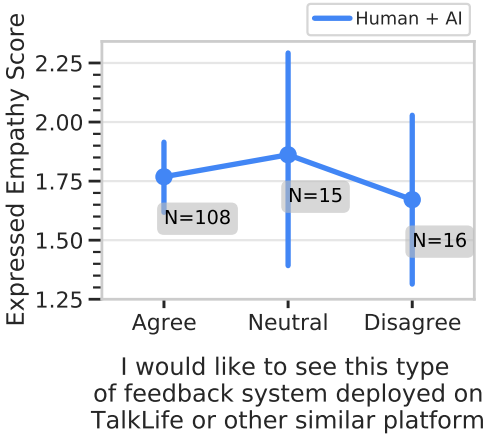
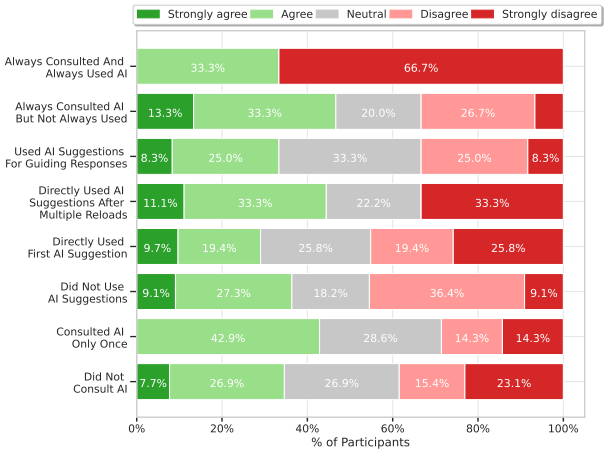
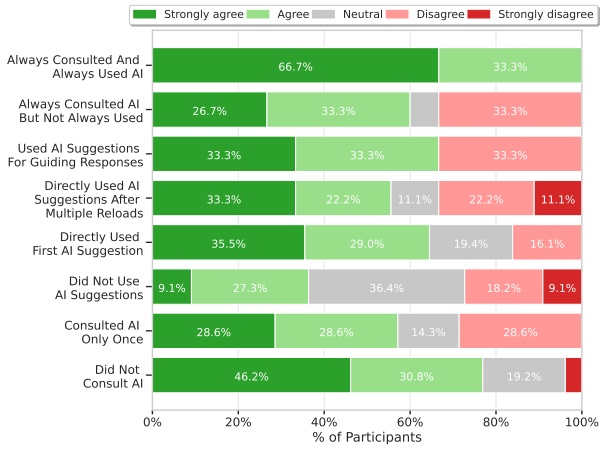


Figure S17. Participant perceptions, as reported in phase IV (post-intervention survey), with different human-AI collaboration categories.

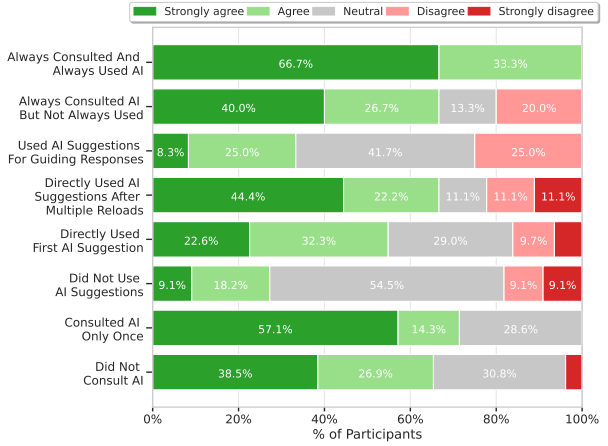
(a) Challenges: Feedback and Suggestions on how to improve my responses would have helped me write better responses



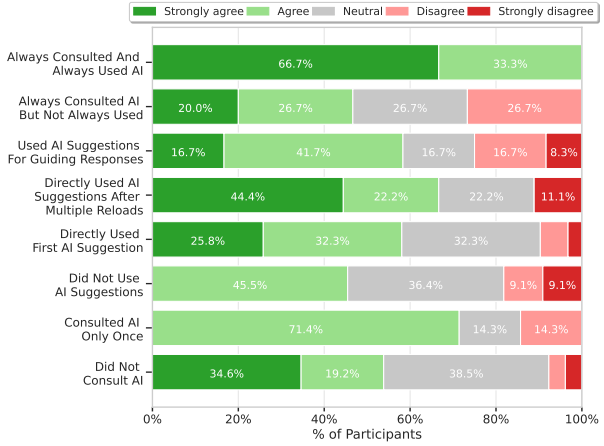
(b) Overall helpfulness: The feedback shown to me was helpful in improving my responses



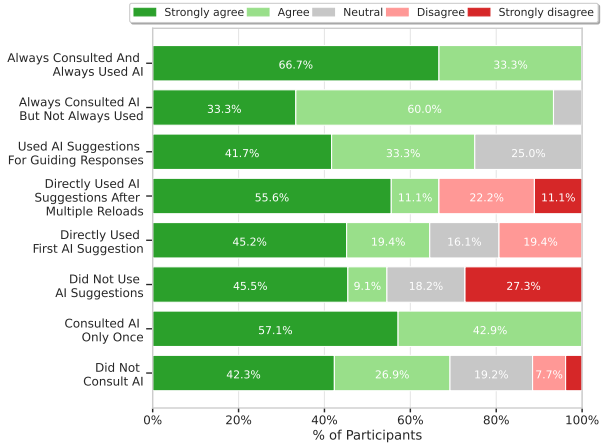
(c) Helpfulness for empathy: The feedback shown to me was helpful in making my responses more empathic



(d) Actionability: The feedback shown to me was easy to incorporate into the final response



(e) Self-efficacy: I feel more confident at writing supportive responses after this study



(f) Intention to adopt: I would like to see this type of feedback system deployed on TalkLife or other similar platforms

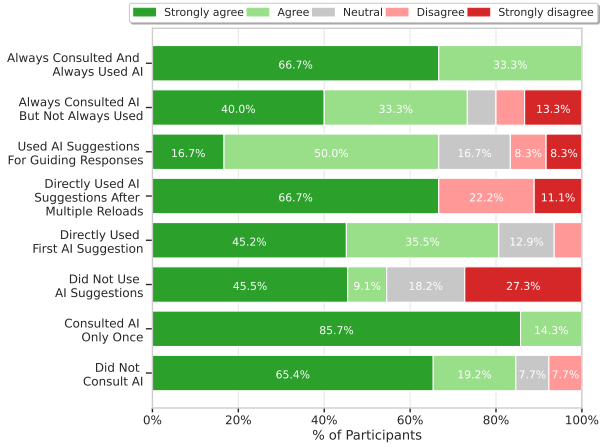


Figure S18. Consent form used in our study.

Disclaimer

Thank you for your interest in our study!

- We are **researchers at the University of Washington**, studying peer-to-peer support platforms.
- This study is **not being conducted by TalkLife**. TalkLife is not responsible for any risks/benefits associated with this study.
- As part of this study, we will be collecting (a) **textual responses to support seeking posts**, and (b) **answers to survey questions** describing your background and your assessment of our study. The data from this study will be uploaded to a **secure platform accessible only to the research team**. All collected data will be exclusively used for research purposes. Also, data will only be analyzed in aggregate.
- This study has been determined to be exempt from IRB approval under University of Washington IRB ID STUDY00012706. For concerns or questions, please contact ashshar@cs.washington.edu.
- This is **not a live conversation** with users. Your responses will not be posted on TalkLife or any other online platform.
- Your participation in this study is **completely voluntary**. You are free to release/quit the study at any time. Refusing to be in the experiment or stopping participation will involve no penalty or loss of benefits to which you are otherwise entitled.
- Poor-quality data (e.g., same response to all posts) **may be removed** without compensation.
- If you are a US citizen or a permanent US resident, you will receive an **Amazon gift card worth 5 USD** after completing this study.

  ○ **Yes**, I'm a US citizen or a permanent US resident.
  ○ **No**, I'm neither a US citizen nor a permanent US resident.
  ○ Prefer not to say.

Consent

☐ I agree to participate in this study. I also understand that only US citizens or permanent US residents can be compensated.

Accept and Continue

Figure S19. Form used for collecting demographics and background of participants [phase I: pre-intervention survey].

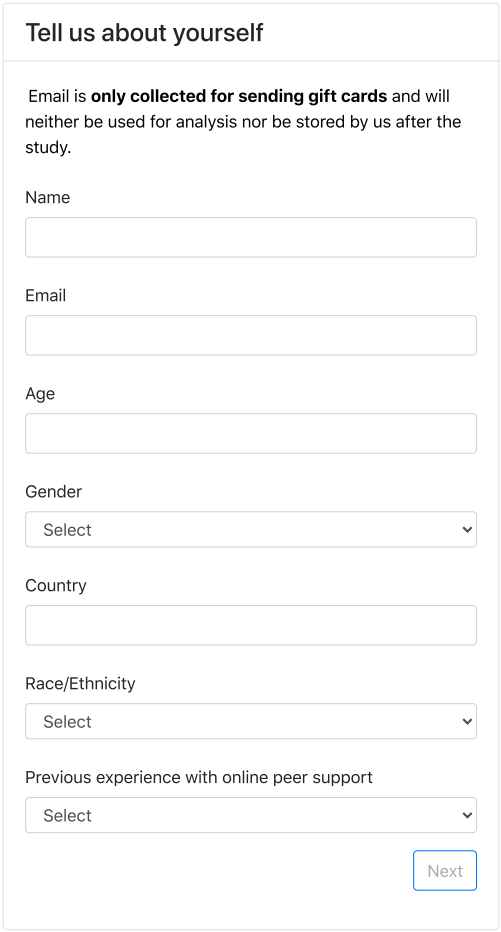


Figure S20. Onboarding survey used for collecting perceptions of participants [phase I: pre-intervention survey].

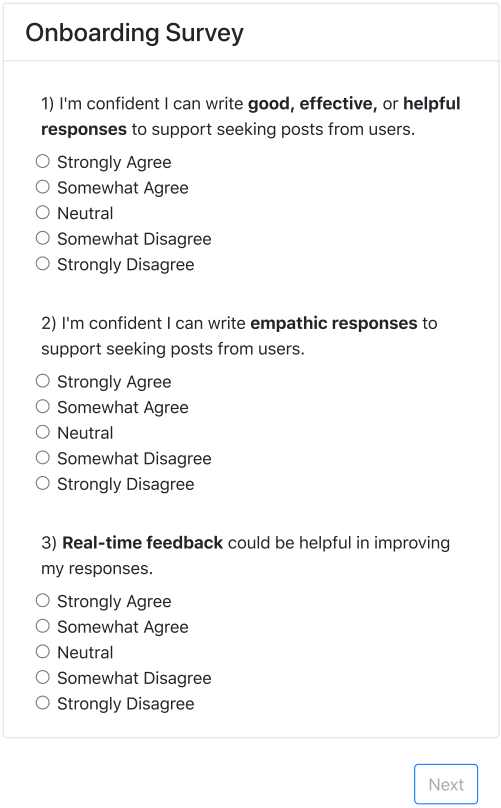


Figure S21. Instructions shown to the control group participants [phase II: empathy training and instructions]. Continued on the next page (1/2).

Instructions

The study will involve writing textual responses to support seeking posts and answering survey questions. The entire study is expected to take ~30 minutes.

Content Warning

The study contains posts including but not limited to self-harm and suicidal ideation, which may be disturbing to you. If you have concerns or questions, please send us an email (ashshar@cs.washington.edu). If you have strong negative reactions to some of the content, please reach out at crisis text line.

What will you do?

During the main part of the study, you will be shown online mental health posts through which people seek support. We call them "Seeker Posts" (writers of these posts are called "Seekers"). For each seeker post, you will be asked to **write a supportive, empathic response**.

Also, at various steps, you will be asked **survey questions** describing your background and your assessment of our study.

Expressing empathy in responses

A key component of your responses should be **empathy** -- You should try and express empathy towards the seeker in your responses.

Empathy

Empathy is the ability to **understand** or **feel** the emotions and experiences of others. Empathic responses typically involve:

- Reacting with emotions felt after reading a post (e.g., *I feel sorry for you*)
- Communicating an understanding of feelings and experiences (e.g., *This must be terrifying*)
- Improving understanding by exploring feelings and experiences (e.g., *Are you feeling alone right now?*)

Examples of empathic responses

- **Seeker Post:** My whole family hates me.
- **Empathic Response:** I'm sorry to hear about your situation. If that happened to me, I would feel really isolated.

- **Seeker Post:** I feel like nobody cares about my existence.
- **Empathic Response:** It's hard to find others who can relate. I feel the same.

- **Seeker Post:** I can't deal with this part of my bipolar. I need help.
- **Empathic Response:** Being manic is no fun. It's scary! I'm sorry to hear this is troubling you. Try to relax. Anyone you can talk to?

Next

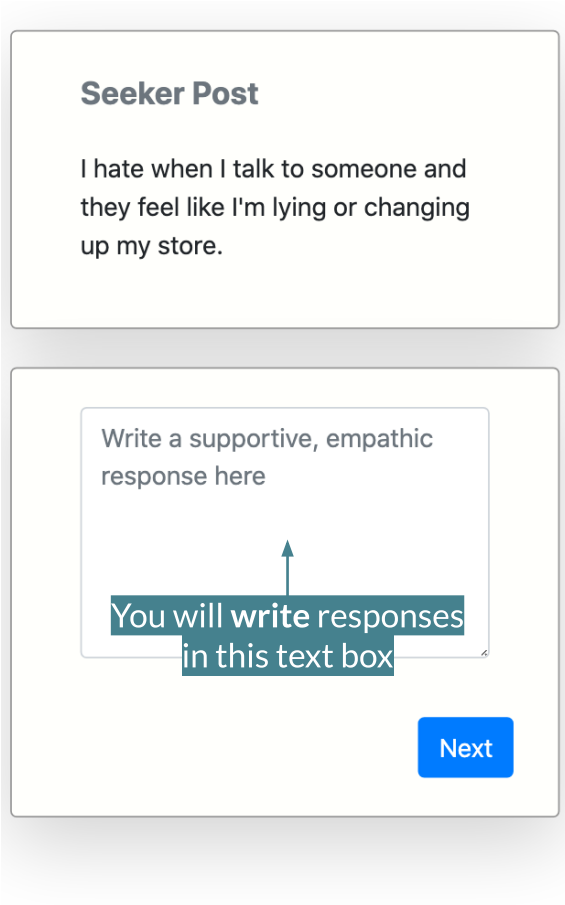
Figure S22. Instructions shown to the control group participants [phase II: empathy training and instructions] (2/2).

Thank you for filling out the survey!

Next, you will **write responses to 10 support seeking posts**. Here, we will give you an overview of our interface.

The Interface

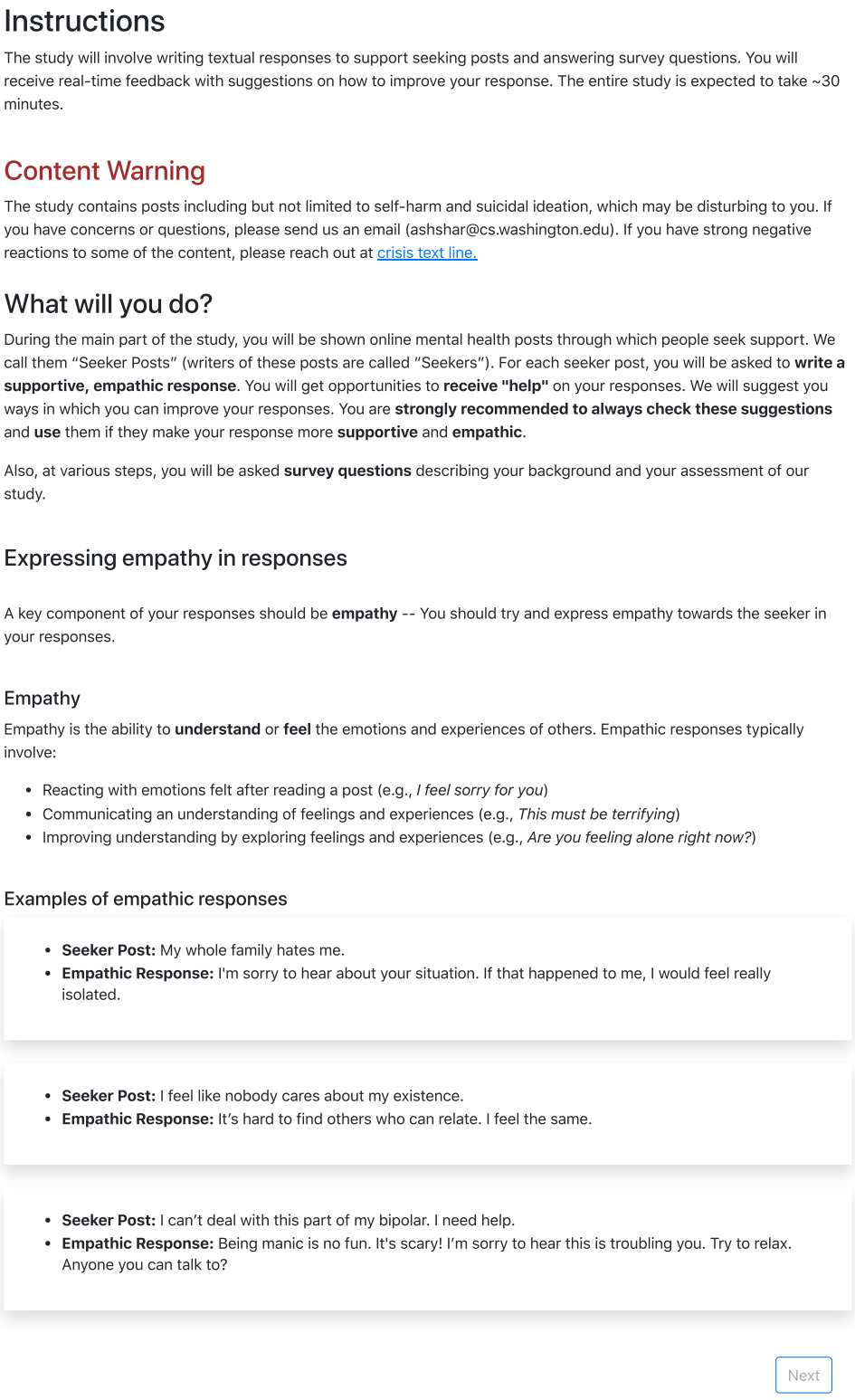
This is how the main interface will look:



You will read the seeker post and **write a supportive, empathic response** in the space provided.

Start Study

Figure S23. Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (1/6).



Instructions

The study will involve writing textual responses to support seeking posts and answering survey questions. You will receive real-time feedback with suggestions on how to improve your response. The entire study is expected to take ~30 minutes.

Content Warning

The study contains posts including but not limited to self-harm and suicidal ideation, which may be disturbing to you. If you have concerns or questions, please send us an email (ashshar@cs.washington.edu). If you have strong negative reactions to some of the content, please reach out at crisis text line.

What will you do?

During the main part of the study, you will be shown online mental health posts through which people seek support. We call them "Seeker Posts" (writers of these posts are called "Seekers"). For each seeker post, you will be asked to **write a supportive, empathic response**. You will get opportunities to **receive "help"** on your responses. We will suggest you ways in which you can improve your responses. You are **strongly recommended to always check these suggestions** and **use** them if they make your response more **supportive** and **empathic**.

Also, at various steps, you will be asked **survey questions** describing your background and your assessment of our study.

Expressing empathy in responses

A key component of your responses should be **empathy** -- You should try and express empathy towards the seeker in your responses.

Empathy

Empathy is the ability to **understand** or **feel** the emotions and experiences of others. Empathic responses typically involve:

- Reacting with emotions felt after reading a post (e.g., *I feel sorry for you*)
- Communicating an understanding of feelings and experiences (e.g., *This must be terrifying*)
- Improving understanding by exploring feelings and experiences (e.g., *Are you feeling alone right now?*)

Examples of empathic responses

- **Seeker Post:** My whole family hates me.
- **Empathic Response:** I'm sorry to hear about your situation. If that happened to me, I would feel really isolated.

- **Seeker Post:** I feel like nobody cares about my existence.
- **Empathic Response:** It's hard to find others who can relate. I feel the same.

- **Seeker Post:** I can't deal with this part of my bipolar. I need help.
- **Empathic Response:** Being manic is no fun. It's scary! I'm sorry to hear this is troubling you. Try to relax. Anyone you can talk to?

Next

Figure S24. Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (2/6).

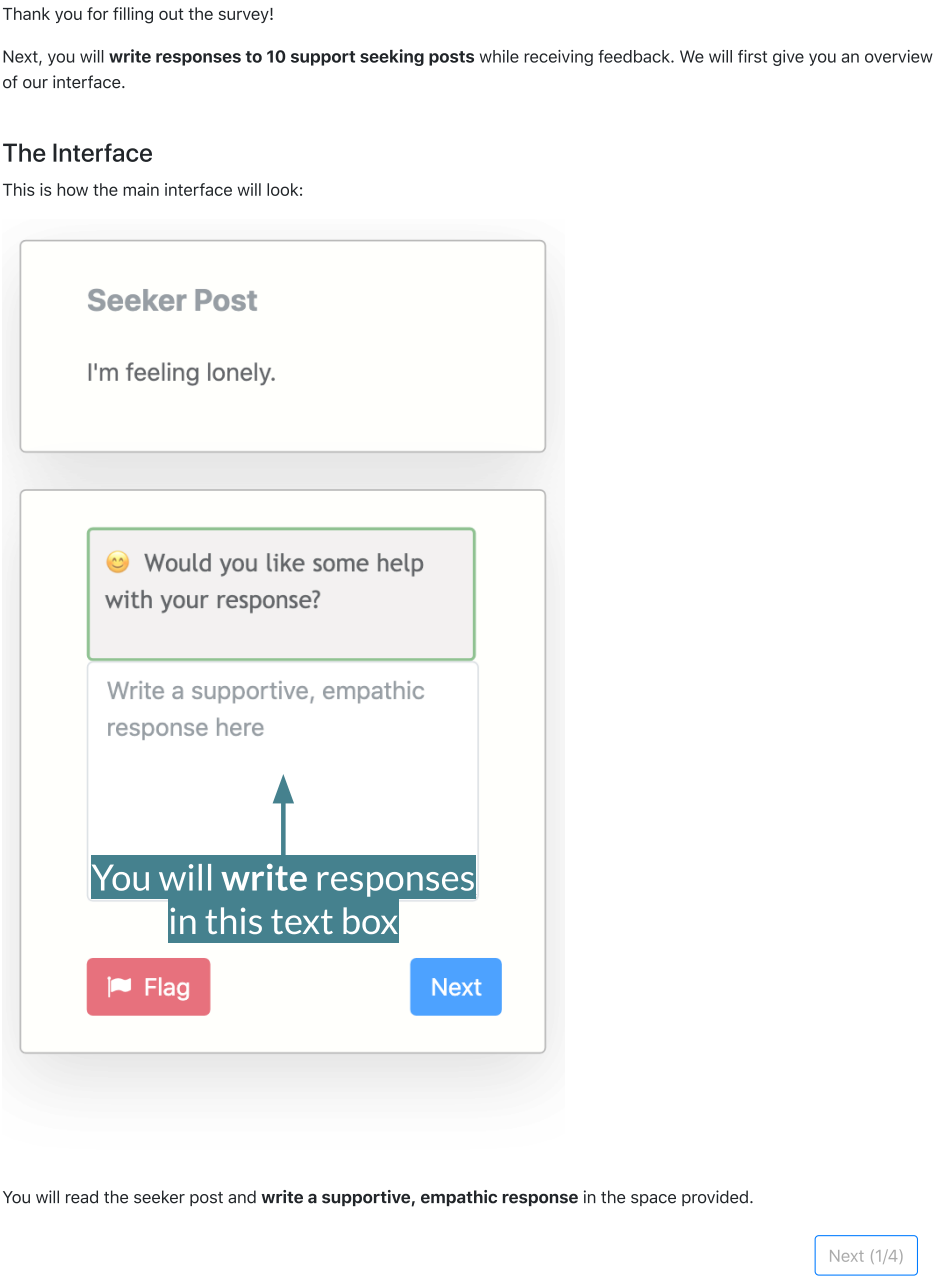


Figure S25. Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (3/6).

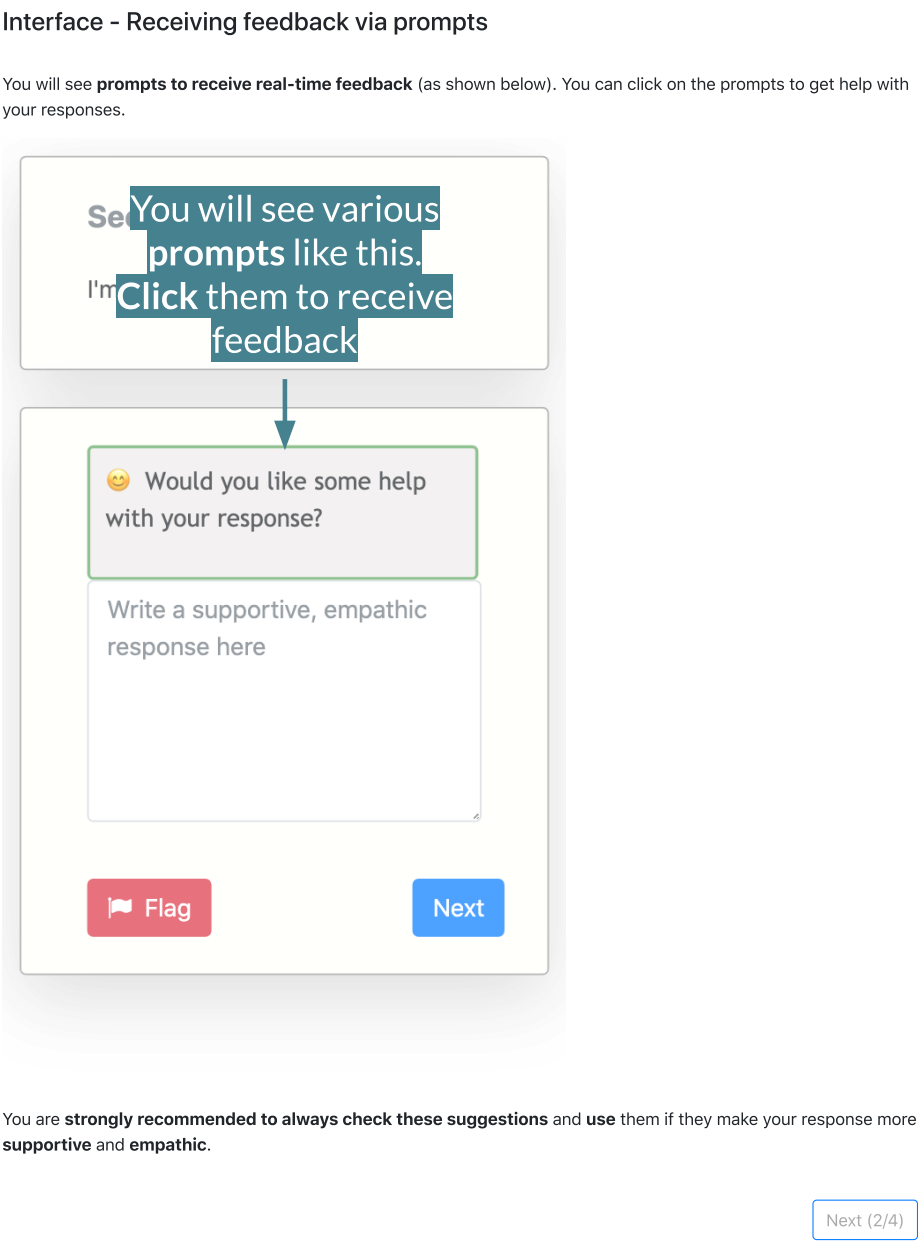


Figure S26. Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (4/6).

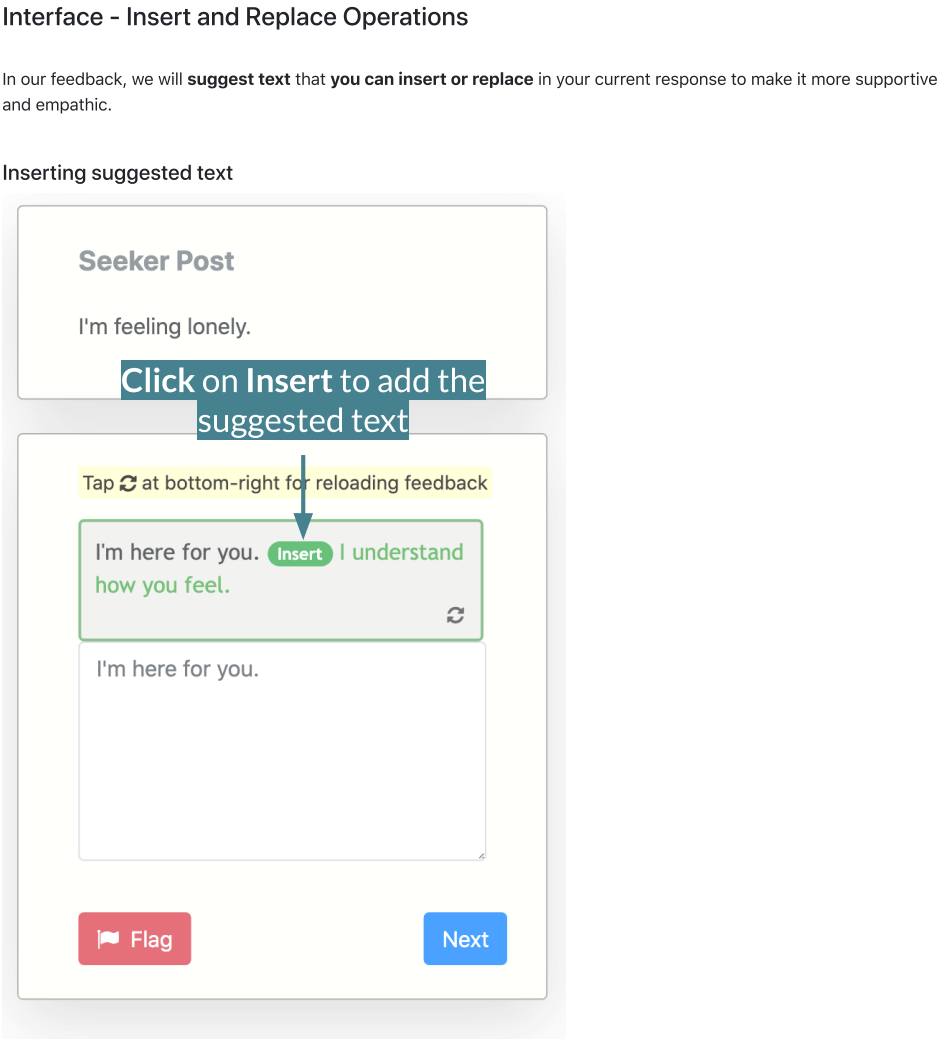


Figure S27. Instructions shown to the treatment group participants [phase II: empathy training and instructions]. Continued on the next page (5/6).

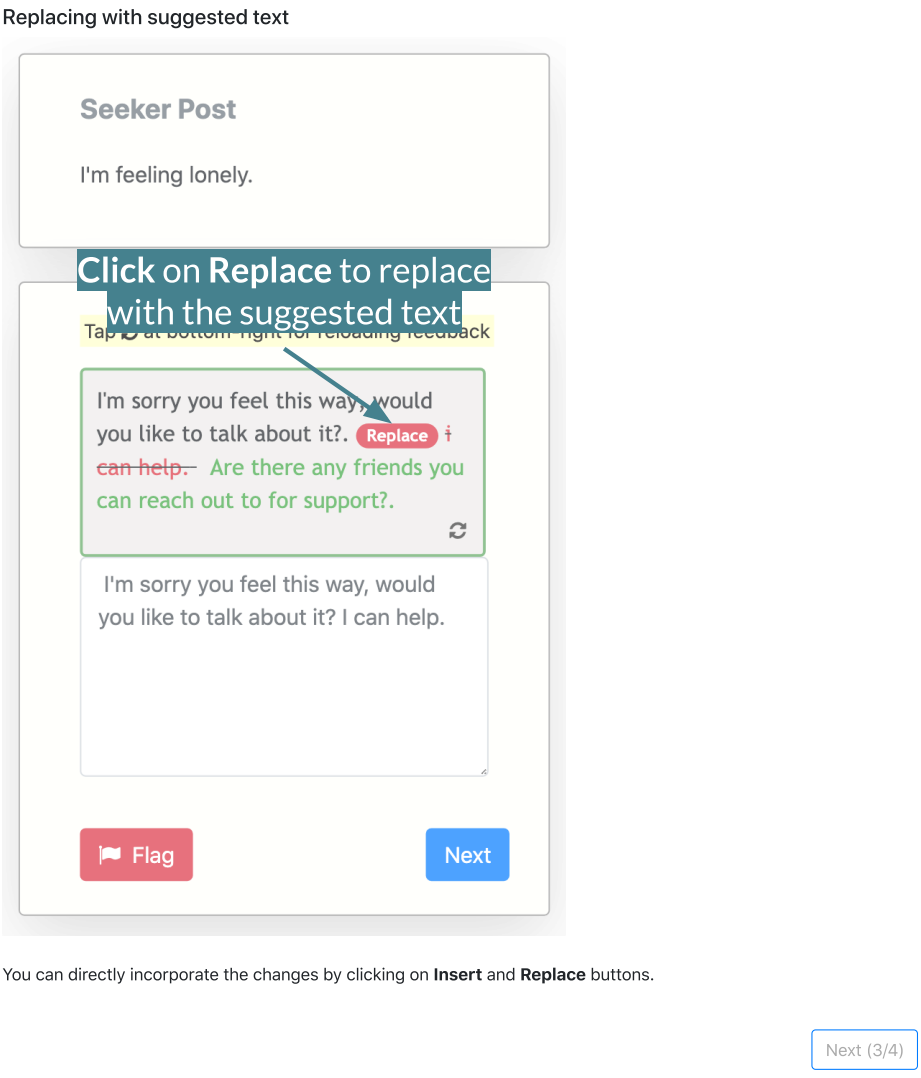


Figure S28. Instructions shown to the treatment group participants [phase II: empathy training and instructions] (6/6).

Interface - Bad feedback

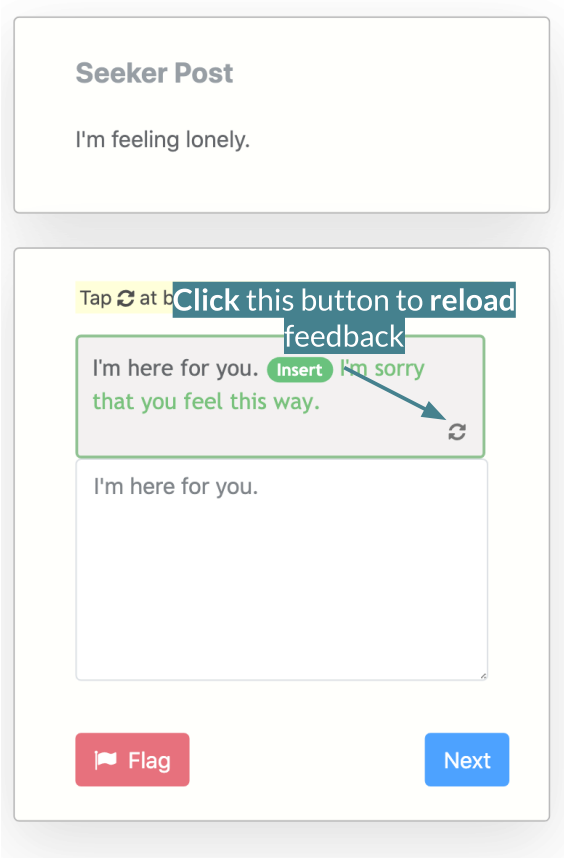
Our feedback will **not** always be perfect. If the feedback is **bad** or **inappropriate**, you may refine, reload, or report the feedback.

Refine

You may need to refine the feedback to **correct grammar** or **content**. You should **make relevant changes** such that the feebdack can be appropriately integrated in your final response.

Reload

Also, whenever the feedback is bad, you can use the ↻ button (see below) to **reload** and **get new feedback**.



You can reload **multiple times** till you see feedback that is helpful.

Report

If you see feedback that is **inappropriate** or **toxic**, you can report it using the **Flag** button.

We will now start the study!

Start (4/4)

Figure S29. An example workflow for Human Only (control) participants [phase III: write supportive, empathic responses]. **(a)** Participant is asked to write a supportive, empathic response. **(b)** Participant starts writing the response.

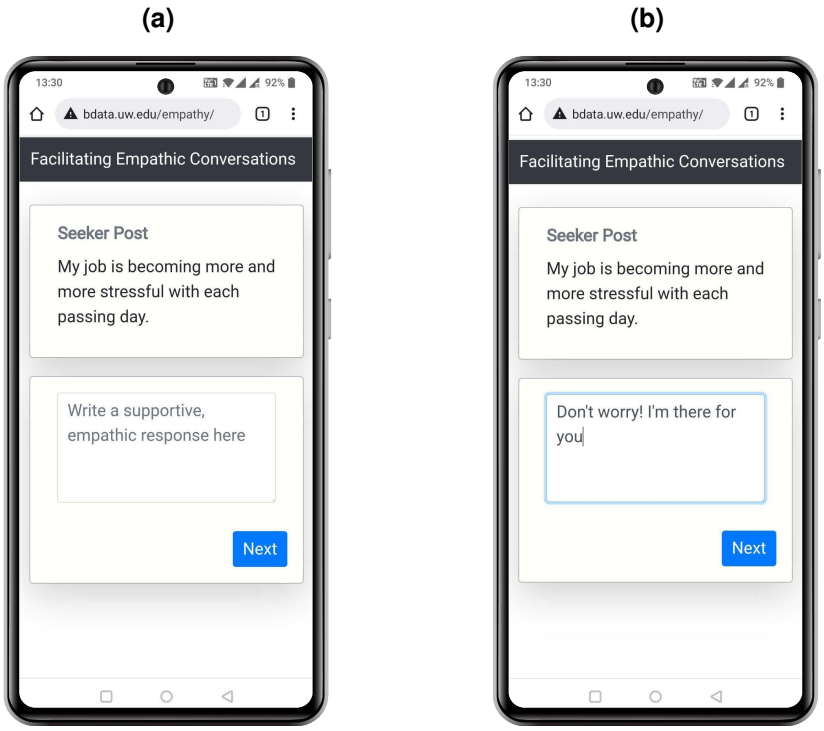


Figure S30. An example workflow for Human + AI (treatment) participants [phase III: write supportive, empathic responses]. **(a)** Participant is asked to write a supportive, empathic response and given an option to receive feedback. **(b)** Participant starts writing the response. **(c)** Participant clicks on the prompt to request feedback from HAILEY. **(d)** Participant accepts the suggested changes and gets an option to request more feedback. **(e)** Participant continues editing the response and requests more feedback as needed. **(f)** When the response is already highly empathic, the participant simply receives a positive feedback.

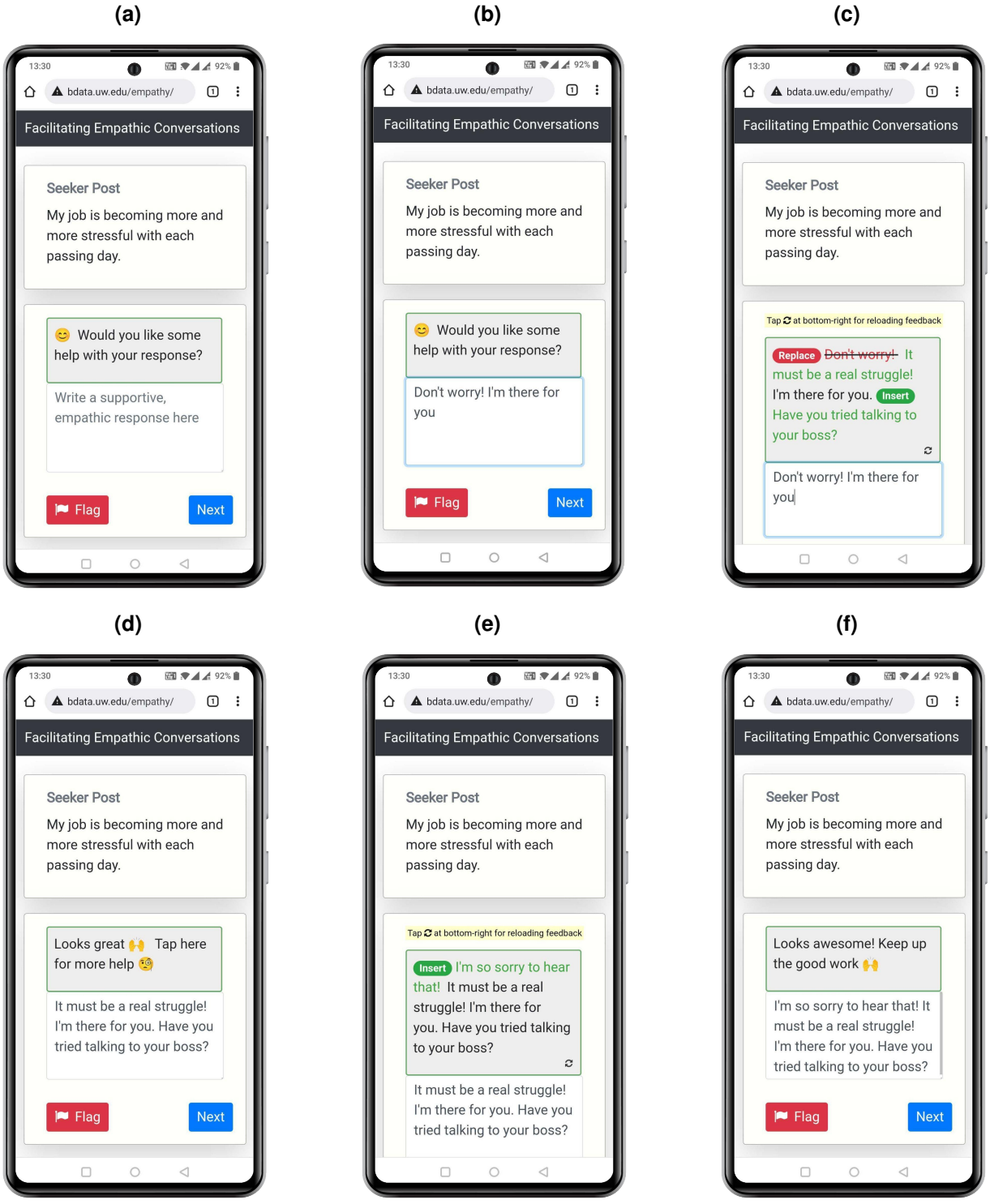


Figure S31. Interface for flagging feedback [phase III: write supportive, empathic responses].

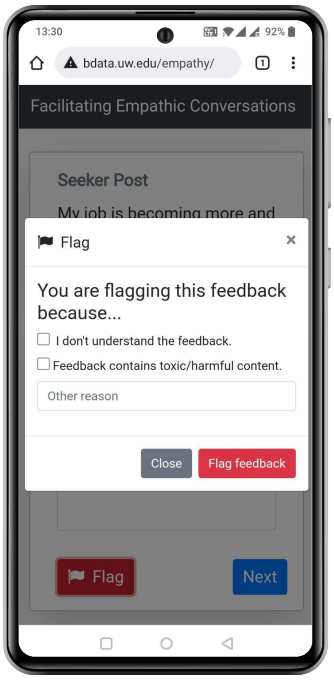


Figure S32. Exit survey used for collecting perceptions of control group participants [phase IV: post-intervention survey].

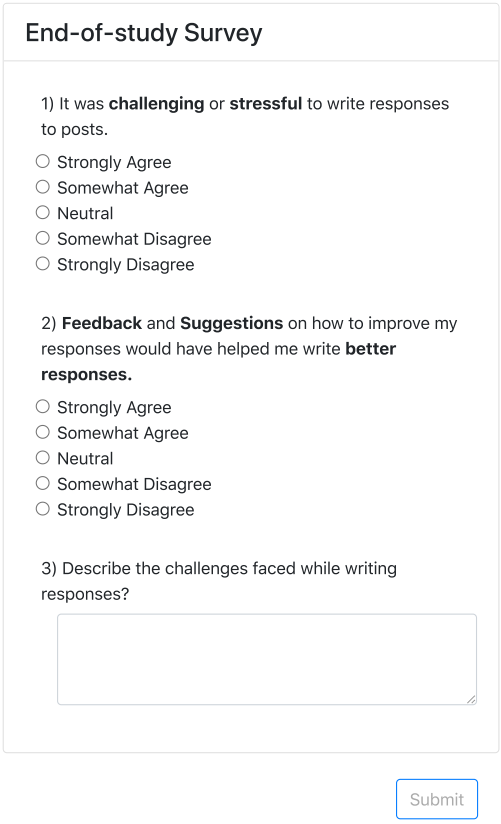
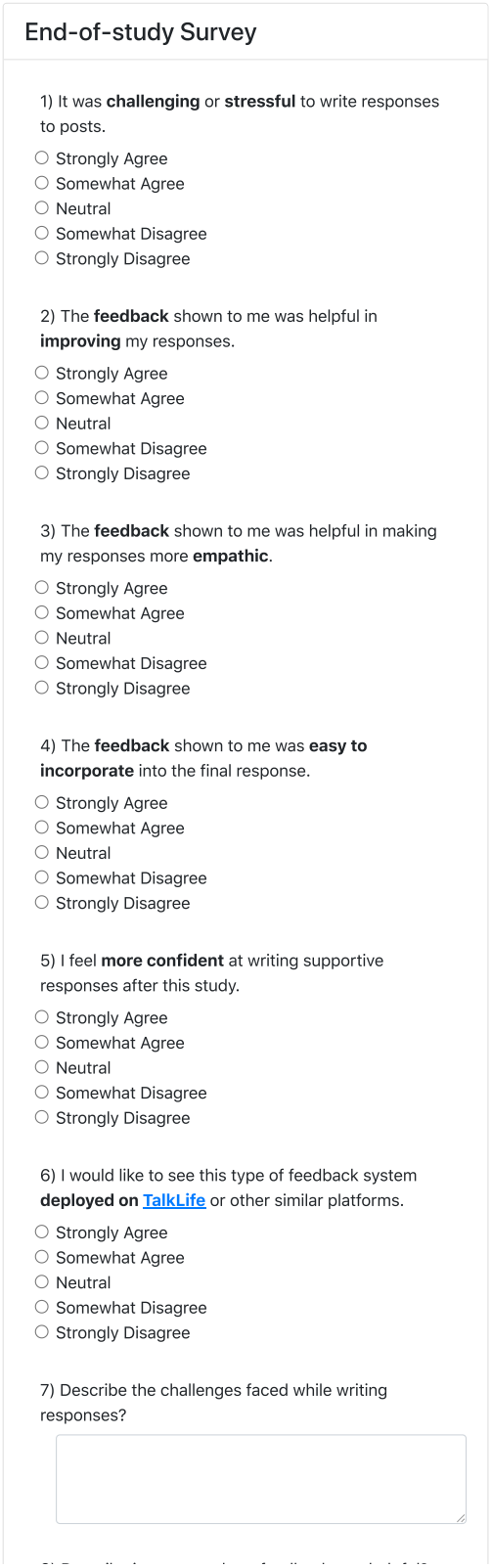


Figure S33. Exit survey used for collecting perceptions of treatment group participants [phase IV: post-intervention survey]. Continued on the next page (1/2).



End-of-study Survey

1) It was **challenging** or **stressful** to write responses to posts.

- ○ Strongly Agree
- ○ Somewhat Agree
- ○ Neutral
- ○ Somewhat Disagree
- ○ Strongly Disagree

2) The **feedback** shown to me was helpful in **improving** my responses.

- ○ Strongly Agree
- ○ Somewhat Agree
- ○ Neutral
- ○ Somewhat Disagree
- ○ Strongly Disagree

3) The **feedback** shown to me was helpful in making my responses more **empathic**.

- ○ Strongly Agree
- ○ Somewhat Agree
- ○ Neutral
- ○ Somewhat Disagree
- ○ Strongly Disagree

4) The **feedback** shown to me was **easy to incorporate** into the final response.

- ○ Strongly Agree
- ○ Somewhat Agree
- ○ Neutral
- ○ Somewhat Disagree
- ○ Strongly Disagree

5) I feel **more confident** at writing supportive responses after this study.

- ○ Strongly Agree
- ○ Somewhat Agree
- ○ Neutral
- ○ Somewhat Disagree
- ○ Strongly Disagree

6) I would like to see this type of feedback system **deployed on** TalkLife or other similar platforms.

- ○ Strongly Agree
- ○ Somewhat Agree
- ○ Neutral
- ○ Somewhat Disagree
- ○ Strongly Disagree

7) Describe the challenges faced while writing responses?

Figure S34. Exit survey used for collecting perceptions of treatment group participants [phase IV: post-intervention survey] (2/2).

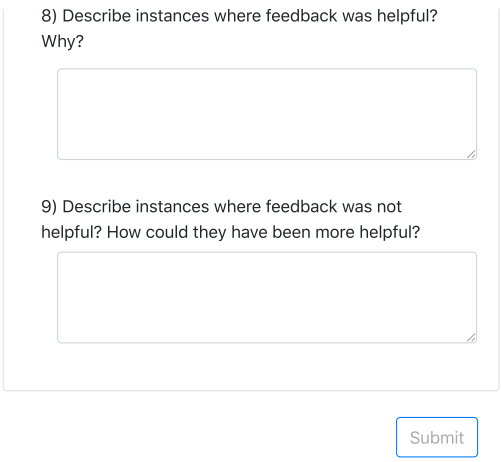


Figure S35. Consent form used for human evaluation of responses.

Disclaimer

Thank you for your interest in our study!

- We are **researchers at the University of Washington**, studying peer-to-peer support platforms.
- This study is **not being conducted by TalkLife**. TalkLife is not responsible for any risks/benefits associated with this study.
- As part of this study, we will be collecting **ratings to online mental health support interactions**. The data from this study will be uploaded to a **secure platform accessible only to the research team**. All collected data will be exclusively used for research purposes. Also, data will only be analyzed in aggregate.
- This study has been determined to be exempt from IRB approval under University of Washington IRB ID STUDY00012706. For concerns or questions, please contact ashshar@cs.washington.edu.
- Your participation in this study is **completely voluntary**. You are free to release/quit the study at any time. Refusing to be in the experiment or stopping participation will involve no penalty or loss of benefits to which you are otherwise entitled.
- Poor-quality data (e.g., same answers to all posts) **may be removed** without compensation.
- If you are a US citizen or a permanent US resident, you will receive an **Amazon gift card worth 5 USD** after completing this study.
  - ○ **Yes**, I'm a US citizen or a permanent US resident.
  - ○ **No**, I'm neither a US citizen nor a permanent US resident.
  - ○ Prefer not to say.
- **Note: Top-2 participants** (based on inter-rater agreement) will receive an **additional gift card worth 25 USD.**

Consent

☐ I agree to participate in this study. I also understand that only US citizens or permanent US residents can be compensated.

Accept and Continue

Figure S36. Instructions for human evaluation of responses.

Instructions

Read and evaluate online mental health support interactions.

Content Warning

The study contains posts including but not limited to self-harm and suicidal ideation, which may be disturbing to you. If you have concerns or questions, please send us an email (ashshar@cs.washington.edu). If you have strong negative reactions to some of the content, please reach out at crisis text line.

What will you read?

You will be shown three posts to read - a seeker post and two response posts. Details below:

- **Seeker Post:** This would be typically a mental health support seeking post, posted online by a user in distress. The writer of this post is called *Seeker.*
- **Response Post A:** This is a response/reply posted in response to the seeker post, usually in an attempt to provide mental health support to the seeker.
- **Response Post B:** This is *another* response/reply posted in response to the *same* seeker post, usually in an attempt to provide mental health support to the seeker.

What will you do?

Your task will be to evaluate the two response posts A and B. In particular, for each set of posts, you will answer the following questions:

- Which response is more **empathic** (regardless of appropriateness)?
  - Response A
  - Both are similar
  - Response B
- Which response is more **appropriate/relevant** to the seeker post (regardless of empathy)?
  - Response A
  - Both are similar
  - Response B
- Response A is...
  - Written by a human
  - Generated by a computer
  - Combination of both
- Response B is...
  - Written by a human
  - Generated by a computer
  - Combination of both

You will read **30** such posts and answer the associated questions. **We discourage the use of the "*Both are similar*" option. Only use it when the posts are actually similar and there is nothing to distinguish the two.**

Next

Figure S37. Interface for human evaluation of responses.

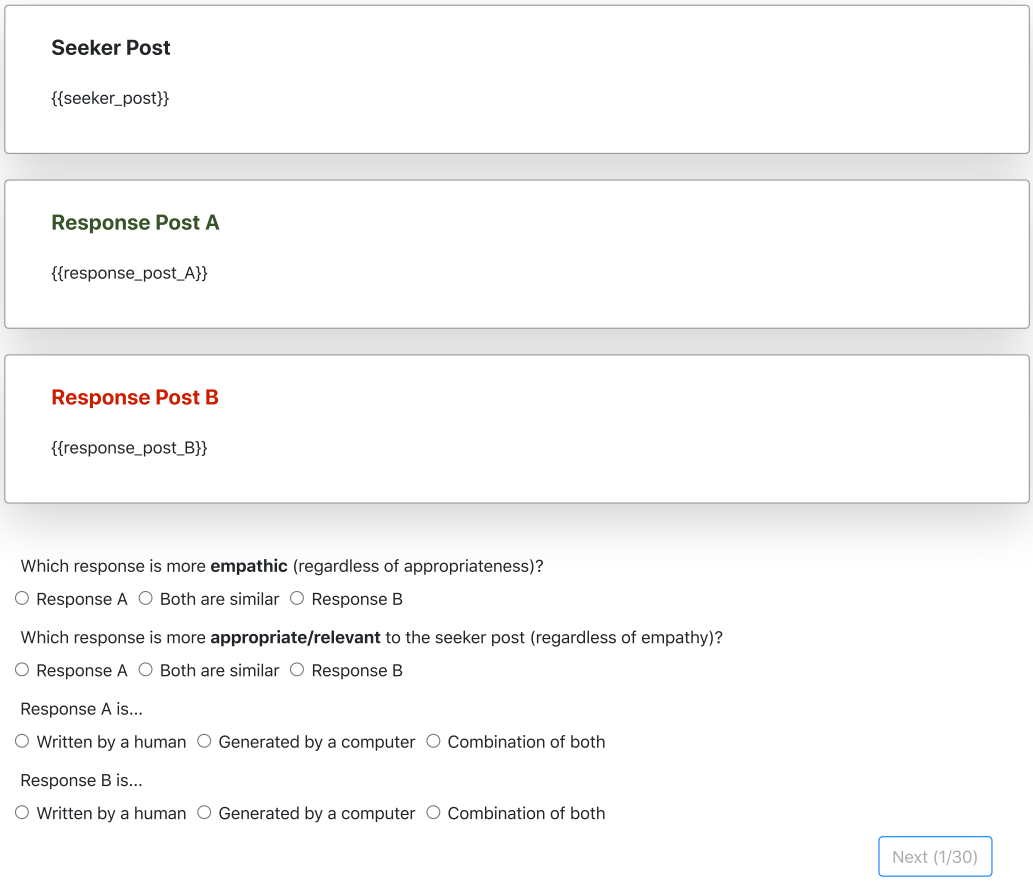


Figure S38. An overview of PARTNER, the deep reinforcement learning model that HAILEY uses. Figure adapted from Sharma et al.[47]