AI Against Anxiety: Efficacy, Mechanisms, and Demand for Voice-Based AI Therapy

Bennet Feld[1] & Nicolas Roever[2]

[1]London School of Economics and Political Science, London, UK
[2]University of Cologne, Cologne, Germany

Abstract

Cognitive Behavioral Therapy (CBT) is among the most effective treatments for depression and anxiety, yet its scalability is limited by cost, availability, and stigma—constraints which artificial intelligence (AI) may help overcome. We evaluate the efficacy of interactive CBT fully administered by a voice-based AI and its downstream effects on demand for traditional mental healthcare in a large-scale randomized controlled trial. First, findings support that AI can effectively deliver CBT. Treated participants show large reductions in symptoms of depression and anxiety relative to the control group. Second, treated participants reduce their demand for human therapy, reporting less fear of judgment and greater comfort disclosing to AI than to a human therapist. Thus, while expanding access to treatment, AI may reshape how patients seek and engage with mental health care.

Depression and anxiety disorders are widespread, ranking among the top 10 global causes of disease burden and affecting approximately 15% of the population (*1*). Yet despite the availability of highly effective treatments through pharmacology and psychotherapy for decades (*2–4*), few receive adequate care. Large treatment gaps persist in wealthy, middle-, and low-income settings alike (*5*). For psychotherapy in particular, structural constraints such as high treatment costs and limited therapist availability make closing this gap challenging (*6, 7*). Even cognitive behavioral therapy (CBT), a comparatively brief intervention with relatively well-established efficacy, remains resource-intensive, typically requiring a minimum of 8–12 sessions with a trained therapist (*2*).

Recent advances in generative artificial intelligence (AI) have sparked a debate about whether AI could help alleviate these constraints and narrow the treatment gap. Emerging applications use conversational AI agents to emulate talking therapies such as CBT. These applications leverage the proficiency of modern AI models in emulating human-like interactions (*8, 9*), can be scaled at low cost, and are more readily available than human providers. This raises two immediate questions: (1) Can AI systems successfully apply a treatment protocol like CBT originally designed for human delivery? (2) Which mechanisms can AI systems engage?

Currently, evidence on AI systems' capacity to deliver therapy is limited. First meta-analyses of simpler, reactive AI-based chatbots to promote mental wellbeing are rather promising, documenting improvements in distress (*10, 11*). However, these reviews emphasize that how AI mental health tools affect mental health is still a black box. Specifically, it has remained an open question whether AI systems can engage the same cognitive and behavioral changes targeted by CBT or whether improvements in symptoms are purely due to nonspecific factors such as having an attentive listener. Relatedly, a fundamental objection to the use of AI for CBT remains: AI systems lack genuine empathy, a quality hypothesized to be essential for therapeutic success (*12*).

Moreover, if AI could provide effective treatment and offer a credible clinical alternative, a third question arises: (3) Who demands AI therapy? Closing the treatment gap depends not only on technological feasibility but also on patient uptake. Individuals vary in how comfortable they are using technology in general and some may prefer human-guided therapy, particularly for emotionally challenging discussions (*13*). At the same time, AI-based delivery may appeal to other individuals by mitigating well-documented psychological barriers to seeking treatment, such as stigma, the fear of negative evaluation, and discomfort disclosing to healthcare professionals (*14–16*). If users view AI therapy as equivalent or superior to conventional treatments, adoption could fundamentally alter demand for other mental health services and reshape patient care pathways. Understanding preferences for AI versus human providers is therefore critical for assessing implications for the broader mental healthcare system.

In this study, we jointly examine the three questions of (1) efficacy, (2) mechanisms, and (3) demand in a large-scale ($N = 400$) randomized controlled trial (RCT) of voice-based, AI-led CBT. Importantly, the voice-based format preserves the conversational structure of human-led *talking* therapy, making this technical implementation format a promising delivery candidate. Participants are randomly assigned either to a control group, who receive psychoeducational material, or to the treatment group, who receive access to use a commercial, voice-interactive AI therapy application.[1]

[1]The application, *Sonia*, is currently available on the US App Store; for more information, see `https://www.soniahealth.com/`.

AI-Led CBT Intervention

Therapeutic Content

AI therapy sessions in our intervention are modeled after standard textbook CBT techniques and draw on the five main CBT session building blocks (*17–19*): (1) mood check-in, (2) collaborative agenda setting, (3) review of between-session practice, (4) introduction and rehearsal of core CBT techniques, and (5) assignment of homework. Users can start sessions anytime and personalize their AI therapist by selecting among different voices and interaction styles (e.g., more or less direct, slower or faster paced; see Figures S1a, S1b). The session interface is minimalistic and centers interaction on voice-based conversation (see Figure S1c). We provide more context on the therapeutic approach as well as a transcript of a full session in the Supplementary Text.

Technical Implementation

The system architecture is organized as a set of different large language models (LLMs) each responsible for key components of the CBT process (Figure 1A).

A *Memory System* manages information flow between models. For each user, the system stores all conversations as well as higher-level reflections such as a "case conceptualization", a structured summary of the individual's key therapeutic information (e.g., core beliefs, frequent automatic thoughts). During each session, the system retrieves and updates this information to preserve therapeutic continuity and deliver personalized interventions grounded in an evolving case formulation. Data storage complies with HIPAA, a U.S. law that sets standards for the privacy and security of health information.

Specialized LLMs that are part of the *Planning Module* handle specifics during the delivery of a session, i.e., deciding whether to move on to the next step in the session structure or which CBT technique is appropriate (e.g., proposing a behavioral experiment, making a balanced evidence log).

The *Safety Agent* operates as an independent LLM pipeline that runs parallel to the conversation. It continuously monitors user inputs for a variety of risk factors, including but not limited to suicide-related cues. When an acute crisis is detected, the session immediately shifts focus to a dedicated protocol. Depending on the severity of the crisis, users are encouraged to seek immediate human support, provided with crisis resources, and may have access to new sessions restricted. The thresholds are conservative, meaning the Safety Agent is "erring on the side of caution" under ambiguity.[2]

Experimental Design

Our experimental procedure was pre-registered on asPredicted #258,203 and approved by the ethics committee of the University of Cologne.[3] Figure 1B shows the basic structure of the experiment. We briefly describe the experiment here; for more detail, see the Materials and Methods section in the Supplementary Materials. Participants were recruited via targeted advertisements on social media (Facebook and Instagram) and directed to a baseline questionnaire. Our main eligibility criterion was the Generalized Anxiety Disorder-7 (GAD-7) scale (*21*), restricting the sample to individuals who self-report moderate to high anxiety symptoms.[4]

Following a confirmatory phone screening by research staff, $N = 400$ eligible participants were randomized with equal probability to the treatment or control group.[5] The treatment group was instructed to download and use the AI therapy app. Participants in the control arm were notified that they would receive the AI therapy app after 4 weeks; until then, they were given access to a web-based psychoeducation app based on World Health Organization mental health material (*22*).[6] In the control arm, 97.75% of participants registered for the educational web app. We selected this control condition over a standard waitlist for two reasons.

First, the control materials are drawn from a high-quality source and have been used in multiple peer-reviewed trials reporting reductions in stress and anxiety symptoms (*23, 24*). They consist of simple, widely used techniques (e.g., conscious breathing), making them a plausible counterfactual; psychological information is commonly delivered in text-based formats. Second, while actual engagement with the content was limited (see Figure S3a and Figure S3b), framing the web app as the provision of helpful WHO information likely created some expectations of mental health improvement. This implies the presence of experimenter demand even in the control group, strengthening our ability to isolate the treatment effect.

Main Treatment Effects on Anxiety and Depression Symptoms

Figure 2 shows average GAD-7 and PHQ-8 scores (measuring depression levels; (*26*)) in the treatment and control group, as well as intention-to-treat treatment effects from our pre-registered specification. At two weeks (our pre-registered main outcome), participants assigned to AI therapy show a -3.52-point lower GAD-7 score relative to the control group (95% CI: $[-4.42, -2.62]$). Expressed as a standardized mean difference (SMD), this corresponds to SMD $= -0.63$ (95% CI: $[-0.8, -0.47]$). The estimated treatment effect increases to -4.47 GAD-7 points (95% CI: $[-5.4, -3.54]$) at our four-week follow-up. The treatment group demonstrated substantial symptom reduction, with average GAD-7 scores dropping from 15.9 at baseline to 8.9 after two weeks—a 7-point decrease that far exceeds the 4-point threshold for clinically meaningful change established in the literature (*27*). Because we standardized GAD-7 scores using the full baseline distribution of participants who filled out the baseline survey, these effect sizes reflect expected benefits for the broader population

[2] For a more detailed description and validation of the safety algorithm, see Ref. (*20*).

[3] Link to preregistration: `https://aspredicted.org/sv9ac7.pdf`.

[4] Additional details and a thorough discussion on sample selection and baseline characteristics are provided in the Supplementary Text.

[5] For balance of baseline covariates, see Table S1.

[6] All participants received the same remuneration with no additional incentive for app usage in either group.

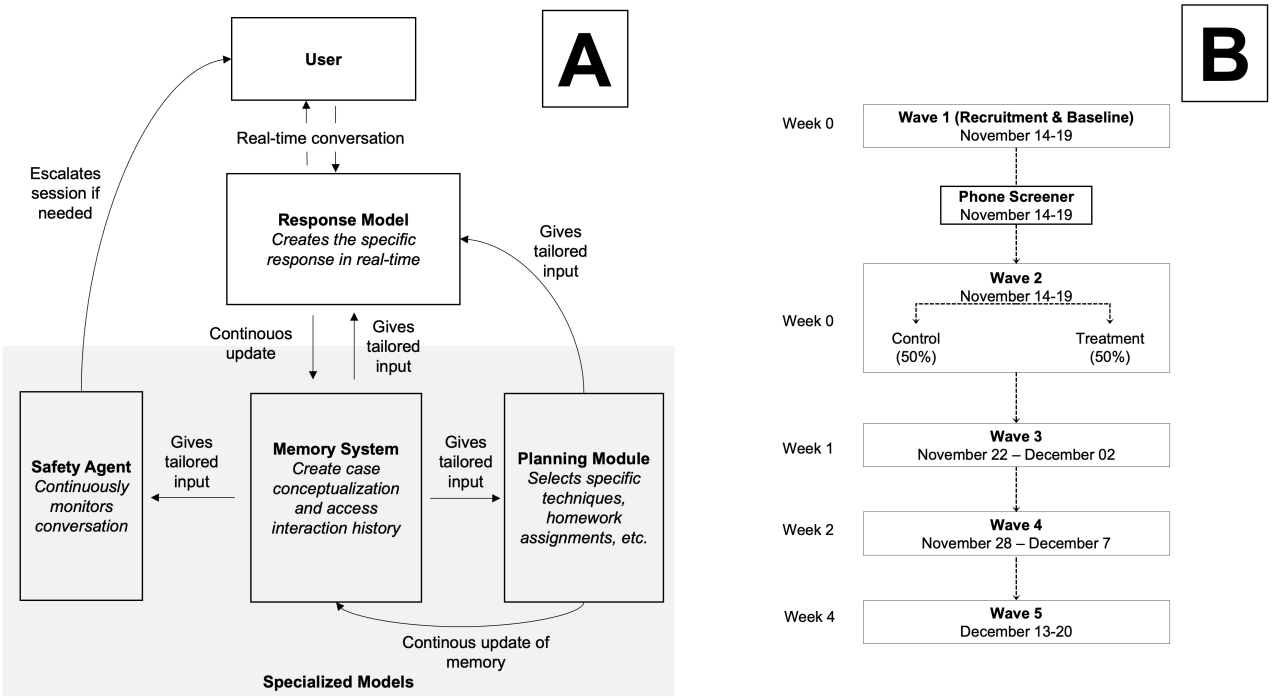


Figure 1: Schematic overview of AI-therapy application architecture and experimental design. (A) This figure shows a schematic overview of the AI therapist's architecture. A response model handles the real-time conversation and interacts with specialized models that intervene in case of safety risks, manage memory, and provide specialized input (e.g., advancing the session structure). (B) This flowchart shows the structure of the RCT conducted in winter 2025.

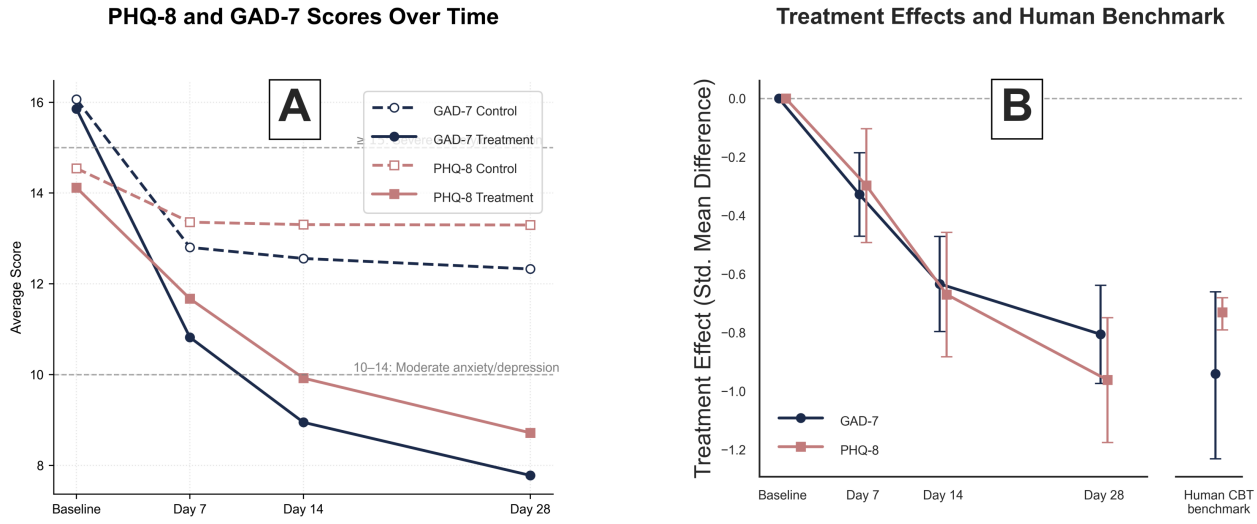


Figure 2: Average development of GAD-7 and PHQ-8 scores in the treatment and control groups and estimated treatment effects over time. Baseline GAD-7 is elicited during the first survey, and baseline PHQ-8 is elicited during the second survey before treatment assignment. (A) shows mean GAD-7 and PHQ-8 trajectories by arm. (B) reports standardized treatment effects at each wave with 95% confidence intervals estimated using our pre-specified ANCOVA model. Standardization is done using the full baseline survey sample for GAD-7 ($N = 10,490$) and using the full randomized sample for PHQ-8 ($N = 400$). The human CBT benchmark is derived from RCTs broadly comparable to our trial conducted with human therapists from the *MetaPsy* database (*25*) (for the exact selection criteria, see Figure S6 in the Supplementary Materials).

of individuals seeking AI-based mental health support. We additionally benchmark our trial against comparable human therapist RCTs from the *MetaPsy* database (*25*) (for selection criteria, see Figure S6). Figure 2B shows that our results at four weeks fall within the range observed in human therapist trials.[7] Furthermore, Supplementary text B.7 shows that treatment effects increase for participants with more sessions completed.

We document several robustness checks in Supplementary Text B.2. Most notably, attrition by the 2-week follow-up is only 2.5% (10 participants) and similarly low at our 4-week follow-up (also 2.5%). The resulting Lee bounds lie close to our main estimates, which alleviates concerns about attrition causing bias. We assess social desirability bias and experimenter demand effects across treatment arms by measuring self- and social-image concerns (*28, 29*). Mann-Whitney U tests show no significant differences between the two treatment arms in self- and social-image concern levels ($p > 0.10$), suggesting that *outcome* differences across arms are not strongly contaminated by these biases.

User Safety

Users in our trial perceive the application to be safe: after two weeks of use, 97% of users reported no safety concerns about the AI therapist relative to traditional chatbots, whereas 3% expressed concerns—two citing privacy risks inherent in AI, two reporting uncertainty in their assessment, and one describing the AI as insufficiently knowledgeable. Throughout the trial, the app's *Safety Agent* determined that four users were having problems not suited for AI help. In those cases, users could not start new conversations with the AI and were pointed to emergency hotlines and message services instead.[8]

Cognitive and Behavioral Mechanisms

To address the limited understanding of mechanisms underlying AI mental health interventions (*10, 12*), we next analyze our comprehensive mechanism data.

First, we consider engagement. Trial participants were compensated exclusively for survey completion, with no explicit incentives for app use. Despite this, engagement was high relative to other online interventions (*30, 31*). Over the 4-week trial period, mean cumulative conversation time with the AI therapist was 3.14 hours (interquartile range: 0.50–4.37), which is equal to three to four standard 53-minute psychotherapy sessions, i.e., the standard conversation time a patient would have over a month with a human therapist.

Second, we examine effects on established constructs targeted by CBT (*32*). Figure 3B shows significant reductions in the frequency of negative automatic thoughts measured by the ATQ-8 at all follow-ups (*33*). We also find significant improvements in two common cognitive biases—negativity bias (expecting adverse outcomes) and selective recall of negative events. In addition, the AI therapist supports behavioral activation; across all follow-ups, treated participants report engaging in more enjoyable and meaningful activities.[9] Downstream social outcomes move in the same direction: treated participants report more frequent social contact (Figure S8) and consistently lower perceived loneliness on the UCLA Loneliness Scale (Figure S9, (*34*)), consistent with changes in mental health operating through behavioral adjustments.

Therapeutic Alliance

Therapeutic alliance—a relationship between therapist and patient characterized by agreement on goals and tasks and an affective bond—has been shown to have a robust, moderate correlation with CBT outcomes (*35–37*).

We assessed alliance with the AI therapist using the Working Alliance Inventory–Short Revised (WAI-SR; (*38*)), a widely used measure in psychotherapy research, and benchmarked results against the inpatient face-to-face sample in Ref. (*39*). Participants reported generally high alliance levels with the AI therapist: The mean total score in our treated sample was 3.75 (benchmark: 3.6). Subscale means were 3.69 for tasks (benchmark: 3.4), 3.81 for goals (benchmark: 3.6), and 3.75 for bond (benchmark: 3.7).[10]

Thus, the AI therapist was perceived to form a relatively strong therapeutic alliance across all three core dimensions. This finding is conceptually interesting: Prior research has established that individuals value empathic communication from an AI less than from a human (*40–42*). Combined with the AI's successful reproduction of CBT's therapeutic effects and mechanisms in our setting, this suggests that CBT efficacy may not fully depend on an empathetic human mind delivering the intervention. Instead, therapeutic benefits may emerge from perceived collaboration and goal alignment between patient and therapist (*43*)—challenging the view that authentic human empathy, while potentially beneficial, is indispensable for effective CBT delivery (*12*).

Preferences for AI Therapy

We next assess demand. Does AI therapy displace other forms of psychological support and treatment, and which population does it serve?

Two mechanisms may plausibly drive effects of the AI therapy treatment on the demand for psychiatric medication and human therapy. The first is mechanical: effective mental health treatment reduces the need for further treatment. The second is substitution: participants in the treatment group have now learned about and experienced AI psychotherapy, which may shift their preferences away from alternatives. This substitution effect reduces the demand for other forms of treatment if participants either had not previously considered AI psychotherapy

[7] Figure S7 indicates that our findings are also consistent with prior studies of AI-based mental health interventions.

[8] Participants of our study were additionally referred to a clinician.

[9] For the survey items we use to measure mechanisms, see the Materials and Methods section in the Supplementary Materials.

[10] Higher WAI-SR scores in our AI setting correlated with greater symptom reduction (Figure S10), replicating therapeutic alliance's role as a transdiagnostic common factor in CBT (*39*).

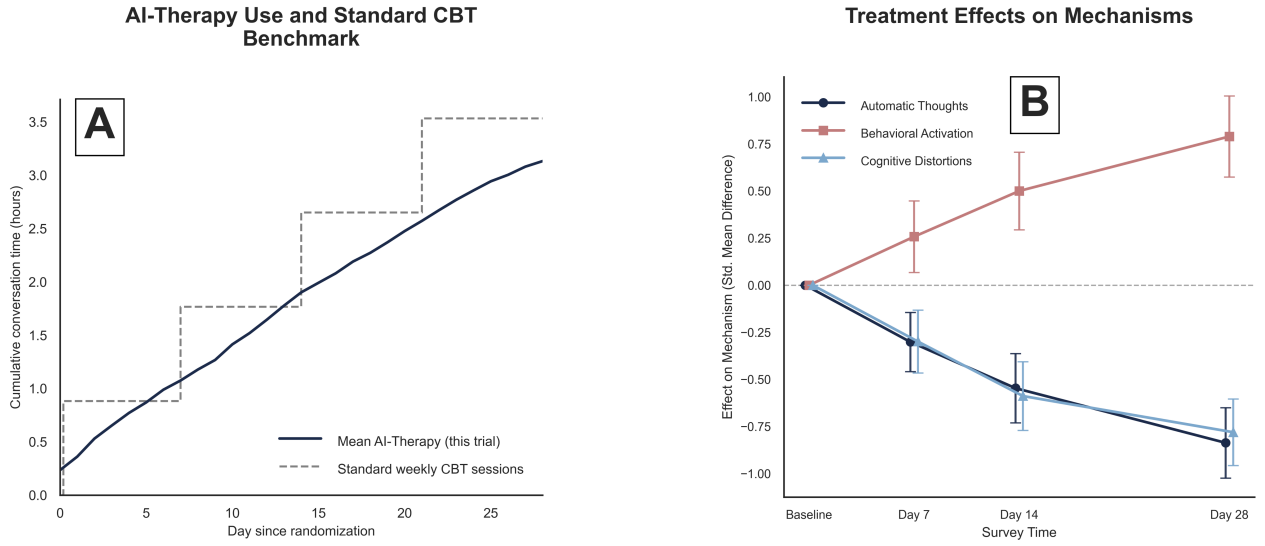


Figure 3: Engagement with the AI therapist and treatment effects on mechanism items over time. (A) shows the mean cumulative conversation time for each day since randomization as well as the standard benchmark for human CBT of one 53-minute session each week. (B) reports standardized treatment effects at each wave with 95% confidence intervals estimated using our pre-specified ANCOVA model. Standardization is done using the baseline data before treatment assignment from the full randomized sample.

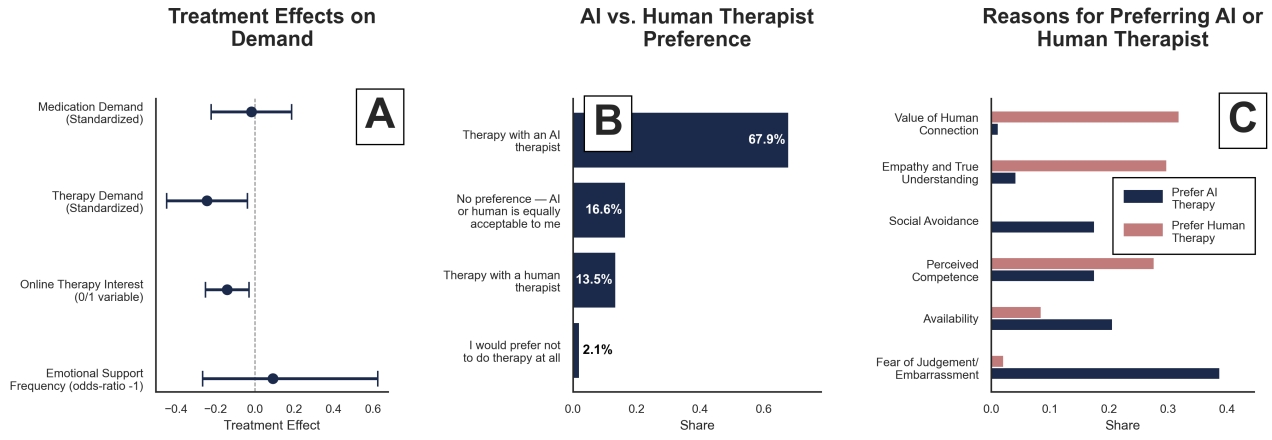


Figure 4: Effects of treatment on demand for human therapy/medication and reasons for preferring AI or human-led therapy. (A) The figure reports intent-to-treat effects of assignment to AI therapy on four outcomes measured at endline (day 28). (1) demand for psychiatric medication and (2) demand for human-led therapy, elicited on 0–100 scales and standardized relative to the control-group distribution. Estimates are from OLS regressions; error bars denote 95% confidence intervals based on cluster-robust standard errors. (3) Willingness to engage in human-led online therapy is measured as a yes/no outcome and estimated using a linear probability model (LPM) with the same specification. (4) Turning to friends for emotional support is measured on a 5-point Likert scale; effects are estimated via ordered logit, plotted as (odds ratio−1). Control variables for all four specifications include prior therapy experience, baseline perceived effectiveness of human therapy, and a mental-health index (mean of standardized PHQ-8 and GAD-7) at endline. (B) shows answers given by treated participants to a single-choice question ("What do you prefer?") at the four week follow-up. (C) shows reasons given by treated participants in an open-text question for why they prefer AI versus human therapy, split by whether the overall answers indicate a preference for human or AI therapy or indifference. Categorization is done using OpenAI's `GPT-5-mini` model. The categories are obtained using the workflow outlined in Section B.6. We only show themes that occurred at least 15 times overall in this figure.

or came to value it more over the course of the trial.

Figure 4A shows the effect of treatment assignment on demand for psychotherapy and pharmacotherapy net of the first, mechanical mechanism by controlling for contemporaneous mental health. While there is no crowding out of psychiatric medication, we find evidence that AI psychotherapy reduces demand for human therapy in our sample through the substitution channel. A reduction in demand for human-led CBT is also reflected in participants' responses to a behavioral survey measure in the week four follow-up: participants were asked whether they wished to receive information via email about online, human-led CBT ("Interest online therapy"). This item carries actual consequences, and is thus plausibly less affected by experimenter demand and social desirability bias (*28, 29*). The effect size on the behavioral item is large: treated participants are 14% less likely to demand information about teletherapy than participants in the control group (p-value $= 0.01$). Consistent with the interpretation that AI therapy is perceived as a substitute precisely for human therapy, we find no reduction of how often people seek emotional support from friends in the treatment group relative to the control group (item "Emotional Support Frequency" in Figure 4A).

What drives these preferences, and for whom does AI present as an alternative to human therapists? Figure 4B shows responses to a single-choice item with four options elicited at endline. Among treated participants, 68% prefer AI-delivered therapy, 14% prefer human delivery, 17% express no preference, and 2% report they do not feel the need for therapy right now. To understand the reason for the surprisingly high share of participants preferring an AI therapist, we used an open-ended question format (*44*). Treated participants are asked whether they prefer AI-delivered or human-delivered therapy and to explain their reasoning. Using unsupervised machine learning, we classified responses into eight distinct thematic categories and report our findings in Figure 4C.[11] Among users who prefer AI, the most frequently cited reasons relate to social anxiety and fear of judgment (e.g., "I feel AI is better in my opinion because I won't have to worry about being judged by a person", "I would rather have the AI therapist because I would feel more comfortable opening up to it rather than a real person").

This finding contrasts with conventional assumptions about therapeutic delivery: for a substantial proportion of our sample, the absence of a human represents an advantage rather than a limitation. While it is often argued that the absence of the human factor and the inability of AI systems to express genuine empathy is a fundamental constraint from a clinical perspective (*12*), many participants in our study frame this feature positively and instead value the reduction in social-evaluative threat over the loss of human connection.[12]

[11] For more detailed information on the classification workflow, see Section B.6. All categories are given in Table S3.

[12] We provide additional evidence supporting this conclusion in Supplementary Text B.3.

Discussion

Three methodological considerations merit discussion and shape the interpretation of the study results: our control group design, our sample recruitment of early adopters, and the brief 4-week intervention period.

(1) Our control condition consisted of a text-based information provision through a web app. Engagement with the web app was significantly lower than with the treatment app (mean visits on web app: 2.1; average minutes spent: 12.4). The treatment effect thus reflects the combined influence of CBT content and engagement with a voice-based AI, which are both core intervention elements of talking therapy. However, we opted for transparent group-assignment disclosure to avoid deception; participants were told at randomization if they were assigned to treatment or control. Treatment-group participants' awareness of their status may generate placebo effects. While this study presents several robustness checks, including tests for experimenter demand, and a positive correlation between engagement and treatment effects suggesting a dose-response relationship, these tests cannot fully rule out such concerns. Thus, we cautiously interpret effect sizes as upper-bound estimates.

(2) Another key design decision was to set the study duration to four weeks, reflecting usage patterns observed among native users of the AI therapy app. Users predominantly engage with the application during periods of acute crisis, followed by attrition as symptoms abate, returning upon the onset of a new crisis. By restricting admission to individuals with current high anxiety levels, the study replicates this "acute phase" usage model. Accordingly, the study focuses on short-term outcomes and does not address longer-term treatment durability or the AI's role in managing symptom recurrence. Given that relapse rates following successful CBT for anxiety disorders are estimated at approximately 14% (*45*), additional longitudinal research is needed to assess whether AI-delivered interventions can sustain remission over time or effectively re-engage users during subsequent episodes.

(3) The sample for this trial was recruited via online advertisements and screened for inclusion criteria favoring positive attitudes toward AI, implying that our participants represent "early adopters" of specialized AI psychotherapy. However, two lines of evidence suggest our findings extend to a broader population. First, adoption of LLMs for emotional and psychological support has become widespread. In a large survey of 21,000 individuals worldwide, 48% report that they have used generative AI tools such as ChatGPT or Gemini in 2024 (*46*). Common applications involve information and advice seeking, including for mental health advice (*47, 48*). Second, we conducted a supplementary survey of $N = 300$ U.S. adults representative by age, gender, and ethnicity to the general U.S. adult population and found broad willingness to engage with AI psychotherapy. The sample was recruited via Prolific, a survey provider commonly used in the social sciences (*49*). Compared to the RCT population in our trial, the survey sample holds generally less favorable views about AI. Nevertheless, when asked whether they would be open to using a specialized

AI for emotional support, 45% agreed, indicating substantial baseline receptiveness to AI-based mental health tools in the general population.

Conclusion

The results of this RCT suggest that voice-based AI therapy can replicate the main treatment effects of human CBT in the short term. We document significant changes in all core CBT mechanisms—reducing automatic thoughts and cognitive distortions while increasing behavioral activation—implying that AI-delivered therapy engages therapeutic mechanisms as human-administered CBT. Our trial shows that users perceive AI therapy as a legitimate or even preferred alternative to human-delivered treatment and population demand data suggest receptivity to such interventions. These results have broad implications for global mental health care.

AI-delivered therapy offers substantial cost advantages. Participants in our trial engaged with the AI therapist for on average 3.14 hours over four weeks. At approximately $100 per hour for human therapy in the United States, equivalent care would cost roughly $314 per participant compared to approximately $3.14 for AI-based delivery.[13] AI therapists may thus provide a viable option for cost-constrained individuals and healthcare systems, with potential extension to subclinical populations. These groups remain largely unserved to date, despite evidence that CBT-based programs promote psychological well-being and work-effectiveness (*50, 51*).

Beyond these economic deliberations, AI therapy can mitigate psychological barriers to treatment. Many participants in our sample demand AI delivery precisely because it eliminates interpersonal judgment and creates less social anxiety. Thus, the mere availability of AI-led CBT may enable accessing populations that were previously difficult to reach through human-delivered care.

Methodologically, this study demonstrates that AI can deliver complex, longitudinal treatments requiring temporal continuity. AI's capacity for standardized delivery at scale enables systematic investigation of which CBT components benefit specific patient subgroups, creating opportunities to optimize therapeutic protocols (*52*).

While encouraging, our results raise important questions for the design of mental healthcare systems. We investigate an acute-use model of AI-delivered therapy that mirrors usage patterns observed in native users: we recruit individuals with high anxiety symptoms and follow their trajectories over a four-week period. Consequently, our study does not address the broader course of a patient's mental health journey, including relapse dynamics, longer-term management, or the integration of AI within stepped-care pathways. A key finding of our trial is that AI availability can crowd out demand for human therapy, implying that some individuals may engage less with mental health professionals due to the existence of AI systems. AI's capacity for mental healthcare, however, is inherently limited. Current AI systems are unlikely to be suited for complex cases that require relational depth, for ambiguous clinical decisions requiring ethical judgment, for rapid symptom escalation, or for suicidal patients. We acknowledge these limitations by design; the application we study refers such cases to human clinicians. This highlights the importance of considering system-level implications, as AI's potential can only be realized within care structures that recognize its limits and ensure effective transitions from AI-based to human care.

[13] The $1 per-session estimate reflects approximate current model costs and is intended for illustrative purposes; costs may decline further as technologies evolve.

Acknowledgments

We thank Nava Ashraf, Oriana Bandiera, Felix Chopra, Julian Herpertz, Thomas Messer, Margot Paul, Marten Ritterath, Chris Roth, Matan Rubin, Sebastian Tonke, the Applied-Micro Seminar at the University of Cologne, and the LSE public economics group for invaluable comments. We are grateful for fruitful cooperation and discussions about their product with Lukas Wolf and Chris Aeberli. We also thank Frauke Gerster, Malte van der Heiden and Moritz Lakenbrink for excellent research assistance.

Funding: B. F. was funded by the Economic and Social Research Council (ESRC); N. R. was funded by the German Academic Scholarship Foundation; participant recruitment and remuneration were partly funded by Bloit Inc., owner of the IP of the tool used in this study, and by the Center for Social and Economic Behavior (C-SEB) of the University of Cologne

Author contributions: B. F. & N. R. Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

Competing interests: The authors declare a competing interest: this study was partially funded and conducted in cooperation with Bloit Inc., the owner of the commercial app under evaluation. N.R. will start consulting work at Bloit Inc. after submission of this paper. See the Materials and Methods section in the Supplementary Materials for further information on how research independence was maintained.

Data and materials availability: Full data and code for this paper are available on GitHub, stored as a stable release at DOI: `10.5281/zenodo.18506142`

Supplementary materials

Materials and Methods
Supplementary Text
Figures S1 to S14
Tables S1 to S3
Preregistration
Survey Instructions

References

1. Institute for Health Metrics and Evaluation, Global Health Data (GHD), https://www.healthdata.org/research-analysis/health-topics/mental-health (2025), accessed: January 28, 2026.

2. P. Cuijpers, *et al.*, Cognitive behavior therapy vs. control conditions, other psychotherapies, pharmacotherapies and combined treatment for depression: a comprehensive meta-analysis including 409 trials with 52,702 patients. *World Psychiatry* **22** (1), 105–115 (2023), doi:https://doi.org/10.1002/wps.21069, https://onlinelibrary.wiley.com/doi/abs/10.1002/wps.21069.

3. S. G. Hofmann, A. Asnaani, I. J. J. Vonk, A. T. Sawyer, A. Fang, The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research* **36** (5), 427–440 (2012), doi:10.1007/s10608-012-9476-1.

4. A. Cipriani, *et al.*, Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: A systematic review and network meta-analysis. *The Lancet* **391** (10128), 1357–1366 (2018), doi:10.1016/S0140-6736(17)32802-7.

5. G. Thornicroft, *et al.*, Undertreatment of people with major depressive disorder in 21 countries. *The British Journal of Psychiatry* **210** (2), 119–124 (2017), doi:10.1192/bjp.bp.116.188078.

6. M. L. Wainberg, *et al.*, Challenges and Opportunities in Global Mental Health: a Research-to-Practice Perspective. *Current Psychiatry Reports* **19** (5), 28 (2017), doi:10.1007/s11920-017-0780-z.

7. J. R. Cummings, *et al.*, Challenges facing mental health systems arising from the COVID-19 pandemic: Evidence from 14 European and North American countries. *Health Policy* **136**, 104878 (2023), doi:10.1016/j.healthpol.2023.104878.

8. C. R. Jones, B. K. Bergen, People cannot distinguish GPT-4 from a human in a Turing test (2024), doi:10.48550/arXiv.2405.08007.

9. C. R. Jones, B. K. Bergen, Large Language Models Pass the Turing Test (2025).

10. H. Li, R. Zhang, Y.-C. Lee, R. E. Kraut, D. C. Mohr, Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine* **6** (1), 236 (2023), doi:10.1038/s41746-023-00979-5.

11. J. Li, *et al.*, Chatbot-Delivered Interventions for Improving Mental Health Among Young People: A Systematic Review and Meta-Analysis. *Worldviews on Evidence-Based Nursing* **22** (4), e70059 (2025), doi:10.1111/wvn.70059.

12. N. Opel, M. Breakspear, Transforming mental health research and care through artificial intelligence. *Science* **391** (6782), 249–258 (2026), doi:10.1126/science.adz9193, https://www.science.org/doi/abs/10.1126/science.adz9193.

13. G. Arbanas, *et al.*, Patients prefer human psychiatrists over chatbots: a cross-sectional study. *Croatian Medical Journal* **66** (1), 13–19 (2025), doi:10.3325/cmj.2025.66.13.

14. S. Clement, *et al.*, What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological Medicine* **45** (1), 11–27 (2015), doi:10.1017/S0033291714000129.

15. C. J. Cronin, M. P. Forsstrom, N. W. Papageorge, What Good Are Treatment Effects Without Treatment? Mental Health and the Reluctance to Use Talk Therapy. *Review of Economic Studies* **92** (3), 1699–1737 (2025), doi:10.1093/restud/rdae061.

16. C. Roth, P. Schwardmann, E. Tripodi, *Depression Stigma*, Discussion Paper DP18919, Centre for Economic Policy Research (2024).

17. D. D. Burns, *Feeling Good: The New Mood Therapy* (William Morrow) (1980).

18. J. S. Beck, *Cognitive Behavior Therapy: Basics and Beyond* (The Guilford Press, New York), third ed. (2021).

19. M. G. Craske, *Cognitive-Behavioral Therapy* (Oxford University Press, Oxford) (2016).

20. S. Weber, *et al.*, Suicide- and crisis-risk detection using large language models in mental-health chatbots. *medRxiv* (2026), doi:10.64898/2026.01.12.26343914, https://www.medrxiv.org/content/early/2026/01/15/2026.01.12.26343914.

21. R. L. Spitzer, K. Kroenke, J. B. W. Williams, B. Löwe, A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine* **166** (10), 1092–1097 (2006), doi:10.1001/archinte.166.10.1092.

22. World Health Organization, *Doing what matters in times of stress: an illustrated guide* (World Health Organization, Geneva) (2020), https://www.who.int/publications/i/item/9789240003927, licence: CC BY-NC-SA 3.0 IGO.

23. C. Acarturk, G. Kurt, Z. Ilkkursun, E. Uygun, A. Karaoglan-Kahilogullari, "Doing What Matters in Times of Stress" to Decrease Psychological Distress During COVID-19: A Randomised Controlled Pilot Trial. *Intervention* **20** (2), 170–178 (2022), doi:10.4103/intv.intv_29_21.

24. P. Yabunga, *et al.*, Effectiveness of barber-facilitated "Doing What Matters in Times of Stress" intervention among urban literate youths in Western Kenya: A cluster randomised trial. *PLOS Global Public Health* **5** (6), e0004712 (2025), doi:10.1371/journal.pgph.0004712.

25. M. Harrer, A. A. Sprenger, P. Kuper, E. Karyotaki, P. Cuijpers, *metapsyData: Access the Meta-Analytic Psychotherapy Databases in R* (2022), `https://data.metapsy.org`.

26. K. Kroenke, *et al.*, The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders* **114** (1-3), 163–173 (2009), epub 2008 Aug 27, doi:10.1016/j.jad.2008.06.026.

27. A. Toussaint, *et al.*, Sensitivity to change and minimal clinically important difference of the 7-item Generalized Anxiety Disorder Questionnaire (GAD-7). *Journal of Affective Disorders* **265**, 395–401 (2020), doi:10.1016/j.jad.2020.01.032.

28. L. Bursztyn, I. K. Haaland, N. Röver, C. Roth, *The Social Desirability Atlas*, Tech. rep., National Bureau of Economic Research (2025).

29. J. De Quidt, J. Haushofer, C. Roth, Measuring and bounding experimenter demand. *American Economic Review* **108** (11), 3266–3302 (2018).

30. M. V. Heinz, *et al.*, Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI* **2** (4) (2025), published March 27, 2025. Copyright © 2025., doi:10.1056/AIoa2400802, `https://doi.org/10.1056/AIoa2400802`.

31. A. Shreekumar, P.-L. Vautrey, Managing Emotions: The Effects of Online Mindfulness Meditation on Mental Health and Economic Behavior (2024), unpublished manuscript, Massachusetts Institute of Technology.

32. A. Norbury, T. U. Hauser, S. M. Fleming, R. J. Dolan, Q. J. M. Huys, Different components of cognitive-behavioral therapy affect specific cognitive mechanisms. *Science Advances* **10** (13), eadk3222 (2024), doi:10.1126/sciadv.adk3222, `https://www.science.org/doi/abs/10.1126/sciadv.adk3222`.

33. R. G. Netemeyer, *et al.*, Psychometric properties of shortened versions of the Automatic Thoughts Questionnaire. *Educational and Psychological Measurement* **62** (1), 111–129 (2002), doi:10.1177/0013164402062001008.

34. M. E. Hughes, L. J. Waite, L. C. Hawkley, J. T. Cacioppo, A short scale for measuring loneliness in large surveys: Results from two population-based studies. *Research on Aging* **26** (6), 655–672 (2004), doi:10.1177/0164027504268574.

35. A. O. Horvath, A. Del Re, C. Flückiger, D. Symonds, Alliance in individual psychotherapy. *Psychotherapy* **48** (1), 9 (2011).

36. P. Cuijpers, M. Reijnders, M. J. Huibers, The role of common factors in psychotherapy outcomes. *Annual review of clinical psychology* **15** (1), 207–231 (2019).

37. R. Elliott, A. C. Bohart, J. C. Watson, D. Murphy, Therapist Empathy and Client Outcome: An Updated Meta-analysis. *Psychotherapy* **55** (4), 399–410 (2018), doi:10.1037/pst0000175.

38. R. L. Hatcher, J. A. Gillaspy, Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy Research* **16** (1), 12–25 (2006), doi:10.1080/10503300500352500.

39. T. Munder, F. Wilmers, R. Leonhart, H. W. Linster, J. Barth, Working Alliance Inventory-Short Revised (WAI-SR): Psychometric properties in outpatients and inpatients. *Clinical Psychology & Psychotherapy* **17** (3), 231–239 (2010), doi:10.1002/cpp.658.

40. M. Rubin, *et al.*, Comparing the value of perceived human versus AI-generated empathy. *Nature Human Behaviour* **9** (11), 2345–2359 (2025), doi:10.1038/s41562-025-02247-w.

41. F. Bernays, M. Henriques Pereira, J. Menges, How Human is AI? Examining the Impact of Emotional Prompts on Artificial and Human and Responsiveness (2026), `https://arxiv.org/abs/2601.05104`.

42. T. Kleinert, M. Waldschütz, J. Blau, M. Heinrichs, B. Schiller, AI outperforms humans in establishing interpersonal closeness in emotionally engaging interactions, but only when labelled as human. *Communications Psychology* (2026), doi:10.1038/s44271-025-00391-7.

43. S. Zilcha-Mano, Is the alliance really therapeutic? Revisiting this question in light of recent methodological advances. *American Psychologist* **72** (4), 311 (2017).

44. I. Haaland, C. Roth, S. Stantcheva, J. Wohlfart, Understanding economic behavior using open-ended survey data. *Journal of Economic Literature* **63** (4), 1244–1280 (2025).

45. H. C. Levy, E. M. O'Bryan, D. F. Tolin, A meta-analysis of relapse rates in cognitive-behavioral therapy for anxiety disorders. *Journal of Anxiety Disorders* **81**, 102407 (2021), doi:https://doi.org/10.1016/j.janxdis.2021.102407, `https://www.sciencedirect.com/science/article/pii/S0887618521000542`.

46. Ipsos, Google, From Innovation to Application: Our Life with AI (Second Annual Global Report) (2025), `https://www.ipsos.com/sites/default/files/ct/news/documents/2025-10/IpsosGoogle_our-life-with-ai_2025.pdf`, global survey report based on 21,000 interviews across 21 countries.

47. A. Chatterji, *et al.*, *How People Use ChatGPT*, NBER Working Paper 34255, National Bureau of Economic Research (2025), https://www.nber.org/papers/w34255.

48. R. K. McBain, *et al.*, Use of Generative AI for Mental Health Advice Among US Adolescents and Young Adults. *JAMA Network Open* **8** (11), e2542281 (2025), doi:10.1001/jamanetworkopen.2025.42281.

49. I. Haaland, C. Roth, J. Wohlfart, Designing information provision experiments. *Journal of economic literature* **61** (1), 3–40 (2023).

50. S. Carolan, P. R. Harris, K. Cavanagh, Improving Employee Well-Being and Effectiveness: Systematic Review and Meta-Analysis of Web-Based Psychological Interventions Delivered in the Workplace. *J Med Internet Res* **19** (7), e271 (2017), doi:10.2196/jmir.7583, http://www.jmir.org/2017/7/e271/.

51. M. Saijonkari, *et al.*, Promotive and preventive interventions for mental health and well-being in adult populations: a systematic umbrella review. *Frontiers in Public Health* **11**, 1201552 (2023), doi:10.3389/fpubh.2023.1201552, https://www.frontiersin.org/articles/10.3389/fpubh.2023.1201552/full.

52. F. Chopra, I. Haaland, C. Roth, N. Röver, Evaluating Behavioral Interventions at Scale with AI (2026).

Supplementary Materials
AI Against Anxiety: Efficacy, Mechanisms, and Demand for Voice-Based AI Therapy

Bennet Feld Nicolas Roever

March 3, 2026

Contents

A Materials and Methods **1**

B Supplementary Text **5**
B.1 RCT Sample Recruitment and Selection 5
B.2 Robustness of Treatment Effects 11
B.3 Engagement-Based Treatment Effect Heterogeneity 16
B.4 Robustness of Crowding Out Result 18
B.5 RCT Sample Preferences Compared to "Representative" Survey 20
B.6 Background on CBT and Treatment Approach 23
B.7 Sample AI-Therapy Session Transcript 25
B.8 Classification Workflow for Open-Ended Text Data 32
B.9 Information Treatment 33
B.10 Clinical Outcomes 35

C Additional Figures **36**

D Additional Tables **46**

E Preregistration **50**

F Survey Instructions **53**
F.1 RCT Survey Instructions Wave 1 53
F.2 RCT Survey Instructions Wave 2 55
F.3 RCT Survey Instructions Wave 3 59
F.4 RCT Survey Instructions Wave 4 63
F.5 RCT Survey Instructions Wave 5 67
F.6 Representative Survey Instructions 71

A Materials and Methods

Experimental Design

We conducted a preregistered, parallel-group randomized controlled trial evaluating a commercial voice-based AI cognitive behavioral therapy application (Sonia) versus passive psychoeducation among U.S. adults with elevated anxiety symptoms. Primary and secondary outcomes were assessed at 1, 2, and 4 weeks post-randomization (Waves 3, 4, and 5). The preregistered primary outcome was standardized GAD-7 score at 2 weeks (Wave 4). The trial was registered prior to data collection under asPredicted #258,203 and approved by the ethics board of the University of Cologne (Application 240088NR). The full survey instructions are given in Section F, and the full data used, as well as the replication package, are available at `DOI: 10.5281/zenodo.18506142`.

Given the sensitive nature of the topic, participants need to give explicit consent at multiple parts of the trial: We ask for spoken consent during the phone screener, and we require participants to give consent at the start of each subsequent wave.

Wave 1 After clicking on our recruitment advertisement, individuals are redirected to a study website to complete a brief baseline questionnaire. Eligibility is determined primarily by participants' scores on the Generalized Anxiety Disorder-7 (GAD-7) scale, a widely used instrument for assessing anxiety symptoms. We include individuals with GAD-7 scores between 13 and 20, ensuring that participants exhibited at least moderate anxiety. Scores of 21 are excluded as well, as this response pattern—selecting the highest option on every item—likely indicates low effort or inattention. We apply the following additional pre-registered exclusion criteria: participants have to (1) complete the survey on an iPhone running iOS 18 or newer, (2) report using an iPhone as their primary device, (3) be at least 18 years old, (4) reside in the United States and complete the survey while located there, (5) not be currently in therapy, (6) not report being very or extremely concerned about privacy when discussing personal experiences with AI, (7) indicate a comfort level of at least 50 on a 0–100 scale when asked about using AI for emotional support, and (8) spend at least 20 seconds completing the GAD-7, to ensure data quality and exclude bots or inattentive responses.

Phone Screener Participants who meet eligibility criteria can schedule a phone call with a member of our research team. In this phone call, we explain the trial procedure to participants and confirm that they have sufficient time to be part of the trial. We also check that they do not currently have suicidal ideation and do not reside in Nevada or Illinois which introduced legislation limiting the provision of AI-based health apps. People who fulfill these criteria are invited via email and text message to the Wave 2 survey. This personal telephone screening primarily serves to reduce attrition, as the direct contact with research staff increases participant commitment and investment in the study.

Wave 2 As pre-registered, recruitment continued until $N = 400$ participants had completed wave 2 and were randomized. Fifty percent of participants were randomly assigned to download the AI emotional support app, and 50% to a control condition. The control group was directed to a web-app containing psychoeducational material.

Before treatment assignment, baseline depression levels were measured using the PHQ-8 scale, and additional mechanism measures were collected. To study the frequency of automatic thoughts, the

ATQ-8 questionnaire was administered.

Behavioral activation was measured using two self-report items assessing activity engagement over the past two weeks. The positive item asked: "During the past two weeks, how often did you do something you genuinely enjoyed or found meaningful?" The negative item asked: "During the past two weeks, how often did you put off or skip activities because you felt low-energy, down, or had negative thoughts?" Answer options were the same for both items ("less than 20% of the time", "20–40% of the time", "40–60% of the time", "60–80% of the time", "over 80% of the time"; coded 1–5), and responses were summed to yield a total behavioral activation score (range: 2–10).

Cognitive distortions, namely bias toward negative information, were assessed with two items capturing threat-related attentional and memory/interpretation biases over the past two weeks. The first item indexed *threat-related attentional bias*: "When something unclear or unexpected happened in the past two weeks, how often did you immediately expect it to turn out badly?" The second item targeted *threat-related memory/interpretation bias*: "How often did you notice your attention being pulled toward potential threats, dangers, or negative cues, rather than neutral or positive information during the past two weeks?" Answer options were the same as for the behavioral activation items and coded from 1 to 5 for each question, with higher scores indicating stronger negativity bias; scores on the two items were summed to form a total cognitive distortion score (range: 2–10).

Waves 3, 4 and 5 Outcome measures are collected at follow-ups after 1, 2 and 4 weeks post randomization. Our main pre-registered mental health outcome is the standardized GAD-7 score at the two-week follow-up. We also collect PHQ-8, as well as all mechanism measures in the follow-up waves.

Payments Participants are informed that they will receive up to $50 for completing all study waves, along with entry into a lottery to win an iPhone. Compensation is identical irrespective of assignment to the treatment or control group. The compensation schedule is communicated at the end of Wave 1, and is structured to increase gradually across survey waves: participants receive $5 for completing the second wave and downloading the *Sonia* app (or completing the first follow-up survey in the control group), $10 for Wave 3, $15 for Wave 4, and $20 for Wave 5.

To reduce attrition, we introduced two unannounced incentives during the trial. Between Waves 1 and 2, all participants in treatment and control groups were notified of a surprise Thanksgiving lottery for an iPhone. Between Waves 4 and 5, a random subset of ten participants received a $10 gift card as an additional retention incentive.

Importantly, compensation is tied exclusively to survey completion rather than engagement with the intervention itself. This design choice ensures that usage behavior reflects genuine engagement with the AI system rather than financial motivation.

Safety Considerations To ensure the protection and well-being of a vulnerable study population, the trial is designed with a multi-layered safety protocol. First, our advertisement clearly states that the project involves testing a novel mental health application. Second, all individuals are shown the experiment procedure and have to indicate that they feel ready to participate in such a trial as part of the baseline survey. Third, during the phone-based screening we check for past-year suicidal ideation prior to randomization, allowing us to exclude participants for whom participation

might pose undue risk. Fourth, during app onboarding to *Sonia*, participants received transparent information about the purpose of the study and the intended role of the app as a supplementary rather than substitutive tool for professional mental health care. Fifth, the *Sonia* app further embeds direct links to clinical resources, ensuring immediate access to human support if needed. Sixth, *Sonia* has an automated "active session abort" mechanism that detects acute distress indicators through natural language processing. When crisis language or severe distress markers are identified, the system immediately terminates the session and redirects users to appropriate professional services, including the 988 Suicide and Crisis Lifeline and local emergency resources.

Researcher Independence This trial was conducted in cooperation with and partly funded by Bloit Inc., the developers of the AI-therapy app. To safeguard scientific independence, we established a binding pre-trial agreement: researchers maintain full autonomy over all analytical and publication decisions, while Bloit Inc. may choose whether to be identified or remain anonymous in research outputs.

Recruitment Details

We recruit participants in November 2025 through targeted social media advertisements on Facebook and Instagram. In total, $10,490$ completed our Wave 1 questionnaire, out of which 893 were eligible and scheduled a phone call. We continued recruiting and phone calls until 400 participants were randomized in the Wave 2 survey using the *Randomizer* feature in Qualtrics. For a more thorough discussion of the sample composition, see Section B.1.

Statistical Analysis

We estimated intention-to-treat effects using analysis of covariance (ANCOVA):

$$Y_i = \beta_0 + \beta_1 \text{Sonia}_i + \beta_2 Y_{i,\text{baseline}} + \beta_3 X_i + \varepsilon_i$$

where Y_i denotes the outcome for participant i at endline, Sonia_i is treatment assignment, $Y_{i,\text{baseline}}$ is baseline outcome value, and X_i includes prespecified covariates (prior therapy experience and perceived probability that therapy would be helpful). Standard errors were heteroskedasticity-robust and clustered at the individual level. For Wave 2 outcomes without baseline values, we controlled for baseline GAD-7.

Robustness analyses included: baseline-carried-forward imputation for attriters; Lee bounds to assess selective attrition; adjustment for baseline imbalances (income, medication use); and controls for social desirability bias and experimenter demand effects.

Heterogeneity analyses examined treatment effect modification by baseline symptom severity, CBT mechanism measures, and prior treatment experience using interaction terms.

Deviations from Pre-Registration

Due to a technical error, we did not collect motivation to work on one's mental health at baseline, which is why we drop this pre-registered covariate. This covariate was included to improve statistical power rather than to control for a confounding variable. Survey responses were included if submitted

within 5 days of the scheduled due date (we did not automatically close the survey; however only one observation required this extension).

B Supplementary Text

B.1 RCT Sample Recruitment and Selection

We recruited participants via targeted advertisements on Facebook and Instagram (Meta Platforms), using the Meta Pixel to track study-website visits and to optimize advertising delivery. Following screening, we transmitted an eligibility indicator (eligible vs. ineligible) back to Meta to improve cost-efficiency by preferentially allocating impressions toward users more likely to meet inclusion criteria.

We discuss the selection process in more detail in Section B.1.1 and then turn to describe the characteristics of our final randomized sample in Section B.1.2.

B.1.1 Sample Selection

The pool of baseline respondents is pre-selected by construction: individuals must first be reached by platform targeting and then opt in by clicking the advertisement. Within this selected pool, we impose four additional eligibility criteria.

Two criteria follow directly from the trial's clinical focus on CBT for anxiety among participants with clinically relevant baseline symptoms. We therefore require a baseline GAD-7 score between 13 and 20; 35.1% of Wave 1 respondents fall outside this range. To avoid spillovers between existing care and the study interventions, we also exclude individuals currently in therapy, who comprise a substantial share of Wave 1 respondents ($> 20\%$; Figure S1).

To mitigate non-compliance, we further restrict eligibility on two variables plausibly predictive of adoption and use of AI-based mental health tools: comfort engaging with an AI system for emotional support (0–100 scale) and perceived privacy risk (5-point Likert scale). We benchmark these distributions against a supplementary survey designed to be representative of U.S. adults by age, gender, and race (Figure S2). In the representative sample, 48.3% report comfort < 50, compared with 24.4% in Wave 1. Similarly, 39.7% report being very or extremely concerned about privacy when speaking to an AI, versus 26.8% in Wave 1.

Taken together, these comparisons indicate that our sample is selected in predictable ways, but it still maps onto a substantial segment of the general population that would plausibly adopt AI-based therapy and emotional support. Thus, we consider the estimates from our study to be informative for a non-trivial population fraction, large enough to be relevant for policy and implementation decisions.

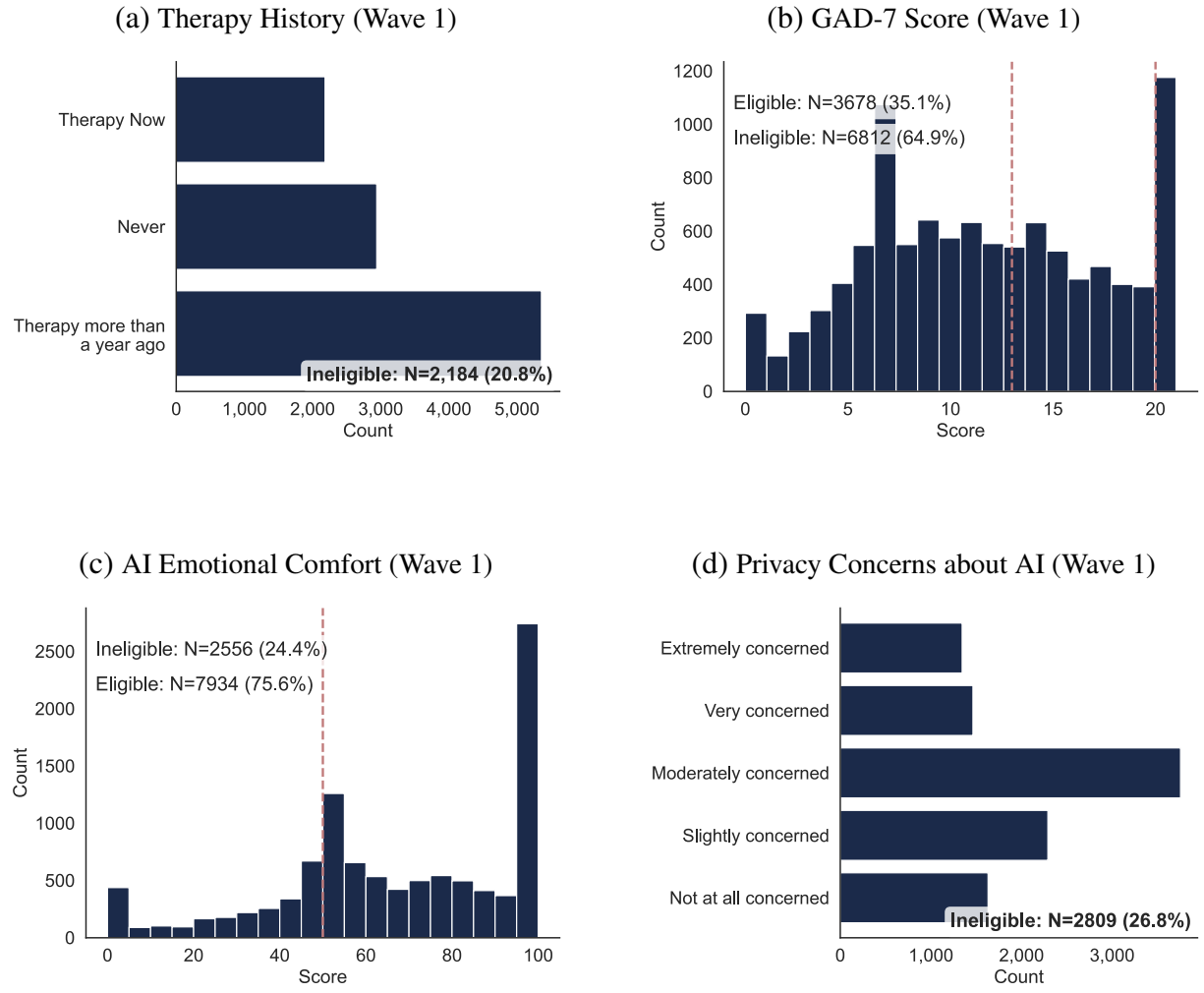
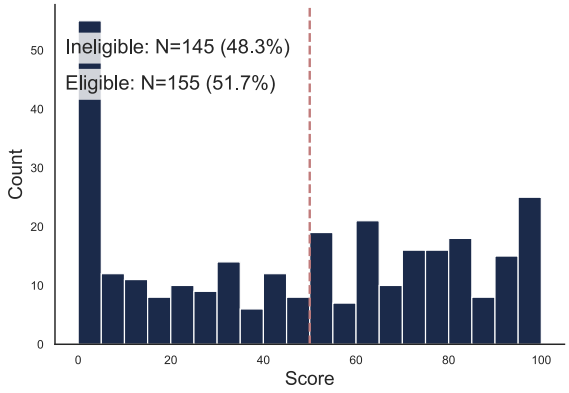


Figure S1: **Eligibility Screener: Wave 1 Full Sample.** This figure summarizes the Wave 1 eligibility criteria across the full sample, showing distributions for therapy history, AI privacy concern, AI emotional comfort, and GAD-7 scores.

(a) AI Emotional Comfort



(b) Privacy Concerns about AI

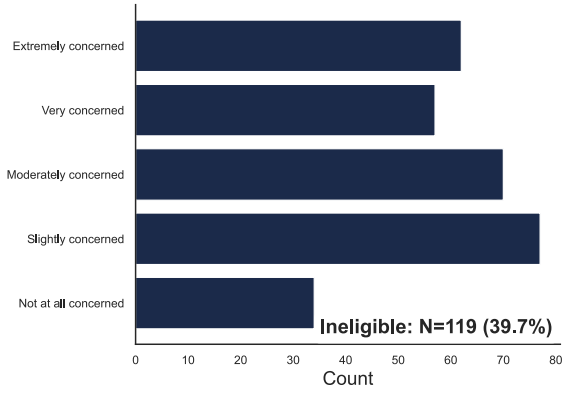


Figure S2: **AI Attitudes: Representative Survey.** This figure summarizes AI-related eligibility criteria from the representative survey, showing distributions for AI emotional comfort and privacy concerns about AI.

B.1.2 Randomized Sample Descriptives

In line with the goal of this study and pre-registered sample selection, the sample represents a population high in symptom burden and high in potential adoption likelihood—the group most relevant for early real-world deployment of AI-based psychological support.

Baseline Mental Health Eligible participants exhibit high baseline symptom severity, corresponding on average to the "severe" range on the GAD-7 scale (GAD-7 ≥ 15). Most participants also report elevated depressive symptoms as measured by the PHQ-8. As expected, higher PHQ-8 scores are associated with lower behavioral activation and greater cognitive distortions at baseline, validating (Figures S3a and S3b).

(a) PHQ-8 and behavioral activation scores elicited in wave 2 with small random jitter for visualization and OLS-regression line.

(b) PHQ-8 and cognitive distortion scores elicited in wave 2 with small random jitter for visualization and OLS-regression line.

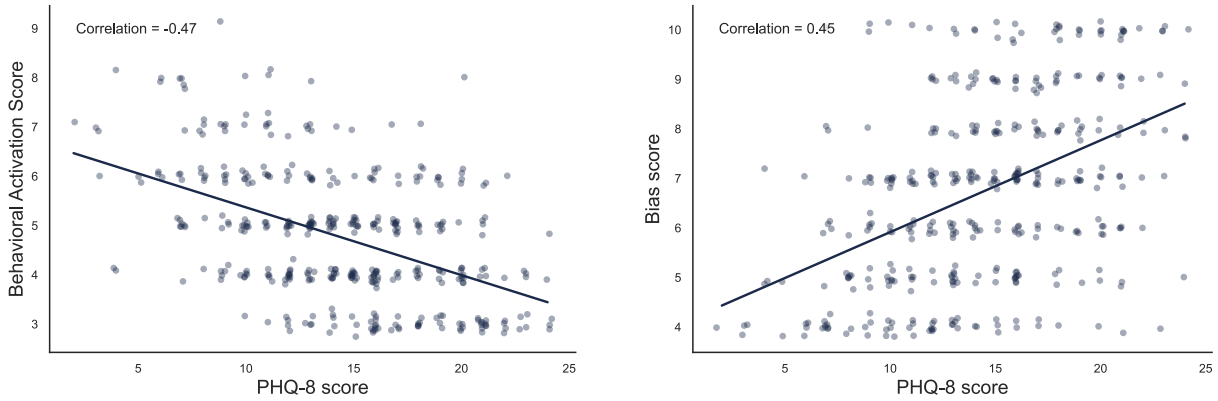


Figure S3: PHQ-8 relationships with behavioral activation and cognitive distortion.

Attitudes towards Psychotherapy Approximately 80% of the sample have previously engaged in psychotherapy and report prior use of psychiatric medication. Despite this, respondents expressed substantial self- and public-stigma related to mental health in our baseline survey, as measured using the instrument from Roth et al. (2024). Attitudes toward conventional therapy are generally favorable: On average, participants believed there was a 57.8% chance that their mental health would improve after a single talking-therapy session.

Baseline Beliefs about AI Given that our recruitment advertisement explicitly described the study as evaluating an AI-based support tool, participants are positively predisposed toward AI. As shown in Figure S4, most respondents report frequent use of large language model (LLM) tools and rate the accuracy of AI-generated medical information as relatively high. They also tend to believe that AI-delivered therapy could be effective for them personally.

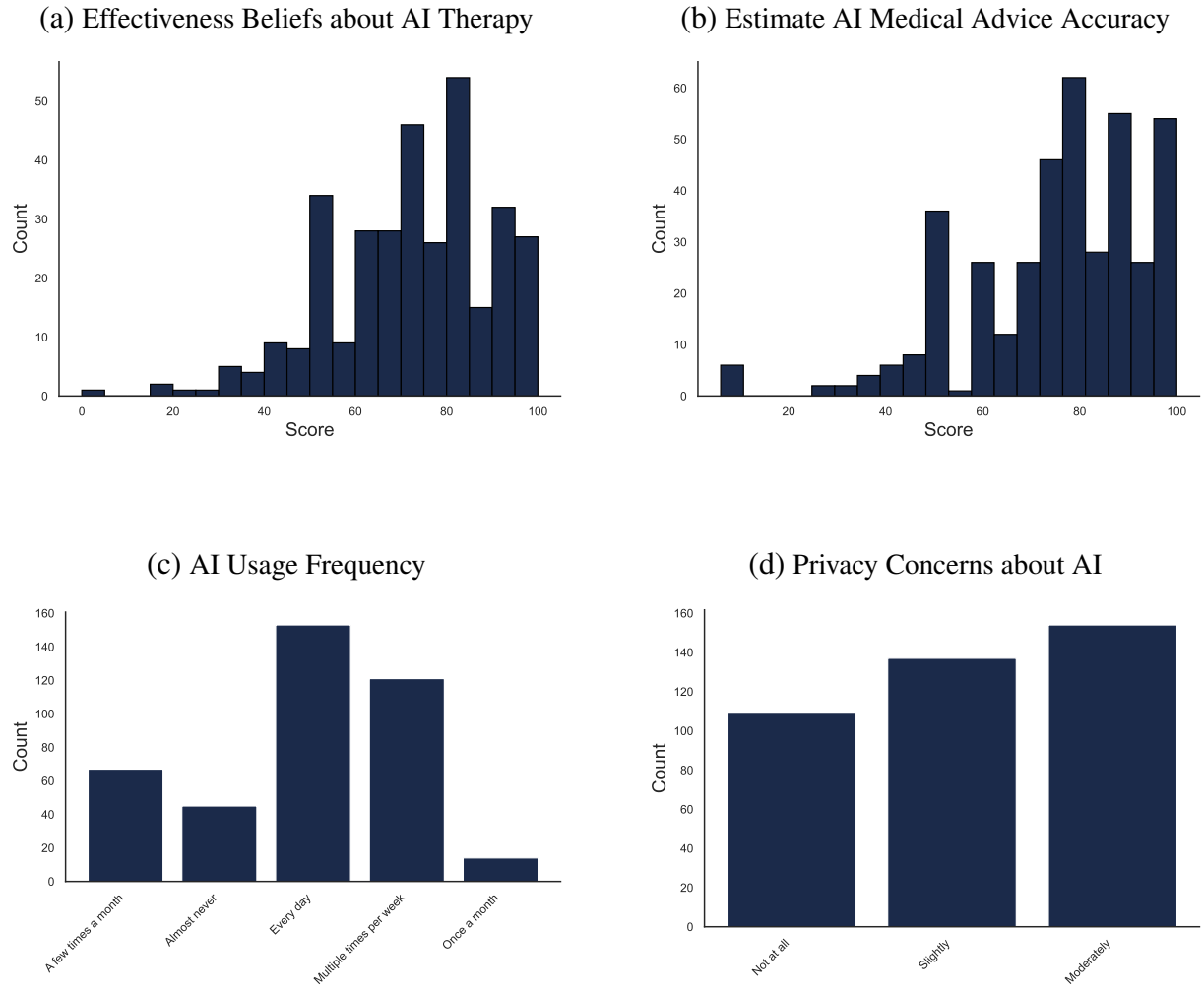


Figure S4: **Baseline descriptives on beliefs about AI in the full randomized sample ($N = 400$).** Participants reported baseline beliefs about AI using brief, single-item measures presented as sliders, a numeric estimate item, and two Likert-type questions. Trust in AI was assessed by asking: “How effective do you think psychotherapy fully delivered by an AI-system is for you personally?”, answered on a 0–100 visual analog scale anchored from "Not at all effective" (0) to "Very effective" (100) (higher scores indicate greater perceived effectiveness/trust in AI-delivered psychotherapy). Perceived AI medical accuracy was assessed with an estimate item from 0 to 100 (higher values indicate greater perceived alignment of AI outputs with scientific consensus). AI usage frequency and the level of privacy concerns about AI were measured on 5-point Likert scales, with responses summarized as category counts in the corresponding panels (higher values indicate more frequent AI use and greater privacy concern, respectively).

Balance of Baseline Covariates In Table S1 we show that participants assigned to treatment and control group have balanced baseline covariates for most variables. To account for imbalances in baseline income and the fraction who has taken psychiatric medication, we also report treatment effects with these additional covariates and show that the treatment effect estimates remain robust.

Table S1: Balance of Baseline Covariates

	Treatment	Control	p-value
N	200	200	–
Age (mean)	39.0	39.2	0.9
Percent Female	76 %	81 %	0.27
Mean Income	40,429	47,242	0.06
Percent Employed	43%	50%	0.16
GAD-7 score (mean)	15.9	16.1	0.34
PHQ-8 score (mean)	14.1	14.5	0.35
Percent Taken Psychiatric Medication	78 %	67 %	0.02
Percent Been to Therapy	76 %	74 %	0.64

Notes: This table reports mean baseline characteristics separately for the treatment and control groups. Female is an indicator equal to 1 for respondents identifying as female. Medication indicates whether a respondent has ever taken medication coded as 1. Employment equals 1 for individuals working full-time or part-time. P-values report two-sample t-tests for equality of means across treatment arms.

B.2 Robustness of Treatment Effects

This section evaluates whether the estimated mental-health benefits of assignment to Sonia are robust to plausible threats to internal validity and to alternative modeling choices. Specifically, it (i) documents stability of the main ANCOVA estimates across specifications that add different baseline covariates, (ii) assesses sensitivity to sample attrition using conservative imputation and Lee bounds Lee (2009), and (iii) assesses potential confounds from social desirability bias and experimenter demand.

Robustness Across Alternative Specifications Table S4 shows that the effect size of standardized GAD-7 in Wave 4 is virtually unchanged across alternative specifications that add different baseline covariates, including income and medication use, which are slightly imbalanced across groups at baseline. Point estimates and standard errors move only minimally, suggesting that the main treatment effect is not sensitive to specification choice or residual baseline imbalances.

Attrition Attrition by the 2-week follow-up—where we measure our primary GAD-7 outcome—is only 2.5% (10 participants), which substantially limits scope for attrition bias. Under a conservative baseline-carried-forward imputation for attriters, treatment effects remain essentially identical to the main specification (Appendix Table A2), and Lee bounds are tightly clustered around the point estimates (Appendix Figure A1), indicating that selective attrition is unlikely to account for the findings Lee (2009).

Social Desirability Bias from Self-Image Concerns A natural concern in the measurement of mental health is social desirability bias (SDB), which can arise from self-image concerns: Individuals are reluctant to acknowledge poor mental health, resulting in a downward bias in self-reported outcome levels. However, this threatens the validity of treatment effect estimates only if the intervention differentially affects individuals' willingness to acknowledge poor mental health, that is, if self-image concerns differ between treated and control groups after treatment. We follow Bursztyn et al. (2025) and assess self-image concerns around mental health in Wave 5. Figure S6a shows the answer distributions for treatment and control; the p-value of a Mann-Whitney-U-Test is at 0.16, alleviating concerns that differential SDB across treatment and control contaminates our effect estimates. We also estimate our main treatment effect controlling for self-image concerns and results remain robust (Table S5 column 1).

Experimenter Demand Another potential concern is experimenter demand effects, which in this context constitute a specific form of SDB: participants in the treatment group may infer that the experimental hypothesis predicts improvement and adjust their responses to align with perceived expectations of the experimenter (the salient social group) De Quidt et al. (2018).
To address this concern, we induce a comparable expectation in the control group by providing psychoeducational materials. In the control arm, 98% of participants registered for the educational web app, although actual engagement with the content was limited (see Figure S3a and Figure S3b). Nevertheless, the framing of the web app conveys an expectation of mental health improvement, implying the presence of experimenter demand even in the control group.
We explicitly test for experimenter demand effects in Wave 5, following best practices from the literature De Quidt et al. (2018); Bursztyn et al. (2025). Participants are asked whether they felt pressured to adjust their responses during the trial and about social-image concerns related to mental

health (Figure S6b) (in our online-survey setting, the experimenters constitute the relevant social audience). We then re-estimate our main treatment effects controlling for social-image concerns (Table S5, column 2) and excluding participants who report feeling pressured to adjust their answers (Table S5, column 3). In both cases, the results remain virtually unchanged, alleviating concerns that experimenter demand effects drive our findings.

Table S2: Attrition Across Waves

	N Treatment	N Control
Wave 1 (November 14th–19th 2025)	10,490	
Wave 2 (November 14–19 2025)	200	200
Wave 3 (November 22 - December 02)	195	198
Wave 4 (November 28 - December 07 2025)	193	197
Wave 5 (December 13 - 20 2025)	194	196

Notes: This table reports the number of observations in the treatment and control arms across survey waves. Wave 1 was conducted prior to randomization and therefore represents the pre-treatment sample. Subsequent waves are post-randomization follow-ups. Counts reflect completed surveys in each wave.

Table S3: Robustness to Attrition: Baseline Mental Health Carried Forward, Wave 4

	GAD-7 (Points)	GAD-7 (Std.)	PHQ-8 (Points)	PHQ-8 (Std.)
Treatment	-3.30***	-0.59***	-2.83***	-0.63***
	(0.46)	(0.08)	(0.48)	(0.11)
Baseline Controls	✓	✓	✓	✓
Obs	400	400	400	400
Adj. R^2	0.21	0.21	0.35	0.35

Notes: This table reports sensitivity analyses for Wave 4 outcomes under a conservative missingness assumption: for participants who attrit prior to Wave 4, outcomes are imputed using their baseline values (baseline carried forward). All specifications control for baseline therapy history and baseline perceived efficacy of human-delivered therapy. Standard errors in parentheses are heteroskedasticity-robust. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

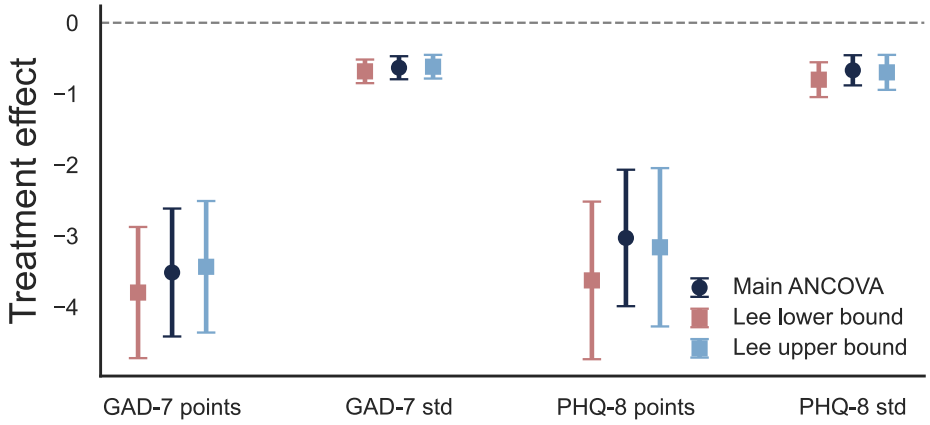


Figure S5: **Main ANCOVA estimates and Lee bounds for the main treatment effect on anxiety (GAD-7) and depression (PHQ-8).** The figure plots intention-to-treat treatment effects of assignment to the AI therapy intervention versus control, estimated using the pre-registered ANCOVA specification (endline outcome regressed on treatment assignment, baseline outcome, and pre-specified baseline covariates), shown separately for outcomes in raw points and standardized units. Point estimates are displayed as markers with vertical error bars denoting 95% confidence intervals, and the dashed horizontal line at 0 indicates no treatment effect.

Table S4: Effect Robustness – Wave 4

	Dependent Variable: Standardized GAD-7 Score in Wave 4					
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.63***	-0.66***	-0.63***	-0.64***	-0.63***	-0.68***
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
WTP Human Therapy / Income						
(Wave 2)		10.41				
		(8.41)				
AI Accuracy Estimate (Wave 1)			-0.00			
			(0.00)			
Comfort with AI						
Emotional Support (Wave 1)				-0.00		
				(0.00)		
AI Privacy Concern (Wave 1)					0.06	
					(0.05)	
Ever Taken Medication (Wave 2)						0.12
						(0.11)
Household Income (Wave 2)						-0.00**
						(0.00)
Baseline Controls	✓	✓	✓	✓	✓	✓
N	390	382	390	390	390	382
Adj. R^2	0.22	0.23	0.22	0.22	0.22	0.24

Notes: This table reports OLS (ANCOVA) estimates of treatment effects on standardized GAD-7 in Wave 4. Standardization uses the full baseline sample. Baseline controls include baseline standardized GAD-7, prior therapy history, and perceived efficacy of human-delivered therapy. Specification (1) corresponds to the main specification. Standard errors (in parentheses) are heteroskedasticity-robust. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table S5: SDB and Experimenter Demand: Robustness Regressions

	(1)	(2)	(3)
Treatment	-0.64***	-0.61***	-0.61***
	(0.08)	(0.08)	(0.08)
Control	Self Image	Social Image	
Baseline Controls	✓	✓	✓
N	386	386	373
Adj. R^2	0.23	0.24	0.24

Notes: Dependent variables are standardized GAD-7 (standardized using the full baseline sample). All specifications control for baseline standardized GAD-7, prior therapy history, and perceived efficacy of human-delivered therapy; specification (1) is the main specification for reference. In specification (3), we exclude respondents who reported adjusting their answers due to perceived experimenter demand. Standard errors (in parentheses) are heteroskedasticity-robust. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

(a) Self-image concerns item (Wave 5). Mann–Whitney U test p=0.16.

(b) Social-image concerns item (Wave 5). Mann–Whitney U test p=0.13.

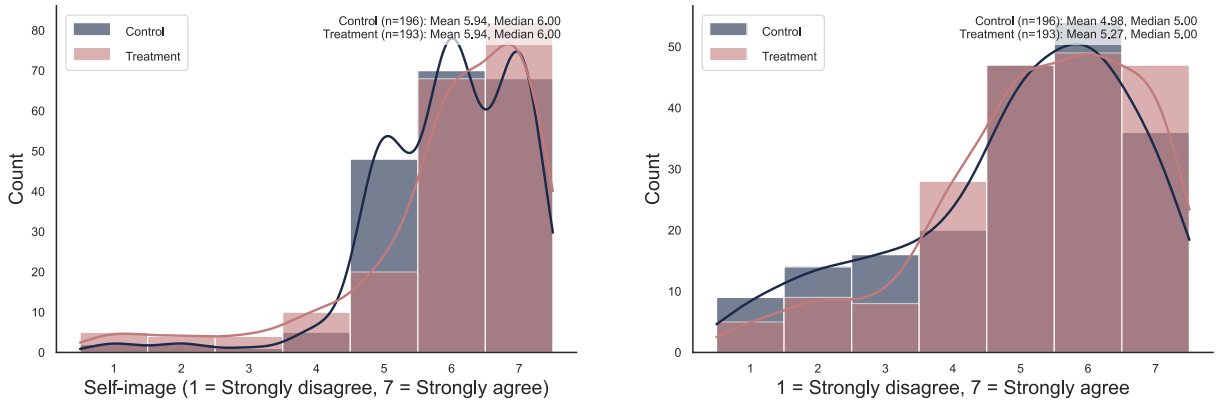


Figure S6: **Self-image and social-image response distributions by treatment arm (Wave 5).**

B.3 Engagement-Based Treatment Effect Heterogeneity

App engagement was not incentivized, and **treatment intensity** was subsequently self-assigned. We report engagement heterogeneity here with a note of caution that these estimates cannot be given a causal interpretation. The core identification problem is reverse causality: participants whose symptoms improve may reduce their use of the app precisely because they feel better, meaning that low usage can reflect therapeutic success rather than disengagement. This mechanical relationship would bias against finding a dose–response gradient and would lead us to understate heterogeneity.

To examine whether greater exposure to the intervention is associated with larger symptom reductions, we interact the treatment indicator with within-arm quartiles of cumulative therapy sessions (defined as conversations of at least five minutes) completed prior to each outcome wave.

Table S6, columns (1)–(3), reports the results. The omitted category is the lowest-usage quartile, so the interaction terms capture the incremental association of higher engagement relative to this reference group. Despite the likely attenuated heterogeneity noted above, we observe a gradient at the two- and four-week follow-ups: participants in the highest two usage quartiles show treatment effects roughly 0.4 standard deviations larger than those in the lowest quartile, consistent with a dose–response pattern.

We also examine whether engagement with supplementary application functionalities, namely morning reflections and evening meditations, referred to here as **"tasks"**, affects treatment responses. These activities can be executed within the app, and the AI therapist may recommend them as between-session exercises during a CBT conversation, though they do not exhaust the full range of homework the therapist may recommend, and users may use them without a homework assignment. Columns (4)–(6) of Table S6 report these results, which indicate that these tasks may increase treatment effects.

These findings of a dose-response pattern help alleviate concerns that the control groups' knowledge of their treatment status causes placebo effects that can amplify treatment effects. While this dose–response pattern is reassuring, it cannot fully rule out expectancy-related placebo effects, as engagement and symptom improvement may both be driven by unobserved motivation or demand characteristics rather than treatment intensity per se.

	Sessions			Tasks		
	Wave 3	Wave 4	Wave 5	Wave 3	Wave 4	Wave 5
Treatment	-0.24**	-0.33**	-0.55***	-0.26***	-0.51***	-0.82***
	(0.11)	(0.17)	(0.15)	(0.06)	(0.06)	(0.15)
Treatment × 25th-50th percentile	-0.15	-0.34*	-0.29	-0.05	-0.04	0.16
	(0.14)	(0.18)	(0.18)	(0.06)	(0.07)	(0.18)
Treatment × 50th-75th percentile	-0.13	-0.41**	-0.39**	-0.23***	-0.21**	-0.24
	(0.15)	(0.19)	(0.18)	(0.08)	(0.10)	(0.18)
Treatment × 75th-100th percentile	-0.10	-0.44**	-0.36**	0.03	-0.26***	0.08
	(0.14)	(0.19)	(0.18)	(0.09)	(0.09)	(0.18)
Baseline Controls	✓	✓	✓	✓	✓	✓
Obs	393	390	390	393	390	390
Adj. R^2	0.20	0.23	0.25	0.20	0.22	0.24

Table S6: **Engagement Heterogeneity.**

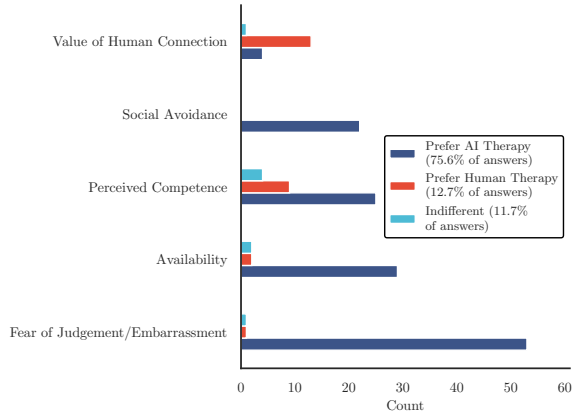
OLS estimates of treatment-effect heterogeneity for standardized GAD-7 scores. Columns 1–3 use session engagement (Waves 3, 4, 5), and Columns 4–6 use task engagement (Waves 3, 4, 5). Each regression includes Treatment, Treatment × engagement-percentile indicators, baseline standardized GAD-7 (Wave 1), therapy history (Wave 1), and perceived efficacy of human therapy (Wave 2). Engagement bins are constructed from cumulative pre-survey usage within the treatment arm, and the omitted interaction category is the lowest engagement group (reference). Interaction labels are: Treatment × 25th–50th percentile, Treatment × 50th–75th percentile, and Treatment × 75th–100th percentile (when identified by the data). Heteroskedasticity-robust (HC3) standard errors are used. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

B.4 Robustness of Crowding Out Result

A central finding of our study is that exposure to AI therapy *reduces* demand for human therapists. Respondents frequently cite lower social anxiety and diminished concerns about judgment as key advantages of AI-led emotional support and therapy. Here, we present additional evidence consistent with this interpretation.

First, people who have not been to therapy before our trial might not have an accurate view of what human therapy actually entails and could thus introduce noise or bias due to inaccurate beliefs about human therapy. As a robustness exercise, we exploit that a large fraction of trial participants attended talking therapy before the onset of the trial. Figure S7a shows the same figure as in the main text, but restricted to treated participants with prior human therapy experience. The same pattern emerges: AI-led therapy is preferred by an even larger share than in the full sample.

(a) Themes in AI vs. Human Therapy Preferences among Respondents with Prior Therapy Experience



(b) Stated Therapy Modality Preference by treatment arm (Wave 4)

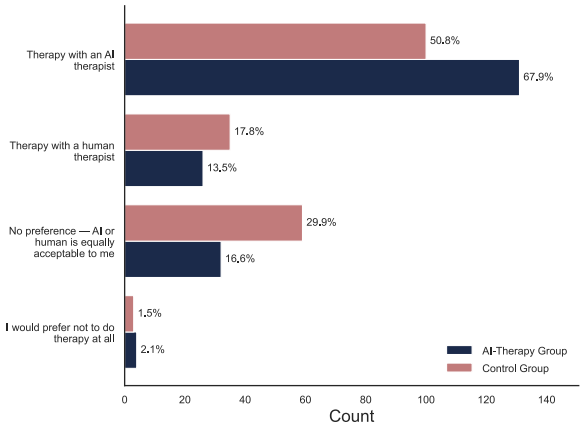
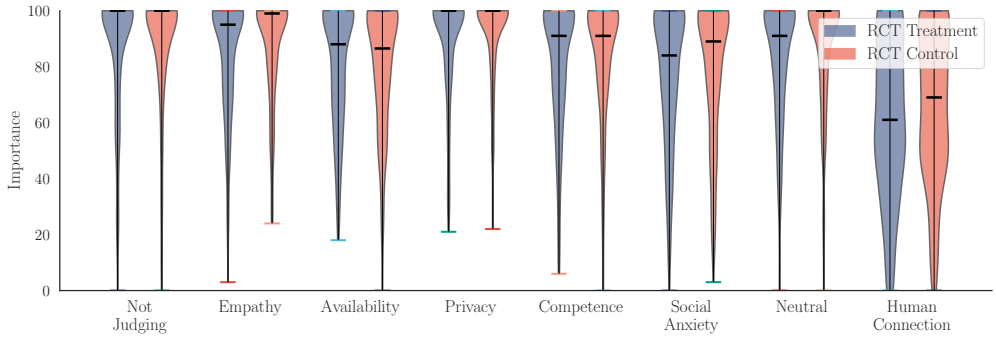


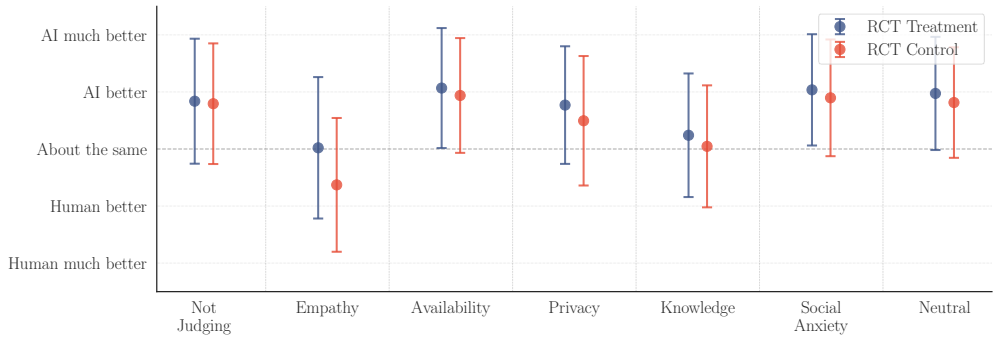
Figure S7: **Preferences for AI versus human therapy.** (A) Coded themes among respondents with prior therapy experience (Wave 3). Thematic categories coded from Wave 3 responses (8 predefined categories); bars show the distribution of coded themes describing preferences between AI-based and human therapy. We only show categories which occur at least 15 times. (B) Stated therapy modality preference by treatment arm at the two-week follow-up (Wave 4). Distribution of respondents' preference for receiving therapy from an AI therapist versus a human therapist, elicited at the two-week follow-up.

Second, in wave 4 we directly elicited preferences in the full sample across four options: preference for AI-led therapy, preference for human-led therapy, indifference between the two, or no need for therapy. Figure S7b shows that AI preference remains widespread under this closed-ended elicitation, mirroring the pattern observed in wave 3 open-text responses. Consistent with our wave 3 coding, 79.02% of respondents classified as preferring AI-led therapy in wave 3 report the same preference when re-elicited in wave 4.

Third, we assess the validity of our thematic categorization by asking trial participants to rate the importance of each identified category on a 0–100 scale. Ratings are uniformly high (Figure S8a),



(a) Emotional support importance by group. Violin plots show the distribution of item ratings for the AI-therapy and control groups with ticks representing means.



(b) Perceived AI vs. human therapist advantages in the RCT control group and a survey sample open to AI therapy. Points show item means on a five-point scale recorded from -2 (human much better) to +2 (AI much better); whiskers represent one standard deviation.

Figure S8: **Emotional support preferences and AI vs. human therapist perceptions.**

suggesting that the extracted themes capture considerations respondents view as salient. In particular, avoiding judgment is rated as highly important (the second-highest item), aligning with wave 3 accounts that emphasize evaluation concerns when seeking emotional support. Finally, Figure S8b shows that participants often rate AI higher than humans across multiple emotional-support subdomains, further indicating strong receptiveness to AI-mediated emotional support and therapy.

B.5 RCT Sample Preferences Compared to "Representative" Survey

Our sample was recruited through social media advertisements promoting an AI-based mental health tool. We subsequently selected participants who reported relatively low concerns about sharing personal information with AI and who exhibited high GAD-7 scores. Thus, advertisement and selection are designed to replicate a sample of early adopters of AI therapy with high need for mental health support. To compare our trial sample to the more general population, we conduct a survey with $N = 300$ respondents on Prolific representative of the US population in terms of gender, age, and ethnicity. In this section, we compare the control group from our trial, which is unconfounded by treatment, to survey participants who are interested in AI psychotherapy and those who are not.

Comparing the RCT control group with survey respondents interested in trying AI therapy in Table S7 shows that RCT participants are generally younger, predominantly female, and less likely to be employed than survey respondents interested in AI therapy. This mirrors typical characteristics of online samples. Owing to the RCT sample selection, they also experience anxiety and depression symptoms at substantially higher rates and are more likely to have taken psychiatric medication or to have been in therapy. While both groups report similar expectations regarding the efficacy of human therapy, there is a pronounced divergence in their estimates of the efficacy of AI therapy and in their comfort with discussing emotional issues with AI.

Comparing survey respondents who are interested in AI therapy with those who are not reveals similar patterns. Respondents who express interest appear to have a greater overall need for psychotherapy, as indicated by higher prevalence of anxiety and depression symptoms. Differences are most pronounced in expectations regarding the efficacy of human therapy and in comfort with disclosing emotional issues to AI, and less so in beliefs about the efficacy of AI therapy.

Interest in AI therapy is not uniformly distributed. Individuals expressing interest are younger, more likely to be female, and more likely to be employed than those who are not interested. They also report higher current symptom levels and greater prior engagement with mental health care, with over half currently taking psychiatric medication and roughly two-thirds having previously attended therapy. Baseline attitudes toward AI further predict interest: respondents who are interested report higher expectations regarding the efficacy of AI therapy and substantially greater emotional comfort disclosing to AI than to humans.

These patterns suggest that selection into AI therapy interest is non-random and driven by both need and favorable predispositions toward AI. The profile of interested individuals—symptomatic, treatment-experienced, and comfortable with AI—is consistent with an early-adopter population. Yet the presence of a substantial group ($N = 120$) expressing interest despite lower symptom levels and less favorable attitudes toward AI suggests that a broader population may be open to AI therapy, perhaps one step behind the early adopters.

We examine preferences for AI versus human therapy in more detail in Figure Figure S9. With respect to beliefs about effectiveness (Panel A), the RCT control group is the least skeptical: only about 23% rate AI therapy as somewhat or much less effective than human therapy, while 11.5% rate it as more effective. The survey respondents not interested in AI therapy are the most skeptical, with roughly half rating AI as less effective. The survey-interested group falls in between, though closer to the RCT sample.

Stated preferences follow a similar pattern (Panel B). The RCT control group exhibits the strongest preference for AI, with 50.5% indicating that they would choose AI therapy. In contrast, the survey respondents not interested in AI therapy show the lowest preference for AI (around 10%) and the strongest preference for human therapy (61.7%). Once again, the survey-interested group falls in between. Notably, when combining respondents who prefer AI with those who consider AI and human therapy equally acceptable, a substantial majority in both the RCT and survey-interested groups appear open to AI therapy.

Perceptions of AI's comparative advantages reveal a striking pattern (Panel C). The RCT control group stands out as markedly more favorable toward AI across nearly all attributes, including availability, privacy, and reducing social anxiety. In contrast, differences between the survey-interested and survey-not-interested groups are surprisingly small. Both survey groups perceive human therapists as superior in empathy and AI as superior in availability and privacy, with the interested group not differing markedly from those not interested.

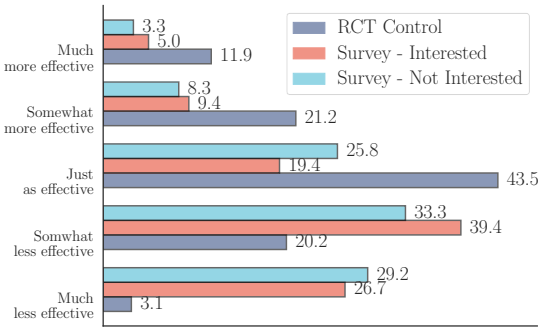
A similar pattern emerges for the importance respondents assign to different therapy attributes (Panel D). The interested and not-interested groups exhibit only small differences in what they value, with both rating empathy, human connection, privacy, and availability as highly important.

Taken together, these findings suggest that while the RCT sample represents early adopters, a considerable share of the wider population expresses openness to AI therapy. Importantly, the interested and not-interested groups differ little in their underlying values or perceptions of AI's relative strengths. This suggests that those currently not interested may represent potential future adopters rather than a fundamentally distinct population, and that uptake may increase as familiarity with AI-delivered therapy grows.

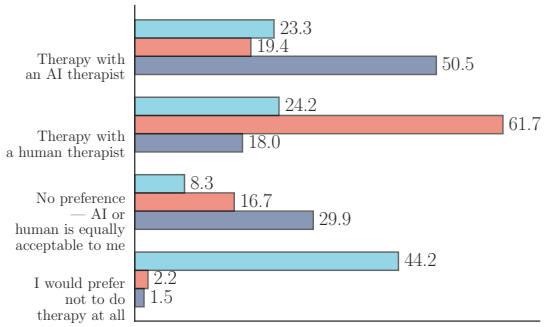
Variable	RCT Control	Survey - Interested	Survey - Not Interested
N	200	180	120
Age (mean)	39.3 (11.6)	42.5 (15.0)	51.8 (17.1)
Female (%)	81.0 (39.0)	54.0 (50.0)	42.0 (50.0)
Employed (%)	50.0 (50.0)	68.0 (47.0)	63.0 (48.0)
GAD-7 score (mean)	16.1 (2.2)	6.0 (5.5)	2.9 (4.1)
PHQ-8 score (mean)	14.6 (4.8)	6.9 (6.2)	2.9 (4.2)
Psychiatric Medication (%)	67.0 (47.0)	56.0 (50.0)	29.0 (46.0)
Been to Therapy (%)	75.0 (43.0)	66.0 (47.0)	31.0 (46.0)
Efficacy Human Therapy (mean, W1)	59.62 (24.09)	60.89 (25.53)	37.61 (29.86)
AI Efficacy Estimate (mean, W1)	76.02 (18.76)	59.61 (23.66)	61.58 (24.97)
AI Emotional Comfort (mean, W1)	83.42 (17.07)	51.34 (32.96)	39.51 (33.61)

Table S7: **Baseline descriptives for the RCT control group compared to survey participants.** The first column contains baseline characteristics for RCT participants assigned to the control group. The second and third columns report the same variables for survey participants split by whether they report being interested in engaging with AI psychotherapy.

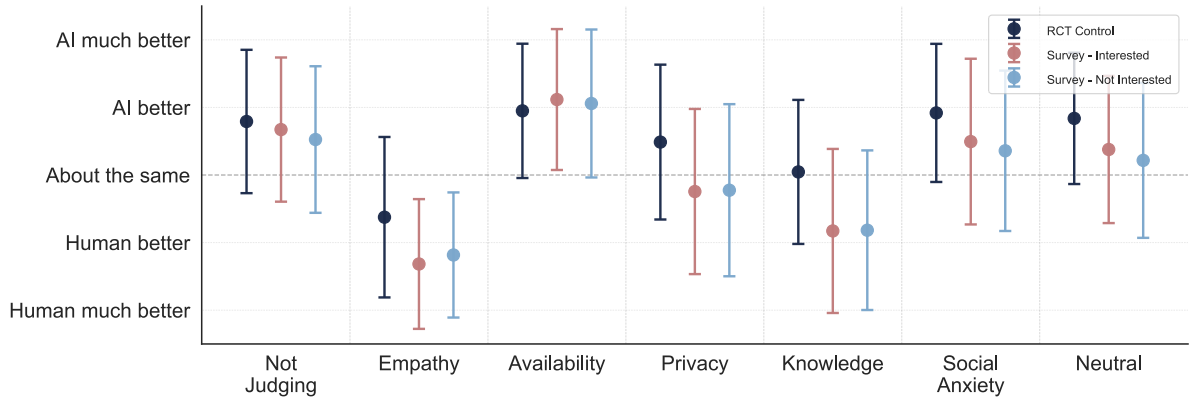
(a) Belief about Effectiveness of AI vs Human as Therapist (% of Responses)



(b) Preference for AI vs Human Therapy (% of Responses)



(c) Assignment of Therapy Attributes as Qualities of AI vs Human



(d) Self-reported Importance of Therapy Attributes in Treatment and Control Group (2 weeks after treatment) and in Survey

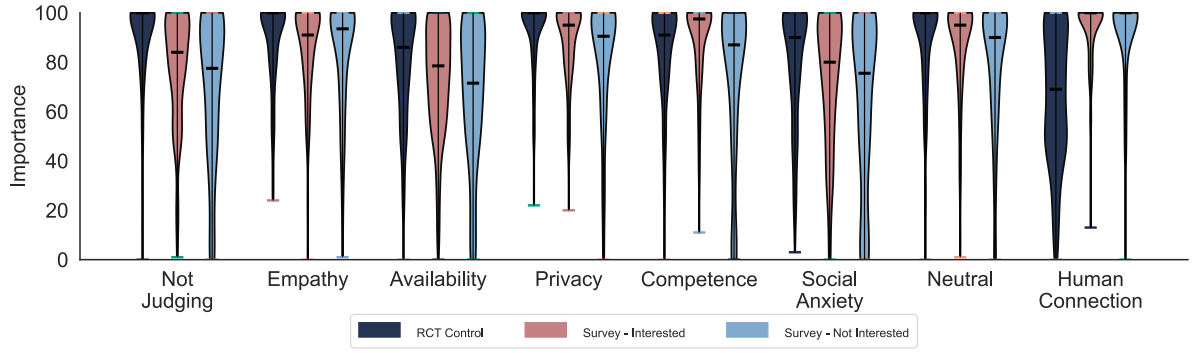


Figure S9: **Therapy Preferences in RCT and Survey.**
(**A**) Responses to "If you compare it to talking with a human therapist, how effective would emotional support from an AI feel to you personally?" (**B**) Responses to "If these were your available options, which one would you choose right now?" (**C**) Whisker plots showing whether respondents believe whether AI or a human therapist is stronger at a given attribute. Points show item means on a five-point scale recorded from -2 (human much better) to +2 (AI much better); whiskers represent one standard deviation. (**D**) Violin plots showing self-reported importance ratings (0-100 scale) for eight therapy characteristics among the RCT treatment group (blue), RCT control group (red), and survey respondents (cyan). Characteristics were identified from open-ended responses in Wave 1 using natural language processing and assessed quantitatively in Wave 4. Black lines indicate medians; violin width represents response density. N=400 (RCT); N=300 (survey).

B.6 Background on CBT and Treatment Approach

The AI-therapy application under study draws on the standard textbook protocol of CBT Burns (1980); Beck (2021); Craske (2016). CBT operates through a core principle: thoughts influence emotions, which drive behaviors, which again influence thoughts in a reinforcing cycle (Figure S10).

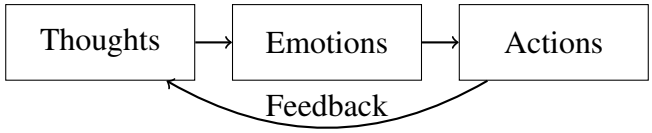


Figure S10: The cognitive model: Thoughts influence emotions, which drive actions, with feedback mechanisms.

Every session a user starts is built around five session blocks; therapeutic interventions are based on the aforementioned cognitive model. We now explain the five blocks in more detail.

(1) Mood Check At the beginning of each session, the system assesses the user's current mood state. This baseline establishes a reference point for session progress and allows the AI to tailor session intensity and focus to the user's current emotional state. For example, if a user reports significantly elevated anxiety, the system may suggest a short breathing exercise before continuing.

(2) Agenda Setting The AI and user then collaboratively establish session priorities. This typically involves reviewing the user's primary concerns on top of mind, identifying specific situations or thoughts to address, as well as past concerns mentioned in previous sessions. At the end of this block, user and AI agree on a structure for the rest of the session. If a user reports workplace stress and difficulty sleeping, the session might focus on identifying anxiety-maintaining thoughts related to work performance, with sleep optimization as a secondary goal.

(3) Review of Between-Session Practice If the session before included homework—typically behavioral experiments, thought records, or exposure exercises—this homework is typically reviewed as a first point after agenda-setting. This review reinforces accountability, identifies obstacles to completion, and extracts learning from real-world practice.

(4) Introduction and Rehearsal of Core CBT Techniques The core therapeutic work occurs in this session block. Depending on the user's presenting problems and therapeutic stage, the AI therapist introduces or deepens CBT techniques such as cognitive restructuring (identifying and challenging unhelpful thoughts), behavioral activation (scheduling valued activities), interoceptive exposure (confronting bodily sensations), or problem-solving. Each technique is explained, modeled through dialogue or example, and practiced within the session. For example, when introducing thought records, the user might work through a real recent situation step-by-step: identifying the triggering event, the automatic thought that arose, the emotion experienced, and then generating alternative, more balanced thoughts.

(5) Homework Assignment If appropriate, a session concludes with assigning specific, concrete homework tasks to extend therapeutic work between sessions. Assignments are collaboratively

determined, achievable within the user's real-world context, and directly tied to session goals. Typical assignments include completing thought records on identified anxiety triggers, scheduling behavioral activation (social contact, exercise, valued activities), conducting behavioral experiments to test anxious predictions, or practicing relaxation or grounding techniques. The system ensures the user understands the assignment, problem-solves potential barriers, and clarifies the rationale linking the homework to their treatment goals.

B.7 Sample AI-Therapy Session Transcript

This section presents a complete transcript from an AI-therapy session to demonstrate how the intervention works. To ensure participant confidentiality, the session shown here was performed by on of the authors ("Ben") following an imaginary case scenario. The session represents a first-time user experience with no stored history, and took about 30 minutes.

AI-Therapist:
Hi, I'm Sonia, your AI therapist. It's nice to meet you and I'm glad you decided to start this journey with me. In this first touchpoint I want to get to know you better and understand how I can best support you going forward. So, what brought you here today?

User:
I think there are multiple of things. So on top of my mind is that I, um, got rejected by a girl for, um, quite recently, um, which is still on my mind, um, because I kind of start, um, asking myself if I'm able to, um, if I'm able to really have long term comp... relationships. Um, the second is that I kind of feel that I'm not really able to relax well and to calm down after work. So I'm currently thinking about work and have the feeling I'm not... so I I don't have the feeling that I'm very stressed, but that I'm, uh, kind of not relaxing enough, which leads me to sometimes have days where I'm super tired and not motivated and kind of not have the spark that I used to have, I think.

AI-Therapist:
That sounds like a really draining mix — carrying the weight of that rejection while feeling like you can't fully recharge from work. What part of all this feels most pressing for you right now?

User:
I think relationships, definitely.

AI-Therapist:
When you think about that rejection and question your ability to have a long-term relationship, what feels hardest about it to you right now?

User:
I think I sometimes have the feeling that many people are already in established relationships, and then I'm kind of, you know, the time's running out for me to find a person.

AI-Therapist:
You know, that worry about time slipping away can really amplify the sting of rejection; what kind of fear sits behind that thought that time's running out?

User:
I guess, like, being lonely for... I I... yeah. It's loneliness.

AI-Therapist:
That feeling of loneliness can cut deep, especially when it seems like others have already figured it out. How do you usually try to handle that sense of loneliness when it shows up?

User:
So I think... I mean, my main thing is kind of seeking contact with other people no matter who it is.

AI-Therapist:
When you reach out to people in those moments, do you usually end up feeling more connected or does it sometimes leave you still feeling alone afterward?

User:
I think it's a mix. It's a fifty fifty. Sometimes, yes, I feel more connected. I'm also distracted. You know what I mean? Um, but... On the other hand, sometimes I'm I'm sitting with a person, and I'm thinking, like, okay. What am I doing here? I'm kind of not spending time with the person... with the people I really want to spend time with, and this is not really also fair to the person I'm hanging out with at that time. You know?

AI-Therapist:
Sounds like part of you seeks connection just to ease the sting of being alone, but ends up feeling even more detached at times. Share a bit about this recent rejection — how long had you known her, and what led up to that moment?

User:
While I have known her for, I don't know, um, three or six months, and, um, I guess I just I just wrote her a text, you know, or I just asked her, you know, we've known each other. We've had a good time with each other. Um, do you maybe wanna wanna wanna date, um, me? And, yeah, she did not respond to that. well.

AI-Therapist:
What was it like for you when she didn't respond well?

User:
That felt really hurt. Yeah. I think that's hurt and sad also. It feels like I've missed an opportunity. I'm missing out on a lot of good time I could have with this person and that that I've just kinda lost this gem.

AI-Therapist:
It really sounds like losing her left you grieving not just a person but the hope of what could have been; what do you think made her feel like such a gem to you?

User:
I I think just the fact that she wanted to spend a lot of time with me. Yeah. And that I enjoyed her company.

AI-Therapist:
What was it about the time you spent together that made you feel so comfortable or understood?

User:
I think that we share the same type of humor, um, yeah, and that we share the same type of feelings about a lot of stuff.

AI-Therapist:
Okay, so what I'm hearing is that being with her felt natural and easy, like you shared a similar outlook and sense of humor. Given that connection, it makes sense the rejection hit hard — what do you think it stirred up in terms of your confidence around relationships?

User:
I think it it kind of... I'm not sure of my confidence. I sometimes feel like I need to work hard for relationships, and and then I feel kind of... if I lose a relationship, you know, I feel like a bit drained and like, okay. I need to work again, and it's, like, tough. I need to go out again when... and sometimes I just don't don't have the energy. I feel like I have the energy to do it.

AI-Therapist:
When you notice that drained feeling after putting effort into relationships, what usually happens next for you?

User:
I feel sad, frustrated, angry at the world.

AI-Therapist:
What happens with that anger when it builds up?

User:
Usually, I have a conversation in my head where I just, like... I'm gonna play a a shouting individual, and then I kinda reflect that this is kinda not a constructive thought to to to use with others or to to to build a conversation with anyone around, and then I just try to let it go, actually.

AI-Therapist:
Ben, when you talk about questioning your ability to have long-term relationships, what's that belief based on? Have there been relationships before that ended, or is it more that things rarely move past the early stages?

User:
I think it's more the second, so the things rarely move past the early stages.

AI-Therapist:
What do you think makes it so hard for things to move past those early stages?

User:
I mean, I'm not super secure around relationships. Um, also, I feel like I don't meet a lot of people where I think it's worth it, and I would like to move past the early stage. And if I meet them, then it's... in in the past, it it just just didn't work out, I guess.

AI-Therapist:
Can you describe what "not super secure" looks like for you when you start getting to know someone?

User:
I think I need very strong signs of validation before I ask someone out. Um, Yeah. Which usually takes a lot of time, and I'm I'm questioning myself a lot if it's the right decision or not. Yeah.

AI-Therapist:
So you tend to wait until you're sure the other person likes you before making a move, which means there's a lot of overthinking and second-guessing in the meantime. That can be exhausting and make you feel like relationships are this uphill effort. When you do get that clear validation from someone, how does that usually change the way you show up with them?

User:
I mean, I think validation definitely helps me, but I'm also seeking it during a relationship. I'm actually doing doing every relationship, I think. not just with, like, romantic partners, but also with friends and... yeah.

AI-Therapist:
What do you think drives that ongoing need for validation across relationships?

User:
That's a good question. Um, I mean, I think part of it is that that we all kind of needed... I think it's a it's a... I really think it's an inherent human, um, desire, which kinda makes sense. You know? Being being lonely is is dangerous for us as humans. So so we kind of need to need to have some some assurance that we're not alone. Um, however, I think that I I need more than others. Whether it's kinda... I feel like it has always been the case, actually. Right? It's always been, like, a trait of mine, if that makes sense.

AI-Therapist:
It makes sense that you'd feel drawn to external reassurance if you've always felt that pull toward needing more of it than others. When you don't get that validation, what kind of thoughts or doubts tend to show up for you?

User:
Let me think of a concrete situation whether it happened. I think I feel weak.

AI-Therapist:
Feeling weak in those moments sounds really painful, almost like your worth depends on whether others give that external signal back to you. What do you usually tell yourself when that feeling comes up?

User:
I try to kinda put the situation into perspective because, I mean, it's normal. I mean, it's very rare that people really give you validation, and and it's okay. You know? People people don't need to give you, um, concrete validation every time. And I usually try to to to kind of, um, get the... raise these thoughts to counter, um, the feeling.

AI-Therapist:
And it seems like you're trying to stay rational about it, yet you mentioned earlier that you feel drained both by work and by relationships; do you think you might just be running on empty in several areas right now?

User:
Yeah. Sometimes it feels that way. That's true.

AI-Therapist:
What does being "drained" actually look like for you day to day?

User:
I think not having... I think it it it shows up in several areas, so not being present with other people and and engaging with other people as much as I wanted. So I can can talk to other people and not truly engage. Um, also, at work, you know, not doing the hard things but doing the easy things, um, and just not having fun at work. I think that's that's a major difference when I'm... when I have energy, I have fun at work. When I don't, I don't. , Yeah.

AI-Therapist:
You really notice it affecting how engaged you are, both with people and at work. It seems like when your energy dips, the spark you usually have fades too. What helps you recover or get some of that spark back, even a little?

User:
Yeah. That's the question. I don't... it's it's... I really don't know.

AI-Therapist:
It's okay to not know yet; recovery isn't always obvious when you're in the middle of depletion. When your energy is low like that, what usually helps you at least stop it from getting worse?

User:
I I I tried to to just shut off work and sleep.

AI-Therapist:
Or when you try to shut off work, do you actually manage to relax, or does your mind keep running anyway?

User:
Yeah. I think I'm pretty... I mean, my mind keeps running, but I'm, I think, okay at not letting it go completely out of hand.

AI-Therapist:
Yeah, even though your mind keeps going, it sounds like you've learned to keep it from snowballing out of control, which takes real effort when you're already tired.

AI-Therapist:
What usually helps you draw that line between work thoughts and your own time so they don't creep back in?

User:
I mean, it's just my meditation practice, I think. I I really have learned to be aware of my thoughts. Yeah.

AI-Therapist:
Alright, since meditation helps you stay aware of your thoughts, what tends to slip through your awareness when you start running on empty anyway?

User:
Well, I think... I still think about work. I think what I can do, how I can how I can improve the output. Yeah.

AI-Therapist:
What kind of work thoughts keep looping in your mind when you try to unwind?

User:
Just to dos. what needs to be done.

AI-Therapist:
What do those to-do thoughts do for you once they start circling in your mind?

User:
I guess they distract me. You know? They give me something to think about.

AI-Therapist:
It sounds like the to-do lists help you fill the space so you don't sit with empty time or uncomfortable feelings. That reminds me of an image I often use: your energy is like a small garden you carry around, and your attention is the watering can. Sometimes when you're tired or lonely, you might end up pouring water on every plant just to keep busy instead of tending the few that actually matter. Would it be alright if I showed you that image on your screen while we talk about it?

User:
Yeah. Sure.

AI-Therapist:
Almost like those tasks give you a sense of purpose when things feel uncertain. Looking at this garden image, which parts of your own "garden" feel like they actually deserve a pour of energy right now?

User:
I think it's my... I think... let me think. I think it's really taking energy and investing into being engaged with people when I meet them.

AI-Therapist:
What would feeling genuinely engaged with people look like for you in those moments?

User:
Not showing up completely drained. Not being tired. Looking people in the eyes, sometimes asking a lot of questions, remembering what they did, how they feel.

AI-Therapist:
So what you're describing sounds like wanting to actually show up in those moments — awake, curious, genuinely tuned in. That ties nicely back to the garden image: being present is like watering the plants that matter with real attention instead of just passing by with an empty can. What kind of small habit or condition tends to help you arrive to those interactions with a bit more energy?

User:
I think taking a short deliberate break between what I've done before and what I'm transitioning in then. Right? So when I've worked a lot and thought really hard about work and then a second later, I'm a hundred... someone... obviously, I'm not super engaged.

AI-Therapist:
So those short, deliberate breaks act like a reset between chapters, giving you space to refill your own can before connecting with others. What tends to make it easiest for you to actually take that pause instead of rushing straight into the next thing?

User:
Well, I think setting the timer, right, or setting setting an alarm ten minutes before I meet this person. Okay. Now I'm stopping.

AI-Therapist:
That's a pretty practical way to draw a boundary for yourself — using a timer makes the pause real, not just theoretical. Over time, those ten minutes could become the moment you metaphorically refill your watering can before you see someone. How did you land on that approach?

User:
I like it. I like it a lot.

AI-Therapist:
You're turning that pause-and-timer trick into a real boundary between worlds, which is clever. Do you ever notice a difference in how your conversations go with people when you've managed to take that break first?

User:
Yeah. Sure. I'm a lot more engaged. I'm a lot more... I I also have a better time with them.

AI-Therapist:
That difference you notice says a lot — it shows how much quality of presence matters to you, not just showing up. Do you think setting those short transition breaks is something you could keep as a regular ritual?

User:
No. That's good.

AI-Therapist:
Feels like you've found a small system that actually respects your limits between work and social time. What do you think helps you keep following through on it when life gets hectic again?

User:
That's a good question. maybe setting the alarm in the morning when I plan my day, and I know that that's gonna be a social interaction.

AI-Therapist:
That makes sense — folding that check-in into your morning routine sets up space for the day instead of hoping it happens on its own. What difference do you hope to feel once those deliberate pauses become a habit?

User:
Well, I I I think if I'm more... if I'm really engaged with other people and and kinda these these irrational feelings of loneliness are are less prevalent.

AI-Therapist:
That makes sense — when you can arrive to social moments with a bit more energy, those lonely thoughts lose some of their grip. The way you're linking self-care and connection shows real awareness of how intertwined they are. It might be worth keeping an eye on whether the same "refill before engaging" approach could help outside of relationships too, like before tougher stretches at work. How does that idea sound to you?

User:
Yeah. That's a good idea. Yeah.

AI-Therapist:
What would make following through on those planned pauses most realistic for you this week?

User:
Oh, I I think it's the same. You know, doing this in the morning and also kinda thinking about this when I plan my week on on Sundays.

AI-Therapist:
That kind of planning shows real self-awareness, Ben. Building those pauses right into your weekly routine could help you protect the energy you're trying to rebuild. As you think about next week, what would tell you that this small shift is actually working for you?

User:
I think I think when I look on my calendar and I realize that I did it, I think this would help me. Yeah.

AI-Therapist:
How will you remind yourself to stick to that plan once things get busy again?

User:
I guess checking in here with you. I could also send, um, a reminder on my phone to remind me in five days again.

AI-Therapist:
Maybe that reminder can act like a quick cue to check your mental energy level before going into something social or demanding. It ties nicely with what we talked about — the small, deliberate pauses that keep you from running on empty. Would you like to use a moment now to think about how this new routine fits into your overall goals for these sessions?

User:
Sure. Yeah. And then we can rev up.

AI-Therapist:
Actually, before we wrap up planning, I'd like to circle back to something from earlier. You said relationships were what felt most pressing, and we haven't really unpacked that part yet. When you asked this woman out and said she didn't respond well, what did she actually say in that moment?

User:
Well, she just just did not want to to see me again. Yeah.

AI-Therapist:
And when she said that, how did you make sense of it afterward?

User:
Well, I didn't. I guess I've been... I attributed it to her, like, kinda being in a maybe not emotionally stable spot right now.

AI-Therapist:
How did hearing that she didn't want to see you again affect how you view yourself in relationships since then?

User:
Well, definitely made me less confident. And, also, led me to to kinda question how I how I spend my relationships in general.

AI-Therapist:
When you say it made you question how you spend your relationships, what exactly have you been reevaluating or wondering about?

User:
I think it's sometimes if I, like, kind of invest too much time into people if they... yeah. And and not really thinking about if they're really worth it.

AI-Therapist:
Well, what you're describing sounds like you're starting to question whether your effort in relationships is really matched or appreciated. That can wear down confidence fast, especially if you often feel like you're the one doing most of the emotional work. How do you usually decide when it's time to keep investing in someone versus stepping back to protect your energy?

User:
Well, tomorrow, if you think, I guess, it's more like... and, I mean, I'm I'm... I don't say no often. I think... yeah.

AI-Therapist:
So it sounds like saying no feels uncomfortable for you, which might make it easy to overextend yourself. What do you think keeps you from saying no when you actually want to?

User:
I mean, I I I think two things. I do think that I'm quite capable of doing a lot of things, so it's not necessarily that I feel like I need to protect myself often. But, I mean, also, I have this belief kind of as really worthwhile helping other people because, first of all, you know, we human beings need to help each other. And second of all, it feels like if if I help a lot of people, then I can draw on their help as well. ,

AI-Therapist:
Well, Ben, I noticed you mentioned wrapping up earlier, and I don't want to push past what feels useful for you — how are you feeling about where we're at right now?

User:
So I'm really feeling good about this this idea of of of taking time before I do social engagements. And, um, I mean, I think it makes sense to to revisit, uh, my relationships in the next session.

AI-Therapist:
Glad that idea about pre-social pauses feels useful and that you'd like to explore relationships more deeply next time. You handled a lot of reflection today, and I'm looking forward to continuing this work with you soon; take care for now and goodbye.

B.8 Classification Workflow for Open-Ended Text Data

This appendix section describes the qualitative classification workflow we use to derive categories from open-ended text data. The procedure follows common workflows in the literature Chopra and Haaland (2023).

Step 1: Thematic clustering First, we apply BERTopic to sort the raw answers into clusters that share a common theme (Grootendorst, 2022). This approach leverages dense text embeddings rather than bag–of–words representations, reduces researcher degrees of freedom in feature choice and pre–processing, and uses a noise–aware clustering procedure that tends to produce semantically coherent topics.

Step 2: Review of clusters Second, we review the resulting thematic clusters one at a time. For each cluster, we retrieve all answers that the model assigns a sufficiently high probability of belonging to that cluster and read them in order of the model score. For every cluster, we write a short synopsis, draft candidate codes and themes, and record illustrative example responses, taking care to construct codes that are grounded in the interview data rather than in prior theoretical expectations.

Step 3: Convergence on coding scheme Third, after reviewing all clusters and compiling preliminary code lists, we consolidate the notes and resolve any ambiguities in the emerging structure. Themes are included in the final coding scheme only if they have a clear conceptual interpretation and are consistently supported by multiple responses. The resulting coding schemes used throughout the paper thus reflect a combination of machine–assisted thematic discovery and researcher judgment following the procedure in Chopra and Haaland (2023).

B.9 Information Treatment

To additionally study whether beliefs about AI efficacy affect trial take-up, we implemented a cross-randomized information treatment in Wave 1. Prior to deciding on participation, all respondents were introduced to a recent study on the accuracy of AI responses in a clinical context, based on Singhal et al. (2023). Participants first read the following description:

> In a study conducted in 2023, licensed clinicians judged whether answers from an artificial intelligence (AI) system, similar to ChatGPT, to standard medical questions were in line with the scientific consensus. **What percent of the AI's answers do you think clinicians judged as being in line with the scientific consensus?**

Participants were then asked to provide a percentage estimate. Immediately afterward, half of the respondents (randomly assigned) received the true study value of **92.6%**. This constitutes our information treatment, denoted by $Info_i = 1$ for treated individuals and 0 otherwise. The respondent's stated belief prior to receiving the information is denoted by $Prior_i$. Our primary outcome variable is an indicator for participation intention, measured at the end of the screen where respondents must choose one of the following options:

- **Yes,** I am ready and want to test the AI-powered app.
- **No,** I cannot commit right now and do not want to test the AI-powered app.

Following the preregistration, we restrict the sample to all participants who reached this decision screen and made a choice and estimate an intention-to-treat (ITT) specification with heteroskedasticity-robust standard errors:

$$Y_i = \beta_0 + \beta_1 Info_i + \beta_2(Prior_i - 92.6) + \varepsilon_i, \tag{S1}$$

where $Y_i = 1$ if participant i indicates willingness to participate. The parameter of interest is β_1, capturing the causal effect of correcting beliefs about AI accuracy on participation intentions.

Table S8 reports the results. We detect no statistically or economically meaningful effect of the information treatment on participation intentions. The estimated coefficient on $Info_i$ is virtually zero. The null result may be driven, however, not by a lack of treatment salience, but by the absence of meaningful scope for behavior to change: almost all participants express willingness to participate irrespective of treatment status. The constant in the regression is approximately **0.98**. With such a ceiling level of baseline demand, the information intervention had virtually no margin on which to operate. Additionally, our sample already consists of AI-open people, as we advertised the study as a study testing a new AI mental health app. Taken together, we think this specific setting was not ideal for studying the effect of trust in AI on take-up and do not consider the null results to be informative.

Table S8: Information Treatment Results

	Participation Intention
Constant	0.977***
	(0.003)
Information Treatment (1=Yes)	-0.005
	(0.004)
Prior AI Efficacy Estimate	0.000***
	(0.000)
N	10416
Adj. R^2	0.003

Notes: This table reports OLS estimates of treatment effects on participation intention. The dependent variable equals one if the participant indicates willingness to join the trial. Standard errors, reported in parentheses, are heteroskedasticity-robust. The information treatment consists of revealing the true percentage of AI answers judged to align with the scientific consensus. Prior beliefs are centered at the true value of the study.

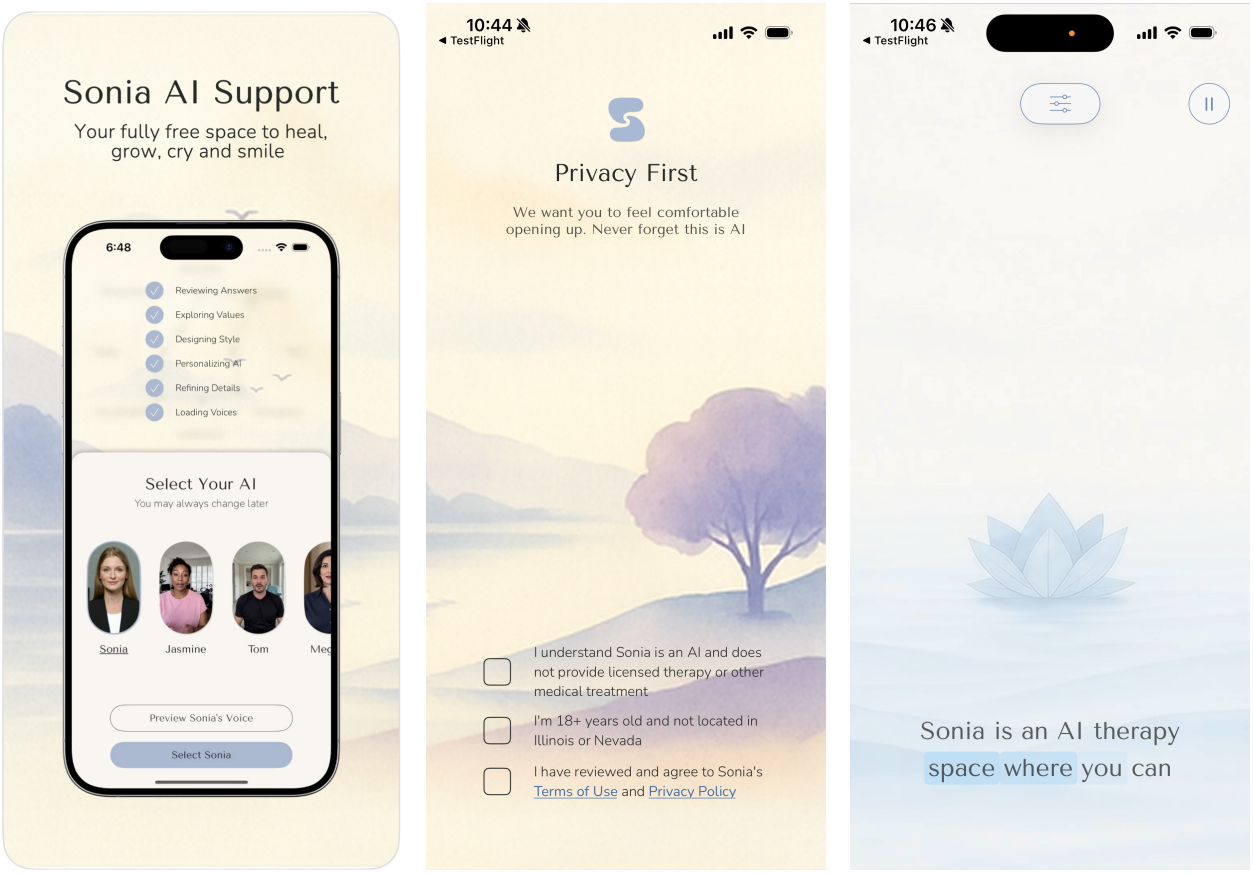
B.10 Clinical Outcomes

Table S9: Clinical outcomes for anxiety and depression at 2-week follow-up.

Outcome	Arm	Percentage (n/N)	Risk diff. vs control (%)	95% CI for risk diff.
a. GAD-7 outcomes				
Caseness (baseline)	Control	100.0%(200/200)	—	—
	Treatment	100.0%(200/200)	0.0	0.0%[0.0, 0.0]
Response ($\geq$ 50% $\downarrow$)	Control	21.0%(42/200)	—	—
	Treatment	48.5%(97/200)	27.5	27.5%[18.6, 36.4]
Remission (2 weeks)	Control	25.0%(50/200)	—	—
	Treatment	52.5%(105/200)	27.5	27.5%[18.3, 36.7]
b. PHQ-8 outcomes				
Caseness (baseline)	Control	84.0%(168/200)	—	—
	Treatment	86.0%(172/200)	2.0	2.0%[−5.0, 9.0]
Response ($\geq$50% $\downarrow$)	Control	10.5%(21/200)	—	—
	Treatment	25.5%(51/200)	15.0	15.0%[7.6, 22.4]
Remission (2 weeks)	Control	23.5%(47/200)	—	—
	Treatment	44.5%(89/200)	21.0	21.0%[11.9, 30.1]

Notes: Caseness is defined as GAD-7 $\geq$13 (anxiety) and PHQ-8 $\geq$10 (depression), corresponding to at least moderate symptom severity. Response is defined as $\geq$50% symptom reduction from baseline among participants meeting baseline caseness criteria. Remission is defined as GAD-7 $\leq$8 and PHQ-8 $\leq$8, indicating subclinical symptom levels. The treatment group (n=200) received access to Sonia, a voice-based AI cognitive behavioral therapy application; the control group (n=200) received psychoeducational materials. Risk differences and 95% confidence intervals (CIs) are estimated using intention-to-treat analysis with pre-specified baseline covariates (prior therapy experience and perceived probability that therapy would be helpful). All participants met GAD-7 caseness criteria by study design. CI = confidence interval; GAD-7 = Generalized Anxiety Disorder-7 scale; PHQ-8 = Patient Health Questionnaire-8.

C Additional Figures



(a) Personalization options (e.g., voice and interaction style) used to tailor the Sonia AI therapist.

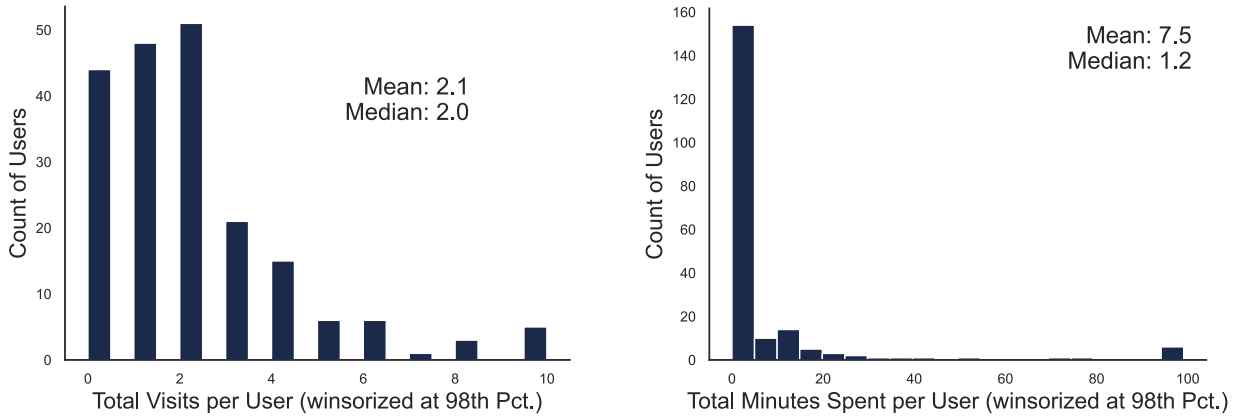
(b) Onboarding privacy notice/disclaimer displayed before first use.

(c) In-session interface during a therapy conversation, showing the minimal, voice-centered layout.

Figure S1: Representative screenshots from the Sonia voice-based AI CBT app illustrating key user-facing elements: personalization settings (left), onboarding privacy disclosure (middle), and the therapy-session interface (right).



Figure S2: Recruitment advertisement displayed on Facebook and Instagram.



(a) Total web app site visits per user (winsorized at the 98th percentile).

(b) Total minutes spent per user on web app (winsorized at the 98th percentile).

Figure S3: Engagement with the passive-control psychoeducation website.

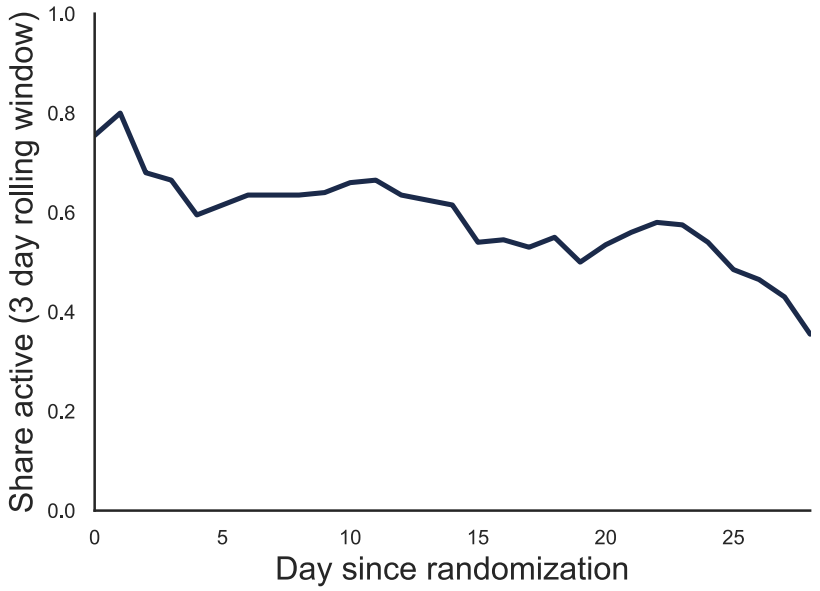


Figure S4: **Engagement over time in the AI-therapy treatment arm.** The curve shows, for each day since randomization (days 0–28), the share of participants with any recorded conversation with the AI therapist in a centered 3-day window (day $t-1$, t, $t+1$).

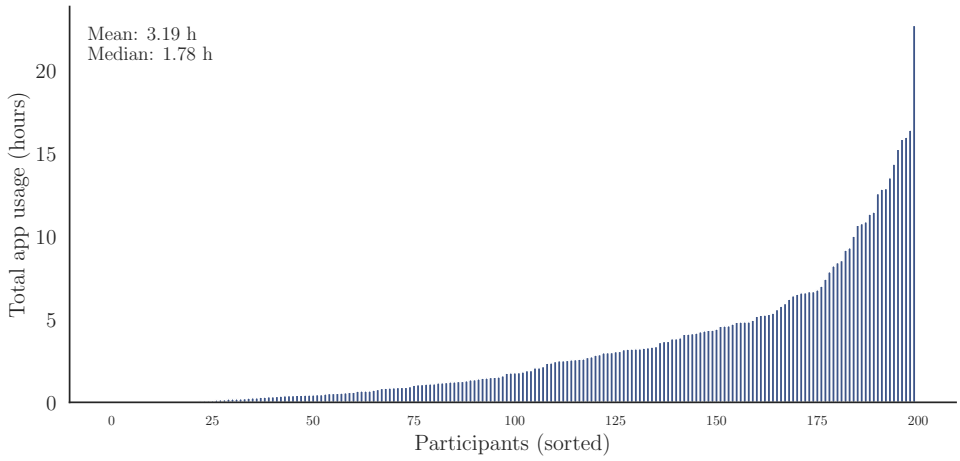


Figure S5: **Total conversation time per user in treatment group.** Total AI-therapist conversations in *Sonia* among participants assigned to the treatment group; each vertical bar represents one individual.

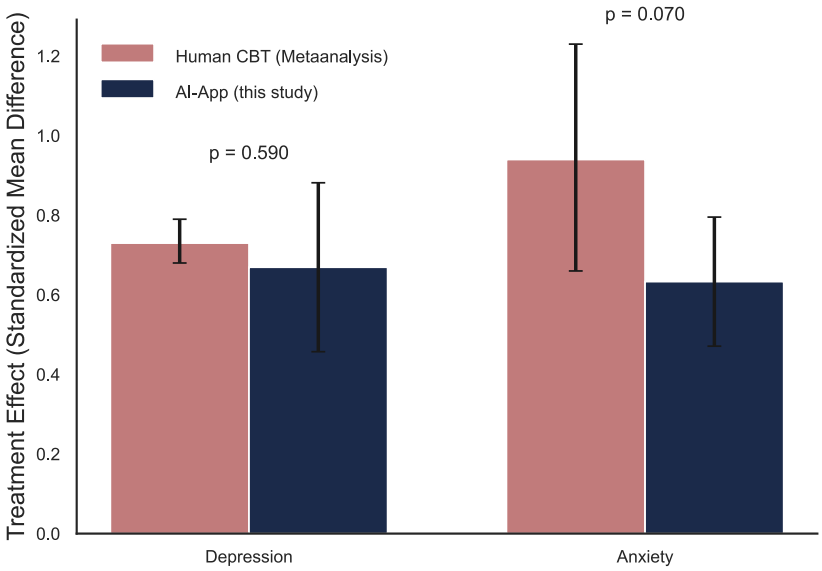
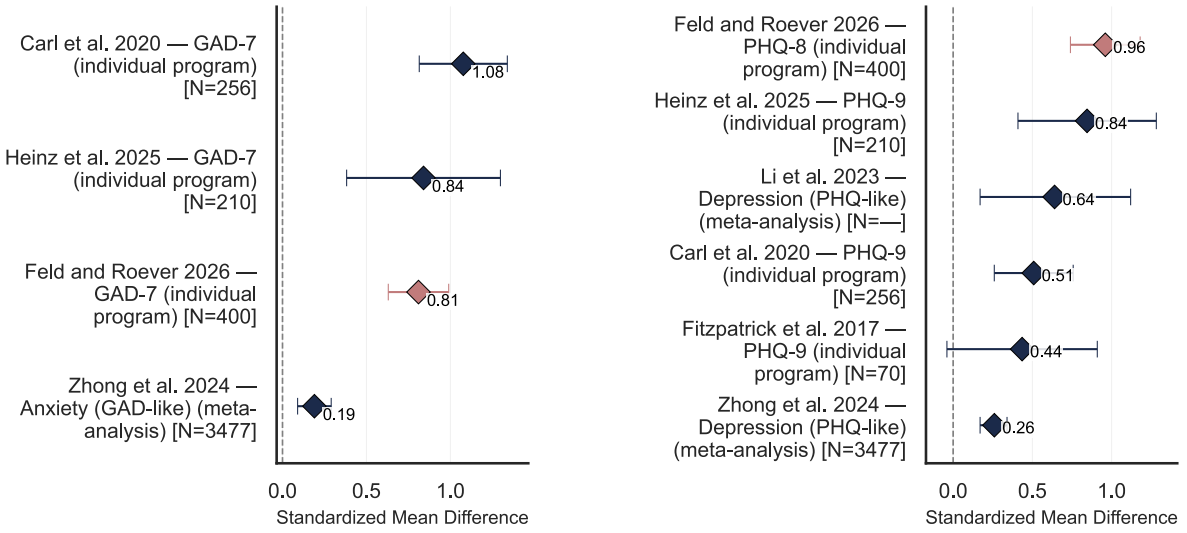


Figure S6: **Standardized treatment effects in the present study versus meta-analytic benchmarks from Harrer et al. (2022).** Points show standardized mean difference estimates with 95% confidence intervals (CIs). The 4-week standardized effect on anxiety symptoms (GAD-7) observed in the present study is plotted alongside a MetaPsy benchmark for individual-format CBT versus waitlist/placebo at 6–8 weeks (dataset v23.0.4; pooled $g = 0.94$, 95% CI 0.66–1.23; $k = 20$; 95% prediction interval -0.20 to 2.08; NNT = 3). For depression, the figure analogously includes a MetaPsy benchmark for individual-format CBT with 1–8 sessions (dataset v24.0.2; pooled $g = 0.73$, 95% CI 0.68–0.79; $k = 577$; 95% prediction interval -0.50 to 1.97; NNT = 4).



(a) Effect sizes for generalized anxiety disorder (GAD) outcomes. We use the 4-week endpoint from Heinz et al. (2025) for reference. For this study (Feld and Roever, 2026), we plot the four week outcome.

(b) Effect sizes for depression outcomes. We use the 4-week endpoint from Heinz et al. Heinz et al. (2025) for reference. For this study (Feld and Roever, 2026), we plot the four week outcome.

Figure S7: Benchmarks for Effect Sizes

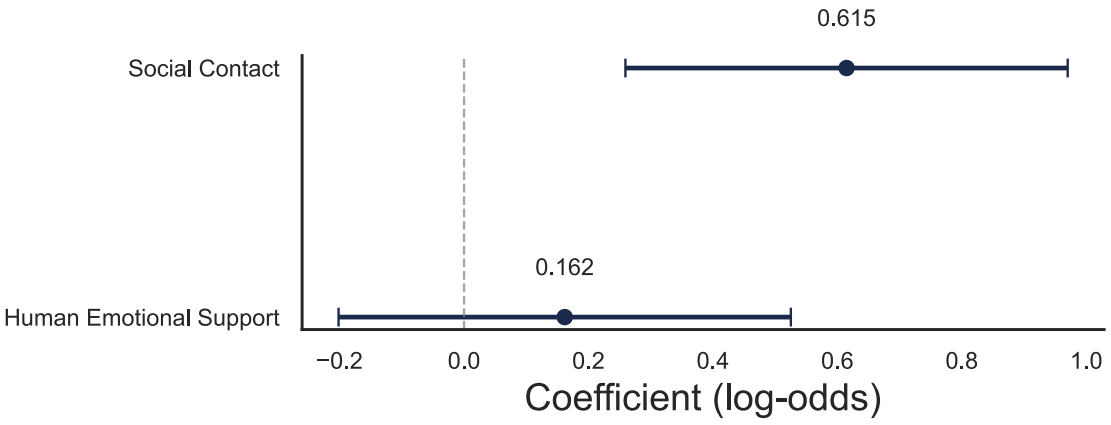


Figure S8: **Treatment effects on social outcomes at Wave 5.** Coefficients are from ordered logistic regressions of social contact frequency and perceived human emotional support on treatment assignment, controlling for baseline therapy history and efficacy of human therapy. Points denote estimated treatment effects (log-odds) and whiskers indicate 95% confidence intervals.

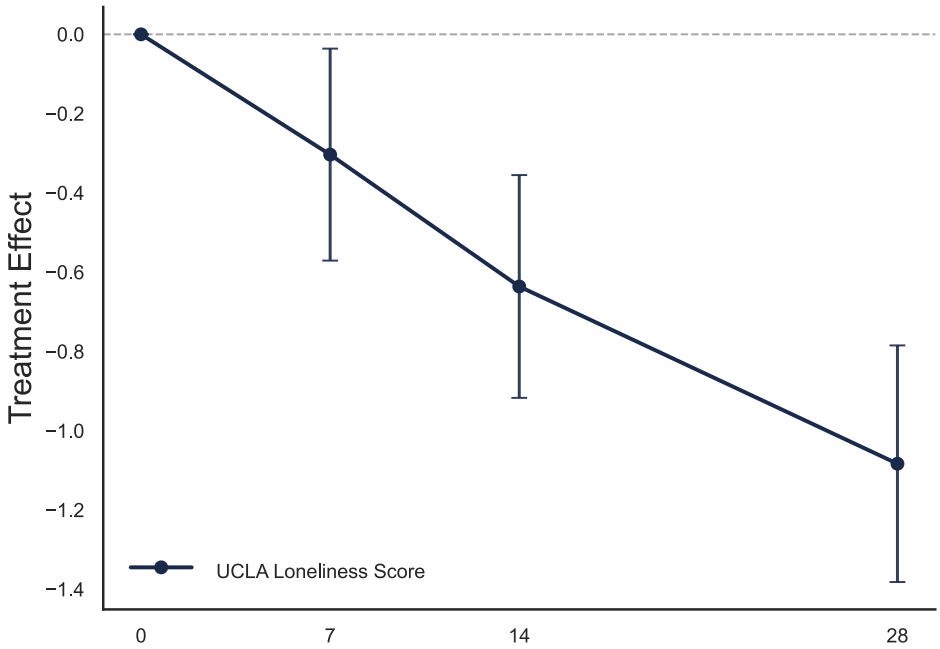


Figure S9: **Treatment effects on loneliness (UCLA Loneliness scale).** Points plot estimated intention-to-treat effects of assignment to Sonia on UCLA Loneliness at each follow-up wave, where the dependent variable is the wave-specific loneliness score. Estimates are obtained from ANCOVA regressions controlling for baseline UCLA Loneliness, prior therapy history, and baseline beliefs about the efficacy of human-delivered therapy; vertical lines denote 95% confidence intervals based on heteroskedasticity-robust (HC3) standard errors. Negative values indicate lower loneliness relative to the control group.

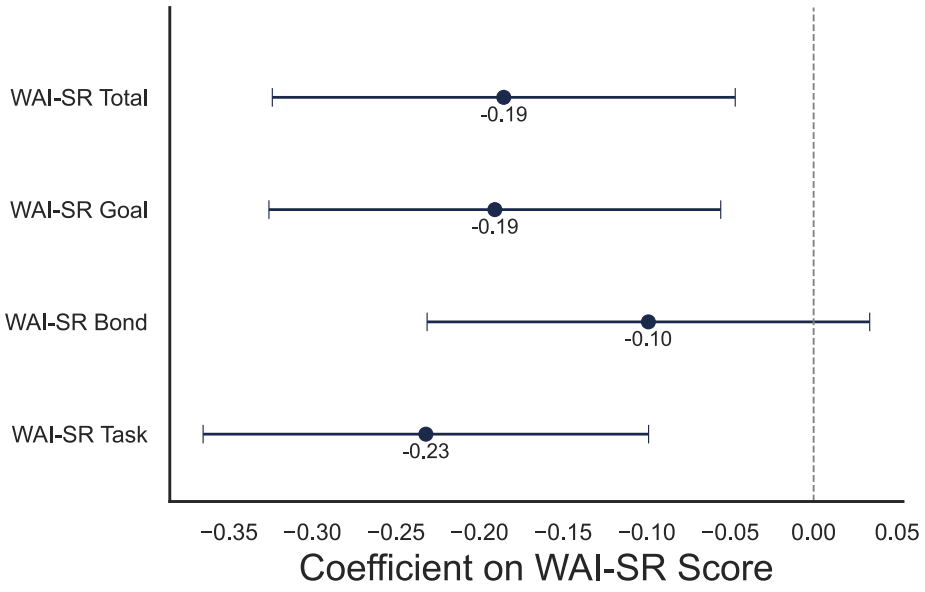


Figure S10: **WAI-SR heterogeneity in mental health outcomes (wave 4).** Points show OLS coefficients for standardized WAI-SR total and subscale scores (goal, bond, task) predicting wave-4 mental health, controlling for baseline mental health. Mental health is measured using the mean of standardized GAD-7 and PHQ-8 values. Error bars are 95% confidence intervals; standard errors are heteroskedasticity robust.

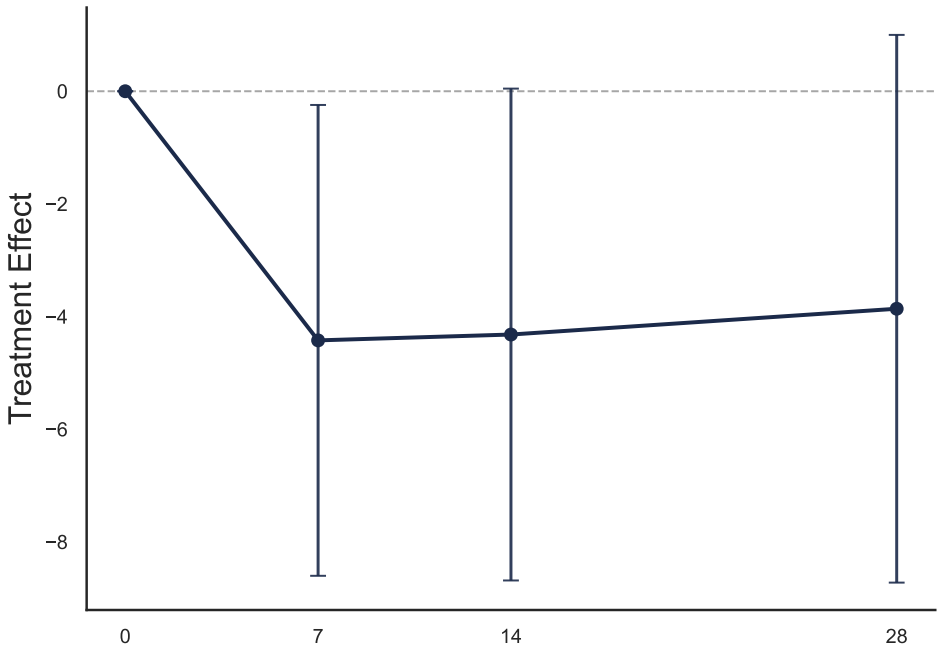
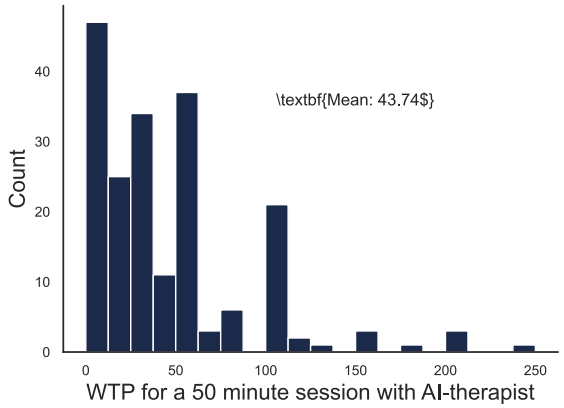
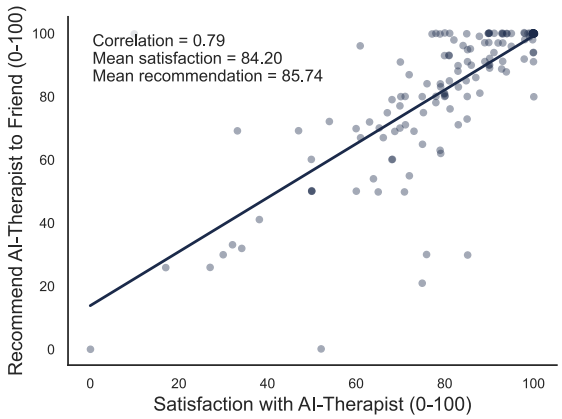


Figure S11: **Treatment effects on beliefs about the efficacy of human-delivered therapy.** The figure plots regression coefficients (treatment–control differences) for the outcome "How likely is your mental health to improve after a session with a human therapist?" measured on a 0–100 scale across follow-up waves on day 7, 14 and 28. Estimates come from ANCOVA specifications that control for the respondent's baseline belief and an indicator for prior therapy experience; standard errors are clustered at the individual level and error bars show 95% confidence intervals.



(a) **Willingness to pay for 50 minute session with AI-therapist among treated participants.** Distribution of stated willingness to pay, with the sample mean annotated.



(b) **Satisfaction and recommendation (wave 3).** Jittered scatter of satisfaction levels (0-100) with AI therapist versus probability of recommendation to friend or family member (0-100) with fitted regression line; panel reports correlation and sample means.

Figure S12: **User valuation and endorsement of AI therapist.** Panels summarize willingness to pay and the association between satisfaction and recommendation in wave 3.

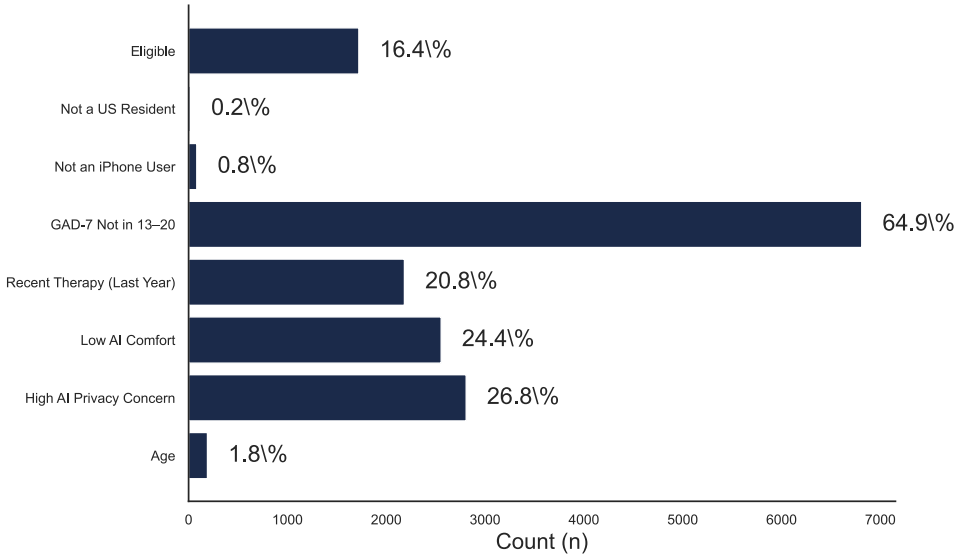


Figure S13: This figure reports screening failures among all $10,490$ participants completing the wave 1 baseline questionnaire, administered immediately following advertisement click-through. Because failure categories are nonexclusive, cumulative failure percentages exceed 100%.

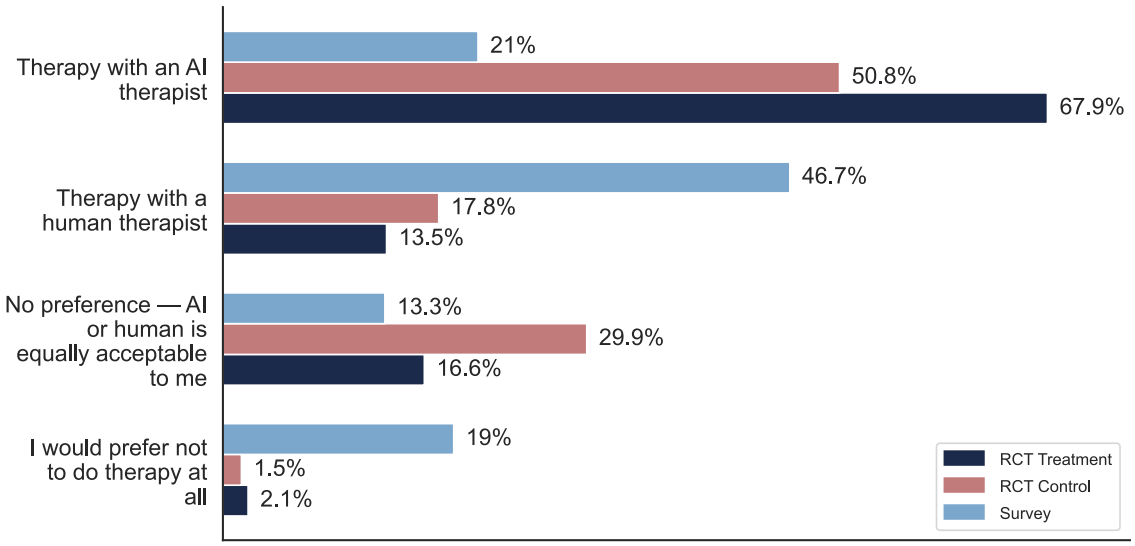


Figure S14: **Overall preference for AI versus human therapy in the RCT and survey samples.** Bars show the distribution of stated preferences for AI therapy versus human therapy across the RCT sample (elicited in wave 4) and the supplementary survey sample from Prolific; percentages are computed within each sample.

D Additional Tables

Table S1: Treatment Effect Heterogeneity

	Dependent Variable: Standardized GAD-7 Score in Wave 4					
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.63***	-0.56***	-0.48***	-0.55***	-0.67***	-0.71***
	(0.08)	(0.11)	(0.11)	(0.11)	(0.10)	(0.11)
Treatment ×						
High Usage	-0.00***					
	(0.00)					
Treatment ×						
High Motivation (Wave 3)		-0.14				
		(0.17)				
Treatment ×						
High AI Correctness Belief (Wave 1)			-0.37**			
			(0.17)			
Treatment ×						
High Comfort with AI Emotional Support (Wave 1)				-0.16		
				(0.16)		
Treatment ×						
High Belief in Human Therapy Efficacy (Wave 2)					0.10	
					(0.17)	
Treatment ×						
High Mental Health Stigma (Wave 2)						0.13
						(0.16)
Baseline Controls	✓	✓	✓	✓	✓	✓
Obs	390	390	390	390	390	390
Adj. R^2	0.22	0.22	0.23	0.22	0.23	0.23

Notes: The dependent variable is standardized GAD-7 score at Wave 4. Each column reports OLS estimates with HC3 heteroskedasticity-robust standard errors in parentheses. Treatment is an indicator for assignment to Sonia. Heterogeneity is tested via interactions between treatment and baseline beliefs/actions: above-median app usage (total hours across duration columns), motivation for mental health (W3), AI efficacy estimate (W1), AI emotional comfort (W1), efficacy of human therapy (W2), and stigma (W2; high = somewhat/strongly agree). All models control for baseline GAD-7 score (standardized, W1), therapy history (W1), and efficacy of human therapy (W2). N and adjusted R^2 are reported. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table S2: Treatment Effect Heterogeneity (continued)

	Dependent Variable: Standardized GAD-7 Score in Wave 4								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	-0.58***	-0.52***	-0.50***	-0.58***	-0.73***	-0.54***	-0.67***	-0.57***	-0.70***
	(0.10)	(0.10)	(0.11)	(0.10)	(0.10)	(0.10)	(0.09)	(0.17)	(0.10)
Treatment ×									
High GAD-7 (Wave 1)	-0.12								
	(0.17)								
Treatment ×									
High PHQ-8 (Wave 2)		-0.26*							
		(0.16)							
Treatment ×									
High Baseline Mental Health Score			-0.27*						
			(0.16)						
Treatment ×									
High ATQ (Wave 2)				-0.15					
				(0.16)					
Treatment ×									
High Behavioral Activation (Wave 2)					0.25				
					(0.17)				
Treatment ×									
High Cognitive Bias (Wave 2)						-0.33*			
						(0.17)			
Treatment ×									
No Prior Therapy Experience							0.13		
							(0.20)		
Treatment ×									
Ever Taken Medication								-0.11	
								(0.19)	
Treatment ×									
High Loneliness (Wave 2)									0.15
									(0.17)
Baseline Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
N	390	390	390	390	390	390	390	390	390
Adj. R^2	0.22	0.28	0.27	0.26	0.26	0.25	0.22	0.22	0.26

Notes: The dependent variable is standardized GAD-7 score at Wave 4. Each column reports OLS estimates with HC3 heteroskedasticity-robust standard errors in parentheses. Treatment is an indicator for assignment to the AI therapy application. Heterogeneity is tested via interactions between treatment and baseline characteristics: above-median mental health severity (W1), GAD-7 (W1), PHQ-8 (W2), ATQ (W2), behavioral activation (W2), cognitive distortion (W2), loneliness (W2), therapy history (W1), and medication status (W2; ever medicated = past or current use). All models control for baseline GAD-7 score (standardized, W1), therapy history (W1), and efficacy of human therapy (W2). N and adjusted R^2 are reported. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

Table S3: AI vs. Human Therapy Themes

Evaluating Factor	**Illustrative Quotes**
Fear of Judgement / Embarrassment	"The AI system I feel doesn't judge you."
	"AI because they aren't anyone to judge."
Value of Human Connection	"Sincerity, personal experiences that can inform their counsel, and the importance of eye-to-eye contact."
	"Yes, a human therapist would be better to actually be able to see a face and receive feedback possibly coming from an actual human experience."
Empathy and True Understanding	"An actual human would naturally be better because of their natural responses in terms of empathy."
Availability	"AI is better cuz it has endless time to spend with and talk to."
	"I like having the AI because I have access to it 24/7 compared to a human you see once or twice in a week."
Indifference	"Either of them is fine."
Don't Know	"Unsure."
Neutrality and Bias	"I would say I don't have a lot of trust in human therapists; a lot of them are biased and go into therapy for the wrong reasons."
	"I do feel like an AI system would be less likely to impart biased information than a human therapist would based on the therapist's belief system."
Perceived Competence	"I feel like AI would be less judgmental and more research and fact based."
	"Maybe an AI, because I use AI tools and they seem to know everything."
Trust and Privacy	"I feel like I can't trust people and people judge you so it may be easier dealing with AI."
	"The AI system is better because it won't judge you nor talk bad about you behind your back."
	"I am not sure. Both have their pros and cons! I like the flexibility of AI but it also feels less secure."
Social Avoidance	"Yes, I would rather have the AI therapist because I would feel more comfortable to open up to it rather than a real person."

Notes: This table shows extracted themes obtained from answers to our Wave 3 survey question, where we ask participants to explain whether they prefer AI-led or human therapy. The workflow for extracting the themes is described in Section B.8.

E Preregistration

AI Emotional Support Waitlist Trial (#258203)

Author(s)
This pre-registration is currently anonymous to enable blind peer-review.
It has 2 authors.

Pre-registered on:
2025/11/14 10:25 (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
We conduct a randomized controlled trial (RCT) to study an AI-based emotional support app. Our focus is on three main areas: (1) changes in anxiety levels, (2) user satisfaction and engagement, and (3) safety. We use a 5-wave survey design: Wave 1 is a screener survey. We assign treatment in Wave 2. Waves 3 (one week after Wave 2), 4 (two weeks after Wave 2), and 5 (four weeks after Wave 2) measure treatment effects.

3) Describe the key dependent variable(s) specifying how they will be measured.
Our primary outcome measures are:

- Changes in standardized GAD-7 after 2 weeks (in Wave 4). Standardization is done using mean and standard deviation from the full baseline data, which we collect in Wave 1.
- A self-report of safety, elicited through an open-ended question in Wave 4.
- Engagement metrics: Number of sessions, number of exchanges in conversations, number of active days, total usage duration.
- User satisfaction metrics: Self-report of satisfaction, willingness to pay for the AI-based emotional support app and willingness to recommend to a friend or family member.

4) How many and which conditions will participants be assigned to?
We employ a two-arm treatment design: (1) The treatment group is incentivised to download and use the AI-based emotional support app. (2) The waitlist group is informed that they will gain access to the AI-based emotional support app after four weeks. In the meantime, they receive access to a web app providing AI-generated educational content on stress management and mental health.

Randomization occurs within Qualtrics during the Wave 2 survey, assigning 50% of participants to each group.

We additionally cross-randomize an information treatment in Wave 1 (see secondary outcomes): 50% of participants see information about the quality of AI-generated health information; the other 50% do not.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will estimate treatment effects using ordinary least squares (OLS) regressions with robust standard errors clustered at the participant level. Our main specification is:

$$Y_i = \beta_0 + \beta_1 AIApp_i + \beta_2 Y_{i,baseline} + \beta_3 X_i + \varepsilon_i,$$

where:

- Y_i is the outcome variable for participant i in a given follow-up wave.
- $Y_{i,baseline}$ is the baseline value of the outcome. For GAD-7, this is elicited in Wave 1; for other outcomes, it refers to the Wave 2 measure.
- $AIApp_i$ equals 1 for participants assigned to the AI-based emotional support app, and 0 for those in the waitlist control group.
- X_i is a vector of baseline controls, including motivation to work on mental health (0–100 scale, Wave 2), prior therapy experience (dummy), and perceived probability that therapy would improve one's mental health (0–100 scale, Wave 2).

Our primary coefficient of interest is β_1. We will use this specification to estimate treatment effects on GAD-7 (main outcome), as well as PHQ-8, a composite mental health score, and mechanisms across all follow-up waves.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
All our primary outcome measures are naturally bounded, so we do not need to pre-register specific outlier handling.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
We recruit participants on Facebook and Instagram through advertisements, where participants are guided to our Wave 1 survey. Participants can complete this survey and are invited to a phone call if they: (1) take the survey on an iPhone running iOS 18 or newer, (2) report using an iPhone as their primary phone, (3) are at least 18 years old, (4) are located in the US when taking the survey and a current US resident, (5) are not currently in therapy, (6) do not self-report being very or extremely concerned when turning to AI for emotional support, (7) report a comfort level of at least 50 (scale 0-100) when asked about turning to AI for emotional support, (8) have a GAD-7 score between 13 and 20. To ensure data quality and prevent bot entries, we will exclude participants who do not meet minimum engagement criteria during the baseline survey (i.e. extremely short completion times inconsistent with attentive participation).

In the phone call, we explain the trial length to participants and confirm that they have sufficient time. We also check that they do not currently have suicidal ideation and do not reside in Nevada or Illinois. People who fulfill these criteria are invited via email to the Wave 2 survey.

All participants who are assigned treatment (i.e. who reach the page where they are instructed to download the respective app) are part of the sample and will be invited to all follow-up surveys.

Based on power calculations, we recruit until we have a sample of 400 participants.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Our secondary outcomes and analyses:

- Changes in willingness to pay for human-led and AI-based emotional support
- A composite mental health score. It is obtained by standardizing PHQ-8 and GAD-7 scores separately using mean and standard deviation in the full baseline sample collected in Wave 2 for PHQ-8 and in Wave 1 for GAD-7, and then summing the standardized PHQ-8 and GAD-7 scores.
- Changes in standardized PHQ-8; standardization is done using the full sample from Wave 2.
- Changes in beliefs in the effectiveness of human-led and AI-based emotional support and psychiatric medication
- The inclination of participants to use psychiatric medication reported on a 1-100 scale and the interest in human-based therapy reported on a 5-point Likert scale as well as whether participants click a link to a teletherapy provider.
- Effect of trust in AI on take-up. We cross-randomize an information treatment in Wave 1; half of the sample receives information that health information has been deemed rather accurate by medical professionals in a study.

The primary specification for the analysis is given by a standard ITT regression with heteroskedasticity robust standard errors:

$Y_i = \beta_0 + \beta_1 Info_i + \beta_2 (Prior_i - 92.6) + \varepsilon_i,$

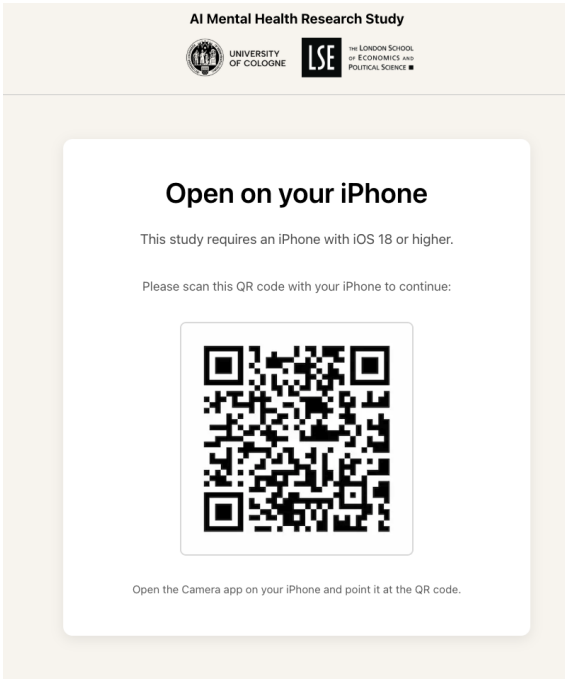
where Y_i equals 1 if the participant indicates willingness to participate, $Info_i$ is an indicator of whether the person has seen the actual information or not, and $Prior_i$ is the prior belief of person i. Primary parameter of interest is β_1; any other potential analyses of heterogeneity and using more downstream measures are considered exploratory.

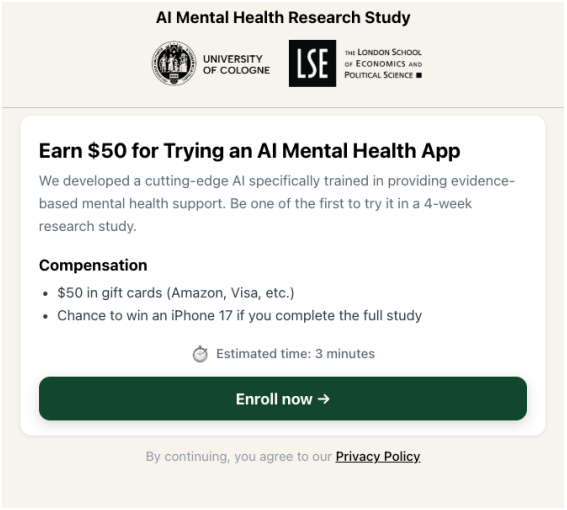
The relevant sample for this analysis comprises all participants who get to the trial information page and click on either the button that they are willing to participate or that they are not willing to participate.

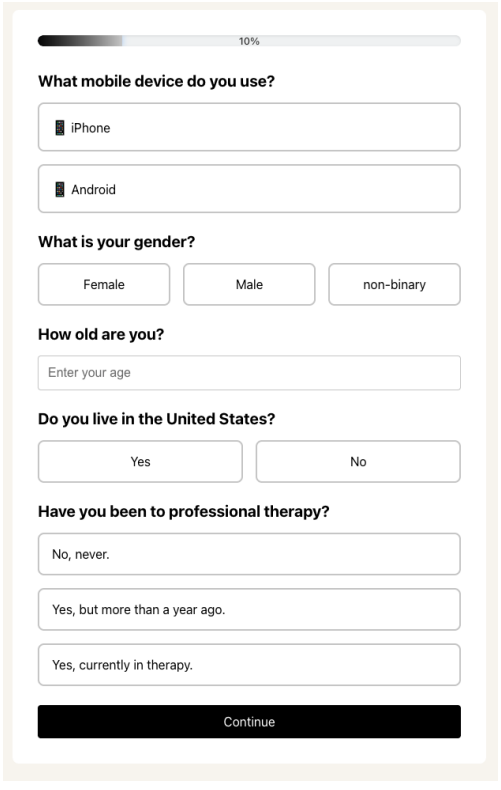
- Mechanisms of the app (automatic thoughts, behavioral activation, cognitive biases) all measured by scales elicited in all waves
- We also test heterogeneity of treatment effects by mechanisms, cross-randomization, and engagement, although we acknowledge that we might be underpowered.
- Estimation of Complier Average Causal Effect (CACE)
- Effects on loneliness, as measured by the 3-item UCLA Loneliness Scale
- Predicted hypotheses of the study and social desirability beliefs on mental health self-reports to analyze the size of experimenter demand effects and social desirability bias.

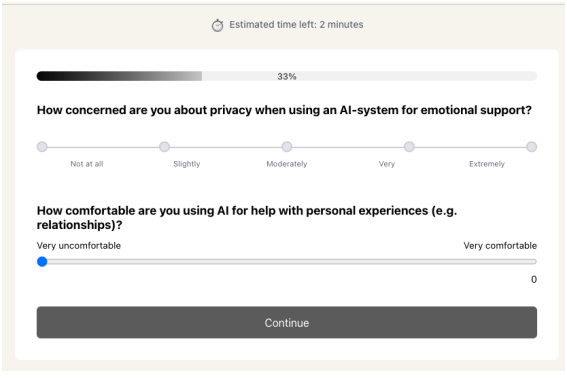
F Survey Instructions

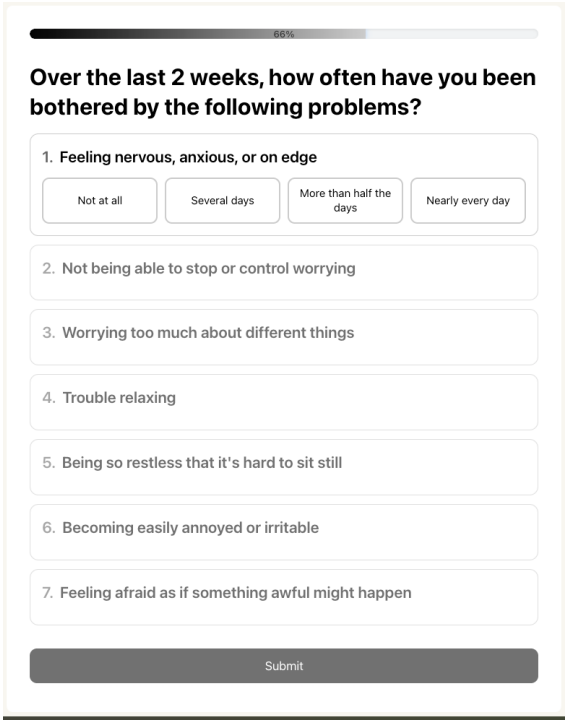
F.1 RCT Survey Instructions Wave 1

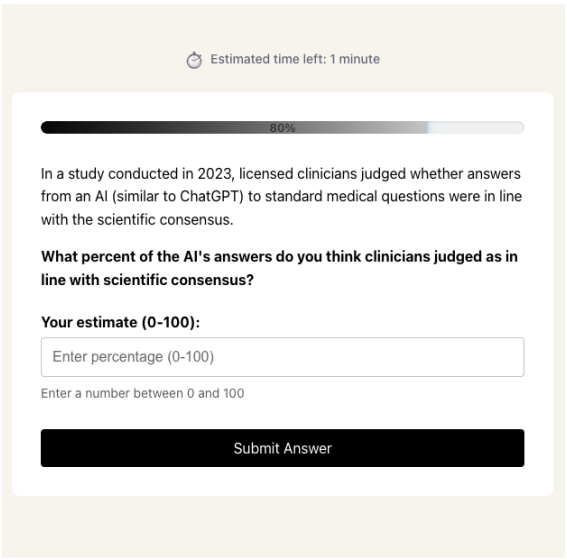


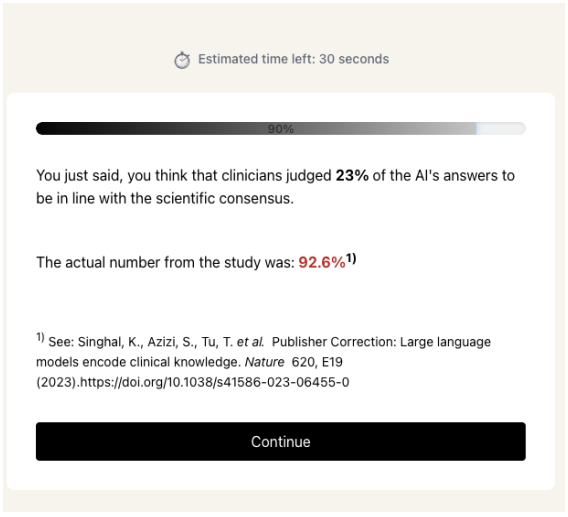


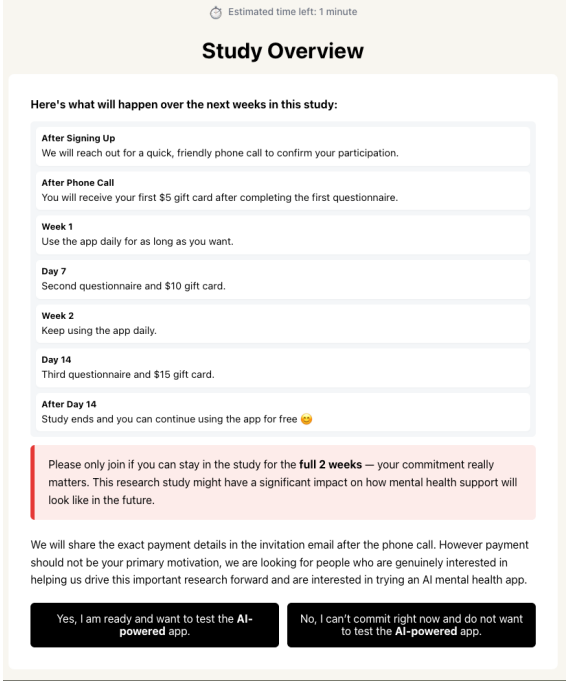





66%
Over the last 2 weeks, how often have you been bothered by the following problems?
1. Feeling nervous, anxious, or on edge
Not at all
Several days
More than half the days
Nearly every day
2. Not being able to stop or control worrying
3. Worrying too much about different things
4. Trouble relaxing
5. Being so restless that it's hard to sit still
6. Becoming easily annoyed or irritable
7. Feeling afraid as if something awful might happen
Submit


Estimated time left: 1 minute
80%
In a study conducted in 2023, licensed clinicians judged whether answers from an AI (similar to ChatGPT) to standard medical questions were in line with the scientific consensus.
What percent of the AI's answers do you think clinicians judged as in line with scientific consensus?
Your estimate (0-100):
Enter percentage (0-100)
Enter a number between 0 and 100
Submit Answer


Estimated time left: 30 seconds
90%
You just said, you think that clinicians judged 23% of the AI's answers to be in line with the scientific consensus.
The actual number from the study was: 92.6%1)
1) See: Singhal, K., Azizi, S., Tu, T. et al. Publisher Correction: Large language models encode clinical knowledge. Nature 620, E19 (2023).https://doi.org/10.1038/s41586-023-06455-0
Continue


Estimated time left: 1 minute
Study Overview
Here's what will happen over the next weeks in this study:
After Signing Up
We will reach out for a quick, friendly phone call to confirm your participation.
After Phone Call
You will receive your first $5 gift card after completing the first questionnaire.
Week 1
Use the app daily for as long as you want.
Day 7
Second questionnaire and $10 gift card.
Week 2
Keep using the app daily.
Day 14
Third questionnaire and $15 gift card.
After Day 14
Study ends and you can continue using the app for free
Please only join if you can stay in the study for the full 2 weeks — your commitment really matters. This research study might have a significant impact on how mental health support will look like in the future.
We will share the exact payment details in the invitation email after the phone call. However payment should not be your primary motivation, we are looking for people who are genuinely interested in helping us drive this important research forward and are interested in trying an AI mental health app.
Yes, I am ready and want to test the AI-powered app.
No, I can't commit right now and do not want to test the AI-powered app.

F.2 RCT Survey Instructions Wave 2

Consent

Do you consent to the processing of your data and our privacy policy?

▶ *Click **here** to display the full privacy policy.*

Declaration of Consent

I hereby consent to the processing of my personal data. I can withdraw my consent at any time. I agree that my consent also extends to special categories of personal data.

- ○ I have read the above and consent to take part in this study
- ○ I do not wish to participate

PHQ_8

Over the last 2 weeks, how often have you been bothered by any of the following problems?

	Not at all	Several days	More than half the days	Nearly every day
Little interest or pleasure in doing things	○	○	○	○
Feeling down, depressed, or hopeless	○	○	○	○
Trouble falling or staying asleep, or sleeping too much	○	○	○	○
Feeling tired or having little energy	○	○	○	○

	Not at all	Several days	More than half the days	Nearly every day
Poor appetite or overeating	○	○	○	○
Feeling bad about yourself, or that you are a failure, or have let yourself or your family down	○	○	○	○
Trouble concentrating on things, such as reading the newspaper or watching television	○	○	○	○

	Not at all	Several days	More than half the days	Nearly every day
Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	○	○	○	○

mechanisms_therapy v002

Please indicate how frequently you have had each of the following thoughts in the past two weeks.

	Not at all	Sometimes	About half the time	Most of the time	All the time
I'm no good.	○	○	○	○	○
I'm so disappointed in myself.	○	○	○	○	○
What's wrong with me?	○	○	○	○	○
I'm worthless.	○	○	○	○	○
I feel so helpless.	○	○	○	○	○
Something has to change.	○	○	○	○	○
My future is bleak.	○	○	○	○	○
I can't finish anything.	○	○	○	○	○

During the past two week, how often did you do something you genuinely enjoyed or found meaningful?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

During the past two weeks, how often did you put off or skip activities because you felt low-energy, down, or had negative thoughts?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time

- ○ 60–80% of the time
- ○ Over 80% of the time

When something unclear or unexpected happened the past two weeks, how often did you immediately expect it to turn out badly?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

How often did you notice your attention being pulled toward potential threats, dangers, or negative cues, rather than neutral or positive information during the past two weeks?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

What is the maximum amount that you would be willing to pay for a 50 minute **in person** therapy session with a licensed psychologist?

Amount in $ [ ]

If you did **in person therapy** with a licensed psychologist, what do you think is the probability that your mental health would improve?

Percent (0-100): [ ]

Do you agree or disagree with the following statement?

"Most people will judge me as weak if I admit I have mental health problems."

- ○ Strongly Agree
- ○ Somewhat Agree
- ○ Neutral
- ○ Somewhat disagree
- ○ Strongly disagree

Please indicate your experience with **prescribed** mental health medication (for example Prozac, Xanax, Seroquel):

- ○ [ ] I am currently taking medication, please tell us which:
- ○ I have taken medication in the past, but not currently
- ○ I have never taken medication

Loneliness

Please answer the following questions:

	Hardly ever	Some of the time	Often
How often do you feel that you lack companionship?	○	○	○
How often do you feel left out?	○	○	○
How often do you feel isolated from others?	○	○	○

Trust in AI

How effective do you think psychotherapy fully delivered by an AI-system is **for you personally?**

Not at all effective — Very effective

0 10 20 30 40 50 60 70 80 90 100 [ ]

How often do you use large language models (like ChatGPT, Google Gemini) and other AI tools?

- ○ Every day
- ○ Multiple times per week
- ○ A few times a month
- ○ Once a month
- ○ Almost never

Demographics

Choose one or more races that you consider yourself to be

- ☐ White or Caucasian
- ☐ Black or African American
- ☐ American Indian/Native American or Alaska Native
- ☐ Asian
- ☐ Native Hawaiian or Other Pacific Islander

- ☐ Other
- ☐ Prefer not to say

What is the highest level of education you have completed?

- ○ Some high school or less
- ○ High school diploma or GED
- ○ Some college, but no degree
- ○ Associates or technical degree
- ○ Bachelor's degree
- ○ Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS etc.)
- ○ Prefer not to say

What was your annual household income in 2024 in US dollars **before taxes and deductions**?

Note: The household income is the total amount of money earned by every member of your household.

- ○ Less than $15,000
- ○ Between $15,000 and $25,000
- ○ Between $25,000 and $50,000
- ○ Between $50,000 and $75,000
- ○ Between $75,000 and $100,000
- ○ Between $100,000 and $150,000
- ○ Between $150,000 and $200,000
- ○ More than $200,000

What best describes your employment status over the last three months?

- ○ Working full-time
- ○ Working part-time
- ○ Unemployed and looking for work
- ○ A homemaker or stay-at-home parent
- ○ Student
- ○ Retired
- ○ Other

What is your current marital status?

- ○ Married
- ○ Living with a partner
- ○ Widowed
- ○ Divorced/Separated
- ○ Never been married

T3: Sonia

Test an AI Powered Mental Health App

We would like to invite you to try **Sonia**, an AI-powered app designed to help improve well-being by relying on research-based methods.

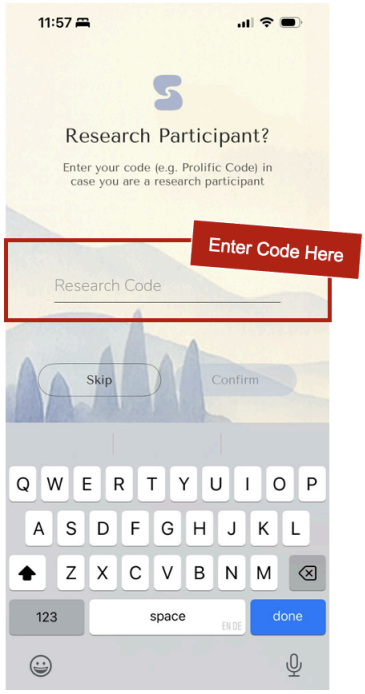
Follow these three steps **now:**

1 Download the app.

Link to App Download

(Link opens in a new tab.)

2 Open the app and enter your personal participant code **${e://Field/participant_code}**



3 Return to this page and continue the survey.

- ○ **Yes,** I want to participate in the trial and have downloaded the app
- ○ **No**, I do not wish to participate

T1: Stress Info NEW

You've been assigned to the "waiting period" group

As part of this study, some participants **— including you —** are randomly assigned to a **waiting period group**. This means you'll be able to **start using the main AI-based anxiety app in 4 weeks**.

This waiting period is essential: it allows us to **compare mental health changes** between participants who use the app immediately and those who start later, helping us understand the app's **true effects**.

During the waiting period of 4 weeks:

- You'll complete surveys and receive **the same compensation** as all other participants.
- You'll get access to **educational mental health material** we developed just for this phase and made available through the website link below.

Thank you for your important contribution to this research!

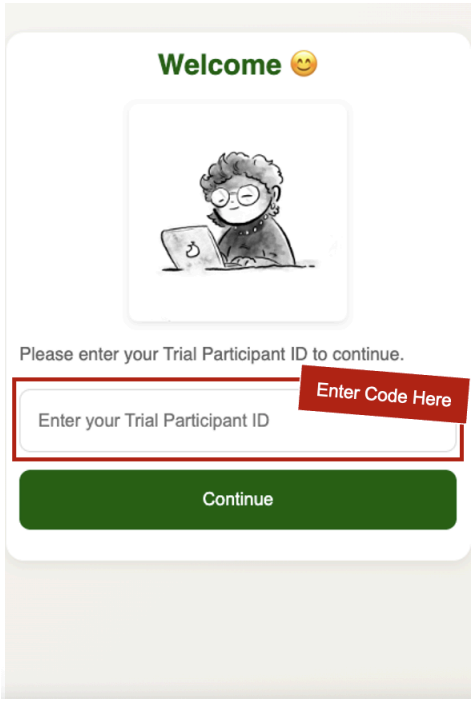
To access the resources you can use during the waiting period, follow these three steps:

1 Access the resources.

Link to Ressources

(Link opens in a new tab.)

2 Enter your personal participant code **${e://Field/participant_code}**



3 Return to this page and continue the survey.

- ○ **Yes,** I want to participate in the trial and have downloaded the app
- ○ **No**, I do not wish to participate

End of Survey 3 treatments

Thank you for participating! 🎉

We'll invite you to a follow-up survey in about 1 week.

We have just sent you an email with all relevant information.

Remember to use the resources we developed for you! 😊

Your Upcoming Surveys:

- **1 Week Survey : $10** for completing the first follow-up survey in 1 week
- **2-Week Survey: $15** for completing the follow-up survey **in 2 weeks**
- **Final Survey: $20** for completing the final survey at the end of the study **in 4 weeks**

Reminder: If you haven't accessed the waiting resources yet, please do it now using **this link**! You may continue using these resources throughout the study period.

Click next to submit your response.

Questions? Feel free to contact us at: team@mental-health-research.com

Thank you for participating! 🎉

We'll invite you to a follow-up survey in about 1 week.
We have just sent you an email with all relevant information.

Remember to use the app Sonia! 😊

Your Bonus Payments for Using Sonia:

- **Start Bonus:** You get **$5** for completing your first session of at least 10 minutes in the app

Your Upcoming Surveys:

- **1 Week Survey: $10** for using the app daily and completing this first follow-up survey
- **2-Week Survey: $15** for completing the follow-up survey **in 2 weeks**

- **Final Survey: $20** for completing the final survey at the end of the study **in 4 weeks**

Reminder: Please do your first 10 minute session in the Sonia app *right now*! Continue using the app throughout the study period.

Click next to submit your response.

Questions? Feel free to contact us at: team@mental-health-research.com

Powered by Qualtrics

F.3 RCT Survey Instructions Wave 3

Consent_Sonia_Trial

Do you consent to the processing of your data and our privacy policy?

▶ *Click **here** to display the full privacy policy.*

Declaration of Consent

I hereby consent to the processing of my personal data. I can withdraw my consent at any time. I agree that my consent also extends to special categories of personal data.

- ○ I have read the above and consent to take part in this study
- ○ I do not wish to participate

Mental GAD7 PHQ8

Over the last two weeks, how often have you been bothered by the following problems?

	Not at all	Several days	More than half the days	Nearly every day
Feeling nervous, anxious, or on edge	○	○	○	○
Not being able to stop or control worrying	○	○	○	○
Worrying too much about different things	○	○	○	○
Trouble relaxing	○	○	○	○

	Not at all	Several days	More than half the days	Nearly every day
Being so restless that it is hard to sit still	○	○	○	○
Becoming easily annoyed or irritable	○	○	○	○
Feeling afraid, as if something awful might happen	○	○	○	○

Over the last 2 weeks, how often have you been bothered by any of the following problems?

	Not at all	Several days	More than half the days	Nearly every day
Little interest or pleasure in doing things	○	○	○	○
Feeling down, depressed, or hopeless	○	○	○	○
Trouble falling or staying asleep, or sleeping too much	○	○	○	○
Feeling tired or having little energy	○	○	○	○

	Not at all	Several days	More than half the days	Nearly every day
Poor appetite or overeating	○	○	○	○
Feeling bad about yourself, or that you are a failure, or have let yourself or your family down	○	○	○	○
Trouble concentrating on things, such as reading the newspaper or watching television	○	○	○	○
Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	○	○	○	○

How motivated are you to work on your mental health from 0 (not motivated at all) to 100 (very motivated)?

0 10 20 30 40 50 60 70 80 90 100

[0]

mechanisms_therapy v002

Please indicate how frequently you have had each of the following thoughts in the past two weeks.

	Not at all	Sometimes	About half the time	Most of the time	All the time
I'm no good.	○	○	○	○	○

	Not at all	Sometimes	About half the time	Most of the time	All the time
I'm so disappointed in myself.	○	○	○	○	○
What's wrong with me?	○	○	○	○	○
I'm worthless.	○	○	○	○	○
I feel so helpless.	○	○	○	○	○
Something has to change.	○	○	○	○	○
My future is bleak.	○	○	○	○	○
I can't finish anything.	○	○	○	○	○

During the past two weeks, how often did you do something you genuinely enjoyed or found meaningful?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

During the past two weeks, how often did you put off or skip activities because you felt low-energy, down, or had negative thoughts?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

When something unclear or unexpected happened the past two weeks, how often did you immediately expect it to turn out badly?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

How often did you notice your attention being pulled toward potential threats, dangers, or negative cues, rather than neutral or positive information during the past two weeks?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

What is the maximum amount that you would be willing to pay for a 50 minute **in person** therapy session with a licensed psychologist?

Amount in $ [ ]

If you did **in person therapy** with a licensed psychologist, what do you think is the probability that your mental health would improve?

Percent (0-100): [ ]

Demand for meds and therapy

How likely are you to **<u>want</u> prescription medication for mental health** (such as Prozac, Xanax, or Seroquel) within the **next month**?

Please answer based on your personal desire for medication, even if you don't have access to it or cannot obtain it.

Very Unlikely … Very Likely

0 10 20 30 40 50 60 70 80 90 100

[ ]

How likely are you to **<u>want</u> psychotherapy from a licensed professional** (human) within the **next month**?

Please answer based on your personal desire for therapy, even if you don't have access to it or cannot obtain it.

Very Unlikely … Very Likely

0 10 20 30 40 50 60 70 80 90 100

[ ]

AI_perceptions

Do you agree or disagree with the following statement?

"Most people will judge me as weak if I admit I have mental health problems."

- Strongly Agree
- Somewhat Agree
- Neutral
- Somewhat disagree
- Strongly disagree

How comfortable are you to share personal feelings with

	Not at all comfortable 0 10 20 30 40 50 60 70 80 90 100 Very comfortable
the **AI-based** app Sonia?	
a **human** mental-health professional (e.g., therapist, counselor)?	

Compared to therapy with a human, how **effective** is emotional support from an AI for you personally?

- Much less effective
- Somwhat less effective
- Just as effective
- Somewhat more effective
- Much more effective

Therapy can come from either a human therapist or an AI system. For you personally, would one of these be better than the other? If yes, why?

User Satisfaction Sonia

How much would you willing to pay for a 50-minute Sonia session?

Amount in $

How satisfied were you with your experience using the Sonia app?

Very unsatisfied 0 10 20 30 40 50 60 70 80 90 100 Very satisfied

How likely are you to recommend the Sonia app to a friend or family member?

Very unlikely 0 10 20 30 40 50 60 70 80 90 100 Very Likely

Therapy can come from either a human therapist or an AI system. For you personally, would one of these be better than the other? If yes, why?

therapeutic_alliance

Please describe which category best describes your own experiences with the Sonia app.

	Seldom	Sometimes	Fairly Often	Very Often
As a result of the sessions I am clearer as to how I might be able to change.				
What I am doing in the sessions gives me new ways of looking at my problem.				
I believe the AI likes me.				
The AI and I collaborate on setting goals for my therapy.				
The AI and I respect each other.				
The AI and I are working towards mutually agreed upon goals.				

	Seldom	Sometimes	Fairly Often	Very Often
I feel that the AI appreciates me.				
The AI and I agree on what is important for me to work on.				

	Seldom	Sometimes	Fairly Often	Very Often
I feel the AI cares about me even when I do things that he/she does not approve of.				
I feel that the things I do in the sessions will help me to accomplish the changes that I want.				
The AI and I have established a good understanding of the kind of changes that would be good for me.				
I believe the way we are working with my problem in the sessions is correct.				

AI_perception_control

Do you agree or disagree with the following statement?

"Most people will judge me as weak if I admit I have mental health problems."

- Strongly Agree
- Somewhat Agree
- Neutral
- Somewhat disagree
- Strongly disagree

How comfortable are you to share personal feelings with

	Not at all comfortable 0 10 20 30 40 50 60 70 80 90 100 Very comfortable
an **AI-based** system	
a **human** mental-health professional (e.g., therapist, counselor)?	

Compared to therapy with a human, how **effective** would emotional support from an AI be?

- Much less effective
- Somwhat less effective
- Just as effective
- Somewhat more effective
- Much more effective

Therapy can come from either a human therapist or an AI system. For you personally, would one of these be better than the other? If yes, why?

Loneliness

Please answer the following questions:

	Hardly ever	Some of the time	Often
How often do you feel that you lack companionship?			
How often do you feel left out?			
How often do you feel isolated from others?			

End of Survey

Thank you for completing the survey! 🎉

We'll invite you to a follow-up survey in about 1 week.

Remember to keep engaging with the app!

Your Upcoming Surveys:

- **2-Week Survey: $15** for completing the follow-up survey we'll send **in 1 week**

- **Final Survey:** **$20** for completing the final survey at the end of the study **in 3 weeks**

📱 **Reminder:** Continue using the app throughout the study period!

Click next to submit your response.

Questions? Feel free to contact us at: team@mental-health-research.com

Block 10

Powered by Qualtrics

F.4 RCT Survey Instructions Wave 4

Consent_Sonia_Trial

Do you consent to the processing of your data and our privacy policy?

▶ *Click **here** to display the full privacy policy.*

Declaration of Consent

I hereby consent to the processing of my personal data. I can withdraw my consent at any time. I agree that my consent also extends to special categories of personal data.

- ○ I have read the above and consent to take part in this study
- ○ I do not wish to participate

Mental GAD7 PHQ8

Over the last two weeks, how often have you been bothered by the following problems?

	Not at all	Several days	More than half the days	Nearly every day
Feeling nervous, anxious, or on edge	○	○	○	○
Not being able to stop or control worrying	○	○	○	○
Worrying too much about different things	○	○	○	○
Trouble relaxing	○	○	○	○

	Not at all	Several days	More than half the days	Nearly every day
Being so restless that it is hard to sit still	○	○	○	○
Becoming easily annoyed or irritable	○	○	○	○
Feeling afraid, as if something awful might happen	○	○	○	○

Over the last 2 weeks, how often have you been bothered by any of the following problems?

	Not at all	Several days	More than half the days	Nearly every day
Little interest or pleasure in doing things	○	○	○	○
Feeling down, depressed, or hopeless	○	○	○	○
Trouble falling or staying asleep, or sleeping too much	○	○	○	○
Feeling tired or having little energy	○	○	○	○

	Not at all	Several days	More than half the days	Nearly every day
Poor appetite or overeating	○	○	○	○
Feeling bad about yourself, or that you are a failure, or have let yourself or your family down	○	○	○	○
Trouble concentrating on things, such as reading the newspaper or watching television	○	○	○	○
Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	○	○	○	○

How motivated are you to work on your mental health from 0 (not motivated at all) to 100 (very motivated)?

0 10 20 30 40 50 60 70 80 90 100

[0]

mechanisms_therapy v002

Please indicate how frequently you have had each of the following thoughts in the past two weeks.

	Not at all	Sometimes	About half the time	Most of the time	All the time
I'm no good.	○	○	○	○	○

	Not at all	Sometimes	About half the time	Most of the time	All the time
I'm so disappointed in myself.	○	○	○	○	○
What's wrong with me?	○	○	○	○	○
I'm worthless.	○	○	○	○	○
I feel so helpless.	○	○	○	○	○
Something has to change.	○	○	○	○	○
My future is bleak.	○	○	○	○	○
I can't finish anything.	○	○	○	○	○

During the past two weeks, how often did you do something you genuinely enjoyed or found meaningful?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

During the past two weeks, how often did you put off or skip activities because you felt low-energy, down, or had negative thoughts?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

When something unclear or unexpected happened the past two weeks, how often did you immediately expect it to turn out badly?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

How often did you notice your attention being pulled toward potential threats, dangers, or negative cues, rather than neutral or positive information during the past two weeks?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

What is the maximum amount that you would be willing to pay for a 50 minute **in person** therapy session with a licensed psychologist?

Amount in $ [ ]

If you did **in person therapy** with a licensed psychologist, what do you think is the probability that your mental health would improve?

Percent (0-100): [ ]

Safety Sonia new

Do you feel Sonia is a safe system **compared to** traditional AI chatbots?

- ○ Yes
- ○ No

You just said you do not feel Sonia is safe in comparison to traditional AI chatbots. Please describe your distressing experiences:

[ ]

therapeutic_alliance

You have been using the Sonia app for two weeks now.
Please describe which category best describes your own experiences with the AI-therapist **${e://Field/therapist_name}.**

	Seldom	Sometimes	Fairly Often	Very Often
As a result of the sessions I am clearer as to how I might be able to change.	○	○	○	○
What I am doing in the sessions gives me new ways of looking at my problem.	○	○	○	○

	Seldom	Sometimes	Fairly Often	Very Often
I believe the AI-therapist ${e://Field/therapist_name} likes me.	○	○	○	○
The AI-therapist ${e://Field/therapist_name} and I collaborate on setting goals for my therapy.	○	○	○	○
The AI-therapist ${e://Field/therapist_name} and I respect each other.	○	○	○	○
The AI-therapist ${e://Field/therapist_name} and I are working towards mutually agreed upon goals.	○	○	○	○
I feel that the AI-therapist ${e://Field/therapist_name} appreciates me.	○	○	○	○
The AI-therapist ${e://Field/therapist_name} and I agree on what is important for me to work on.	○	○	○	○

	Seldom	Sometimes	Fairly Often	Very Often
I feel the AI-therapist ${e://Field/therapist_name} cares about me even when I do things that he/she does not approve of.	○	○	○	○
I feel that the things I do in the sessions will help me to accomplish the changes that I want.	○	○	○	○
The AI-therapist ${e://Field/therapist_name} and I have established a good understanding of the kind of changes that would be good for me.	○	○	○	○
I believe the way we are working with my problem in the sessions is correct.	○	○	○	○

Loneliness

Please answer the following questions:

	Hardly ever	Some of the time	Often
How often do you feel that you lack companionship?	○	○	○
How often do you feel left out?	○	○	○
How often do you feel isolated from others?	○	○	○

End of Survey

Thank you for completing the survey! 🎉

Your reward of **$15** is on its way to you right now.

Your upcoming surveys:

You will receive $20 for completing the final survey at the end of the study in exactly in 2 weeks.

Questions? Feel free to contact us at: *team@mental-health-research.com*

📱 **Reminder:** Continue using the app throughout the study period!

Block 9

How likely are you to recommend the Sonia app to a friend or family member?

Very unlikely — Very Likely

0 10 20 30 40 50 60 70 80 90 100

How satisfied were you with your experience using the Sonia app?

Very unsatisfied — Very satisfied

0 10 20 30 40 50 60 70 80 90 100

How much would you be willing to pay for a 50-minute Sonia session?

Amount in $

Human Therapy Experience

If you compare it to talking with a human therapist, how effective would emotional support from an AI feel **to you personally?**

- ○ Much less effective
- ○ Somwhat less effective
- ○ Just as effective
- ○ Somewhat more effective
- ○ Much more effective

Would you feel comfortable to share personal feelings with **an AI-based system?**

Not at all effective — Very effective

0 10 20 30 40 50 60 70 80 90 100

Did you at any point feel under severe psychological distress or at risk while using the AI support app? Please describe your experience (if any).

- ○ Yes (please describe)
- ○ No

If these were your available options, which one would you choose right now?

- ○ Therapy with a **human** therapist
- ○ Therapy with an **AI** therapist
- ○ No preference — AI or human is equally acceptable to me
- ○ I would prefer not to do therapy at all

In weeks, how long do you think it would take you to find a **human therapist**?

Weeks:

If you had to pay for it yourself, how much do you think a 50-minute **AI-based therapy session** would cost?

Price in US-Dollar:

If you had to pay for it yourself, how much do you think a 50-minute session with a **human therapist** would cost?

Price in US-Dollar:

Price of Human and AI Therapy

In the next questions, we ask you to compare an **AI therapist and a human therapist.** If you have not talked to either a human or AI therapist before, please answer based on what you would expect.

For each of the following, who do you think is **better**: an AI therapist or a human therapist?

	A human therapist is much better	A human therapist is better	Human and AI therapist are about the same	An AI therapist is better	An AI therapist is much better
Not judging or embarrassing you.	○	○	○	○	○
Showing empathy and really understanding you.	○	○	○	○	○
Available to give advice whenever you need it.	○	○	○	○	○
Being trustworthy and keeping what you say private.	○	○	○	○	○
Having strong professional knowledge about mental health.	○	○	○	○	○
Doesn't make you feel socially anxious.	○	○	○	○	○
Treating you with neutrality and without bias.	○	○	○	○	○

“Have you ever had sessions with a human therapist/counselor?”

- ○ Yes
- ○ No

What aspects of your sessions with your therapist/counselor did you find **most valuable?**

What did you **not like** about your sessions with the **human** therapist/counselor?

Block 10

We are interested in **what you personally value** when you receive emotional or psychological support — whether it is from a friend, family member, AI or human therapist.

How important are the following for you when receiving advice?

	Not important at all 0	10	20	30	40	50	60	70	80	90	Very Important 100
Not being judged or embarrassed.											
A human connection.											
Availability to receive advice whenever you need it.											
Professional knowledge about mental health.											
Trustworthiness and privacy.											
Not feeling socially anxious.											
Empathy and feeling truly understood.											
Being treated with neutrality and without bias.											

In a previous survey, you mentioned that you are not currently in **human therapy**. What are the reasons for this?

Powered by Qualtrics

28/01/2026, 12:39 Qualtrics Survey Software

F.5 RCT Survey Instructions Wave 5

Consent_Sonia_Trial

Do you consent to the processing of your data and our privacy policy?

▶ *Click **here** to display the full privacy policy.*

Declaration of Consent

I hereby consent to the processing of my personal data. I can withdraw my consent at any time. I agree that my consent also extends to special categories of personal data.

- ○ I have read the above and consent to take part in this study
- ○ I do not wish to participate

Mental GAD7 PHQ8

Over the last two weeks, how often have you been bothered by the following problems?

	Not at all	Several days	More than half the days	Nearly every day
Feeling nervous, anxious, or on edge	○	○	○	○
Not being able to stop or control worrying	○	○	○	○
Worrying too much about different things	○	○	○	○
Trouble relaxing	○	○	○	○

	Not at all	Several days	More than half the days	Nearly every day
Being so restless that it is hard to sit still	○	○	○	○
Becoming easily annoyed or irritable	○	○	○	○
Feeling afraid, as if something awful might happen	○	○	○	○

Over the last 2 weeks, how often have you been bothered by any of the following problems?

	Not at all	Several days	More than half the days	Nearly every day
Little interest or pleasure in doing things	○	○	○	○
Feeling down, depressed, or hopeless	○	○	○	○
Trouble falling or staying asleep, or sleeping too much	○	○	○	○
Feeling tired or having little energy	○	○	○	○

	Not at all	Several days	More than half the days	Nearly every day
Poor appetite or overeating	○	○	○	○
Feeling bad about yourself, or that you are a failure, or have let yourself or your family down	○	○	○	○
Trouble concentrating on things, such as reading the newspaper or watching television	○	○	○	○
Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	○	○	○	○

How motivated are you to work on your mental health from 0 (not motivated at all) to 100 (very motivated)?

0 10 20 30 40 50 60 70 80 90 100

[0]

mechanisms_therapy v002

Please indicate how frequently you have had each of the following thoughts in the past two weeks.

	Not at all	Sometimes	About half the time	Most of the time	All the time
I'm no good.	○	○	○	○	○

	Not at all	Sometimes	About half the time	Most of the time	All the time
I'm so disappointed in myself.	○	○	○	○	○
What's wrong with me?	○	○	○	○	○
I'm worthless.	○	○	○	○	○
I feel so helpless.	○	○	○	○	○
Something has to change.	○	○	○	○	○
My future is bleak.	○	○	○	○	○
I can't finish anything.	○	○	○	○	○

During the past two weeks, how often did you do something you genuinely enjoyed or found meaningful?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

During the past two weeks, how often did you put off or skip activities because you felt low-energy, down, or had negative thoughts?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

When something unclear or unexpected happened the past two weeks, how often did you immediately expect it to turn out badly?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

How often did you notice your attention being pulled toward potential threats, dangers, or negative cues, rather than neutral or positive information during the past two weeks?

- ○ Less than 20% of the time
- ○ 20–40% of the time
- ○ 40–60% of the time
- ○ 60–80% of the time
- ○ Over 80% of the time

What is the maximum amount that you would be willing to pay for a 50 minute **in person** therapy session with a licensed psychologist?

Amount in $ [ ]

If you did **in person therapy** with a licensed psychologist, what do you think is the probability that your mental health would improve?

Percent (0-100): [ ]

talkspace

Psychotherapy with **human therapists** is also available *online* through video calls.

Would you like us to send you an email with more information about providers of remote video therapy and their offerings?

- ○ Yes
- ○ No

We are not affiliated with any provider.

Human_vs_AI_socialis

During the past 4 weeks, how often did you have meaningful social interaction with other people (in person, by phone, etc.)?

- ○ About once a week or less
- ○ About twice a week
- ○ 3-4 times a week
- ○ 5-6 times a week
- ○ Once a day
- ○ Several times a day
- ○ Many times a day

During the past 4 weeks, how often did you seek emotional support from friends or family?

- About once a week or less
- About twice a week
- 3-4 times a week
- 5-6 times a week
- Once a day
- Several times a day
- Many times a day

Compared to before the trial, I now feel more comfortable seeking emotional support from friends or family.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

Stigma

Do you agree or disagree with the following statement?

"Most people will judge me as weak if I admit I have mental health problems."

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

perceptions_change_Sonia

Did your belief change about what AI-therapy actually is after using the Sonia app?
(For example, this could include changes in your expectations about how helpful and comprehensive AI therapy would be)

- Yes, my beliefs changed after using the Sonia app
- No, my beliefs stayed the same

Did using the Sonia app change how positive or negative you feel about therapy in general?

- I feel much more positive about therapy in general
- I feel slightly more positive about therapy in general

- My feelings stayed the same
- I feel slightly more negative about therapy in general
- I feel much more negative about therapy in general

Control AI Emotional Support

Did you use AI tools (for example ChatGPT, Google Gemini) for emotional support during this trial?

- Very often
- Often
- Sometimes
- Seldom
- Never

Demand for meds and therapy

How likely are you to **want prescription medication for mental health** (such as Prozac, Xanax, or Seroquel) within the **next month**?

Please answer based on your personal desire for medication, even if you don't have access to it or cannot obtain it.

Very Unlikely Very Likely

How likely are you to **want psychotherapy from a licensed professional** (human) within the **next month**?

Please answer based on your personal desire for therapy, even if you don't have access to it or cannot obtain it.

Very Unlikely Very Likely
0 10 20 30 40 50 60 70 80 90 100

Loneliness

Please answer the following questions:

	Hardly ever	Some of the time	Often
How often do you feel that you lack companionship?	○	○	○
How often do you feel left out?	○	○	○
How often do you feel isolated from others?	○	○	○

Human Therapy Experience

Are you currently searching for a **human** therapist?

- Yes
- No

In a previous survey, you told us that you are not currently in **human** therapy. Why? Select all reasons that apply.

- [ ] I don't think I need it.
- [ ] I can't find any therapist.
- [ ] I don't feel comfortable talking to a human about my mental health.
- [ ] I can't find a therapist who feels like the right match for me.
- [ ] My family/friends would disapprove.
- [ ] I don't have enough energy for therapy.
- [ ] I prefer digital help/help from an AI.
- [ ] I don't have time.
- [ ] I can't afford it.
- [ ] None of the above
- [ ] I just found a human therapist

During this trial, did you do any of the following?

- [ ] Take medication medication for mental health (e.g. antidepressants)
- [ ] Began or restart human-led psychotherapy
- [ ] None of the above

Have you *ever* had sessions with a human therapist/counselor?

- Yes
- No

What aspects of your sessions with your **human** therapist/counselor did you find **most valuable?**

What did you **not like** about your sessions with the **human** therapist/counselor?

ExpDemand and SDB

Do you agree or disagree with the following statement?

"I am able to admit **to myself** when I feel down, anxious, or overwhelmed."

- Strongly disagree
- Disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Agree
- Strongly agree

Do you agree or disagree with the following statement?

"Among my closest friends and immediate family, it is **acceptable to say** when I feel down, anxious, or overwhelmed."

- Strongly disagree
- Disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Agree
- Strongly agree

Did you adjust any of your answers in this or previous surveys because you thought we wanted you to answer in a particular way?

- ○ Yes
- ○ No

What do you think was the purpose of this study and the earlier ones you were invited to by us?

End_of_survey

Thank you so much for being part of this study! 🎉

We've just sent your **$20 gift card** as a thank-you for completing this final survey.

Before you finish, we'd love to hear from you — is there anything you'd like to share about your experience with the study?

Powered by Qualtrics

F.6 Representative Survey Instructions

Consent_Sonia_Trial

Welcome!

I'm not a robot reCAPTCHA Privacy - Terms

Do you consent to the processing of your data and our privacy policy?

▶ *Click **here** to display the full privacy policy.*

Declaration of Consent
I hereby consent to the processing of my personal data. I can withdraw my consent at any time. I agree that my consent also extends to special categories of personal data.

- ○ I have read the above and consent to take part in this study
- ○ I do not wish to participate

Screener: Human attention check

The next question is about the following problem. In questionnaires like ours, sometimes there are participants who do not carefully read the questions and just quickly click through the survey. This means that there are a lot of random answers which compromise the results of research studies. To show that you read our questions carefully, please choose both "Very strongly interested" and "Not at all interested" as your answer in this question.

How interested are you in sports?

- ☐ Very strongly interested
- ☐ Very interested
- ☐ A little bit interested
- ☐ Almost not interested
- ☐ Not at all interested

Screener: LLM audio check

Please follow the instructions in the audio to answer the question below.
This is to identify bots, as they cannot interpret audio easily.

0:00 / 0:01

Your answer: _____

GAD7

Over the last two weeks, how often have you been bothered by the following problems?

	Not at all	Several days	More than half the days	Nearly every day
Feeling nervous, anxious, or on edge	○	○	○	○
Not being able to stop or control worrying	○	○	○	○
Worrying too much about different things	○	○	○	○
Trouble relaxing	○	○	○	○

	Not at all	Several days	More than half the days	Nearly every day
Being so restless that it is hard to sit still	○	○	○	○
Becoming easily annoyed or irritable	○	○	○	○
Feeling afraid, as if something awful might happen	○	○	○	○

Treatment Experience and Demand

Please indicate your experience with **prescribed** mental health medication (for example Prozac, Xanax, Seroquel):

- ○ _____ I am currently taking medication, please tell us which:
- ○ I have taken medication in the past, but not currently
- ○ I have never taken medication

Have you been to professional therapy?

- ○ No, never
- ○ Yes, but more than a year ago.
- ○ Yes, currently in therapy.

Would you be interested in starting therapy with a human therapist, assuming no practical barriers (cost, availability, etc)?

- ○ Yes
- ○ No

Are you currently searching for a human therapist?

- ○ Yes
- ○ No

How likely are you to **want prescription medication for mental health** (such as Prozac, Xanax, or Seroquel) within the **next month**?
Please answer based on your personal desire for medication, even if you don't have access to it or cannot obtain it.

Very Unlikely … Very Likely
0 10 20 30 40 50 60 70 80 90 100
_____

How likely are you to **want psychotherapy from a licensed professional** (human) within the **next month**?
Please answer based on your personal desire for therapy, even if you don't have access to it or cannot obtain it.

Very Unlikely … Very Likely
0 10 20 30 40 50 60 70 80 90 100
_____

Beliefs About Human

If you did **in person therapy** with a licensed psychologist, what do you think is the probability that your mental health would improve?

Percent (0-100): _____

Info AI Systems and Beliefs about AI Therapy

AI-Based Emotional Support and Therapy

In recent years, many people have started using AI tools for emotional and psychological support.

In addition to the general-purpose AI chatbots (e.g. ChatGPT), there exist specialized AI systems developed under clinical guidance to support emotional and psychological well-being.

They often come in the form of a mobile app, through which a person can speak or chat with an AI, much like a conversation with a therapist. The following questions explore the role of AI in emotional support and therapy.

If you did therapy with an **AI-based system**, what do you think is the probability that your mental health would improve?

Percent (0-100) _____

Would you be interested in trying a specialized AI (clinically validated, HIPAA compliant) developed specifically for emotional support?

- ○ Yes
- ○ No

Emotional Support General

We are interested in **what you personally value** when you receive emotional or psychological support — whether it is from a friend, family member, AI or human therapist.

How important are the following for you when receiving advice?

Not important at all … Very Important
0 10 20 30 40 50 60 70 80 90 100

- Trustworthiness and privacy. _____
- Not feeling socially anxious. _____
- Availability to receive advice whenever you need it. _____
- Not being judged or embarrassed. _____
- Professional knowledge about mental health. _____
- A human connection. _____
- Being treated with neutrality and without bias. _____

Not important at all — Very Important

0 10 20 30 40 50 60 70 80 90 100

Empathy and feeling truly understood. [ ]

Do you agree or disagree with the following statement?

"Most people will judge me as weak if I admit I have mental health problems."

- Strongly Agree
- Somewhat Agree
- Neutral
- Somewhat disagree
- Strongly disagree

Emotional Support AI vs Human

In the next questions, we ask you to compare an **AI therapist and a human therapist.** If you have not talked to either a human or AI therapist before, please answer based on what you would expect.

For each of the following, who do you think is **better**: an AI therapist or a human therapist?

	A human therapist is much better	A human therapist is better	Human and AI therapist are about the same	An AI therapist is better	An AI therapist is much better
Not judging or embarrassing you.	○	○	○	○	○
Showing empathy and really understanding you.	○	○	○	○	○
Available to give advice whenever you need it.	○	○	○	○	○
Being trustworthy and keeping what you say private.	○	○	○	○	○
Having strong professional knowledge about mental health.	○	○	○	○	○
Doesn't make you feel socially anxious.	○	○	○	○	○
Treating you with neutrality and without bias.	○	○	○	○	○

How comfortable are you to share personal feelings with

Not at all comfortable — Very comfortable

0 10 20 30 40 50 60 70 80 90 100

an **AI-based** system [ ]

a **human** mental-health professional (e.g., therapist, counselor)? [ ]

If you compare it to talking with a human therapist, how effective would emotional support from an AI feel **to you personally?**

- Much less effective
- Somwhat less effective
- Just as effective
- Somewhat more effective
- Much more effective

How concerned are you **about privacy** when using a **human** for emotional support?

- Not at all concerned
- Slightly concerned
- Moderately concerned
- Very concerned
- Extremely concerned

How concerned are you **about privacy** when using an **AI-system** for emotional support?

- Not at all concerned
- Slightly concerned
- Moderately concerned
- Very concerned
- Extremely concerned

How comfortable are you using AI for help with personal experiences (e.g. relationships)?

Very uncomfortable — Very comfortable

0 10 20 30 40 50 60 70 80 90 100

[ ]

Considering you could also use an **AI app**, would you be interested in starting therapy with a **human** therapist, assuming no practical barriers (cost, availability, etc)?

- Yes
- No

If these were your available options, which one would you choose right now?

- Therapy with a **human** therapist
- Therapy with an **AI** therapist
- No preference — AI or human is equally acceptable to me
- I would prefer not to do therapy at all

Do you use any AI tools for emotional support (e.g. venting, therapy-like sessions, sharing and discussing your feelings and emotions, etc.)

- Very often
- Often
- Sometimes
- Seldom
- Never

In a study conducted in 2023, licensed clinicians judged whether answers from an AI (similar to ChatGPT) to standard medical questions were in line with the scientific consensus.

What percent of the AI's answers do you think clinicians judged as in line with scientific consensus?

Your estimate (0–100): [ ]

PHQ8 and Motivation

Over the last 2 weeks, how often have you been bothered by any of the following problems?

	Not at all	Several days	More than half the days	Nearly every day
Little interest or pleasure in doing things	○	○	○	○
Feeling down, depressed, or hopeless	○	○	○	○

	Not at all	Several days	More than half the days	Nearly every day
Trouble falling or staying asleep, or sleeping too much	○	○	○	○
Feeling tired or having little energy	○	○	○	○

	Not at all	Several days	More than half the days	Nearly every day
Poor appetite or overeating	○	○	○	○
Feeling bad about yourself, or that you are a failure, or have let yourself or your family down	○	○	○	○
Trouble concentrating on things, such as reading the newspaper or watching television	○	○	○	○
Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	○	○	○	○

How motivated are you to work on your mental health from 0 (not motivated at all) to 100 (very motivated)?

0 10 20 30 40 50 60 70 80 90 100

[0]

Demographics

What is your gender?

- Male
- Female
- Non-binary
- Other / Prefer not to say

How old are you?

[ ]

Choose one or more races that you consider yourself to be

- ☐ White or Caucasian
- ☐ Black or African American
- ☐ American Indian/Native American or Alaska Native
- ☐ Asian
- ☐ Native Hawaiian or Other Pacific Islander
- ☐ Other
- ☐ Prefer not to say

What is the highest level of education you have completed?

- ○ Some high school or less
- ○ High school diploma or GED
- ○ Some college, but no degree
- ○ Associates or technical degree
- ○ Bachelor's degree
- ○ Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS etc.)
- ○ Prefer not to say

What was your annual household income in 2024 in US dollars **before taxes and deductions**?

Note: The household income is the total amount of money earned by every member of your household.

- ○ Less than $15,000
- ○ Between $15,000 and $25,000
- ○ Between $25,000 and $50,000
- ○ Between $50,000 and $75,000
- ○ Between $75,000 and $100,000
- ○ Between $100,000 and $150,000
- ○ Between $150,000 and $200,000
- ○ More than $200,000

What best describes your employment status over the last three months?

- ○ Working full-time
- ○ Working part-time
- ○ Unemployed and looking for work
- ○ A homemaker or stay-at-home parent
- ○ Student
- ○ Retired
- ○ Other

What is your current marital status?

- ○ Married
- ○ Living with a partner
- ○ Widowed
- ○ Divorced/Separated
- ○ Never been married

Powered by Qualtrics

References

Judith S. Beck. *Cognitive Behavior Therapy: Basics and Beyond*. The Guilford Press, New York, third edition, 2021. ISBN 9781462544196.

David D. Burns. *Feeling Good: The New Mood Therapy*. William Morrow, 1980.

Leonardo Bursztyn, Ingar K Haaland, Nicolas Röver, and Christopher Roth. The social desirability atlas. Technical report, National Bureau of Economic Research, 2025.

Felix Chopra and Ingar Haaland. Conducting qualitative interviews with ai. 2023.

Michelle G. Craske. *Cognitive-Behavioral Therapy*. Oxford University Press, Oxford, 2016.

Jonathan De Quidt, Johannes Haushofer, and Christopher Roth. Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302, 2018.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Mathias Harrer, Antonia A. Sprenger, Paula Kuper, Eirini Karyotaki, and Pim Cuijpers. *metapsyData: Access the Meta-Analytic Psychotherapy Databases in R*, 2022. URL `https://data.metapsy.org`.

Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A. Banta, Abi D. Jewett, Abigail J. Salzhauer, Tess Z. Griffin, and Nicholas C. Jacobson. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI*, 2(4), March 2025. doi: 10.1056/AIoa2400802. URL `https://doi.org/10.1056/AIoa2400802`. Published March 27, 2025. Copyright © 2025.

David S. Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102, 2009. doi: 10.1111/j.1467-937X.2009.00536.x.

Christopher Roth, Peter Schwardmann, and Egon Tripodi. Depression stigma. Discussion Paper DP18919, Centre for Economic Policy Research, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Publisher correction: Large language models encode clinical knowledge. *Nature*, 620(7973):E19, 2023. doi: 10.1038/s41586-023-06455-0.